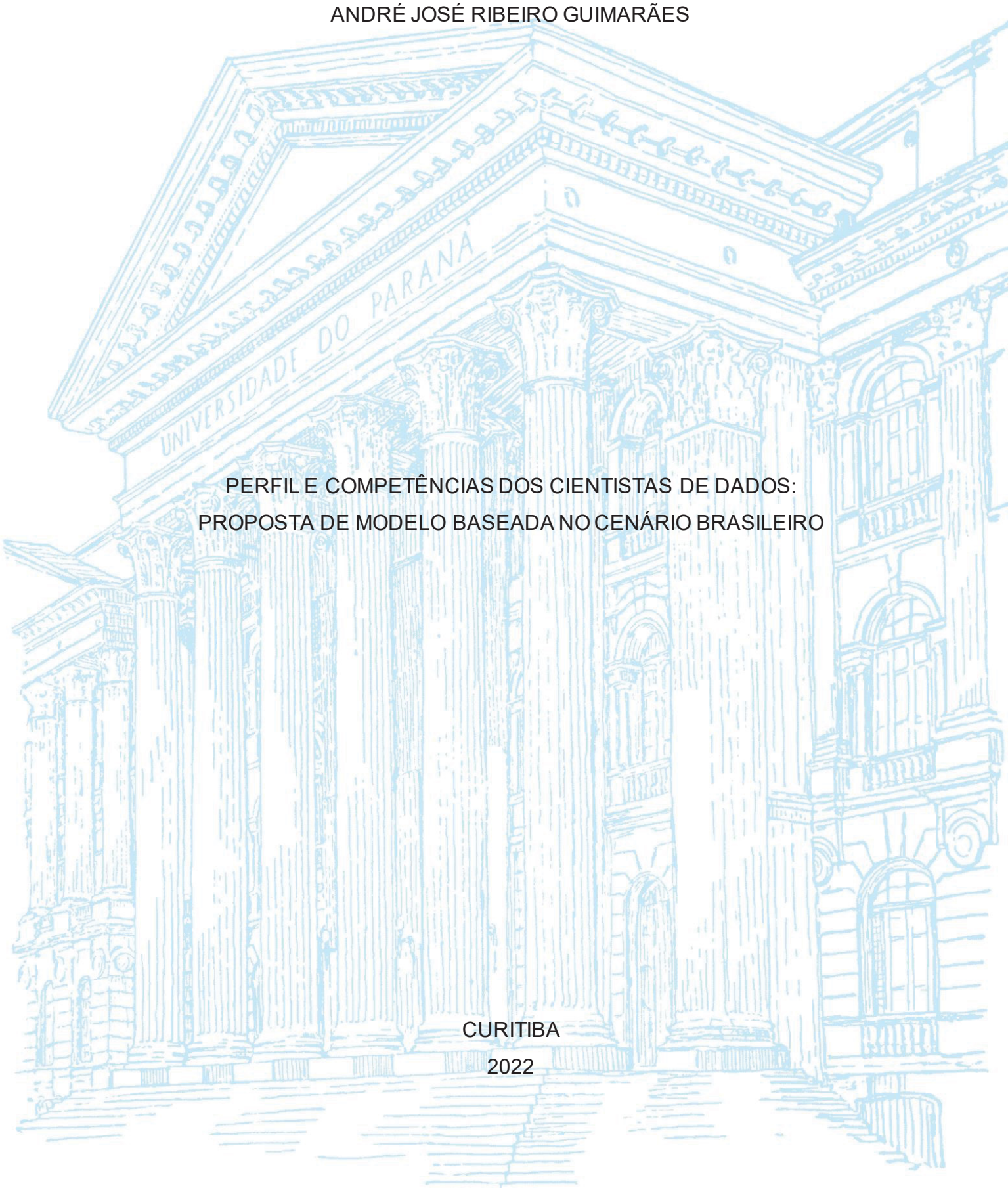
UNIVERSIDADE FEDERAL DO PARANÁ

ANDRÉ JOSÉ RIBEIRO GUIMARÃES



PERFIL E COMPETÊNCIAS DOS CIENTISTAS DE DADOS: PROPOSTA DE MODELO BASEADA NO CENÁRIO BRASILEIRO

CURITIBA

2022

ANDRÉ JOSÉ RIBEIRO GUIMARÃES

PERFIL E COMPETÊNCIAS DOS CIENTISTAS DE DADOS: PROPOSTA DE MODELO BASEADA NO CENÁRIO BRASILEIRO

Tese apresentada ao Programa de Pós-Graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná, como requisito parcial para a obtenção do título de doutor.

Orientadora: Dra. Maria do Carmo Duarte Freitas

Coorientador: Dr. Ricardo Mendes Junior

CURITIBA

2022

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIAS SOCIAIS APLICADAS

Guimarães, André José Ribeiro
Perfil e competências dos cientistas de dados : proposta de modelo baseada no cenário brasileiro / André José Ribeiro Guimarães. – Curitiba, 2022.
1 recurso on-line : PDF.

Tese (Doutorado) – Universidade Federal do Paraná, Setor de Ciências Sociais Aplicadas, Programa de Pós-Graduação em Gestão da Informação.
Orientadora: Profa. Dra. Maria do Carmo Duarte Freitas.
Orientador: Prof. Dr. Ricardo Mendes Junior.

1. Ciência de dados. 2. Mapeamento de competências. 3. Tecnologia da informação. I. Freitas, Maria do Carmo Duarte. II. Mendes Junior, Ricardo. III. Universidade Federal do Paraná. Programa de Pós-Graduação em Gestão da Informação. IV. Título.

Bibliotecária: Maria Lidiane Herculano Graciosa CRB-9/2008

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GESTÃO DA INFORMAÇÃO - 40001016058P1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação GESTÃO DA INFORMAÇÃO da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **ANDRÉ JOSÉ RIBEIRO GUIMARÃES** intitulada: **PERFIL E COMPETÊNCIAS DOS CIENTISTAS DE DADOS: PROPOSTA DE MODELO BASEADA NO CENÁRIO BRASILEIRO**, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 26 de Setembro de 2022.

Assinatura Eletrônica
09/11/2022 14:05:46.0
RICARDO MENDES JUNIOR
Presidente da Banca Examinadora

Assinatura Eletrônica
08/11/2022 15:16:00.0
HELENA DE FÁTIMA NUNES SILVA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
09/11/2022 16:43:18.0
CEZAR KARPINSKI
Avaliador Externo (UNIVERSIDADE FEDERAL DE SANTA CATARINA)

Assinatura Eletrônica
15/11/2022 07:10:24.0
JOSÉ ANTÓNIO BARATA DE OLIVEIRA
Avaliador Externo (UNIVERSIDADE NOVA DE LISBOA - NOVA IMS)

Assinatura Eletrônica
14/11/2022 22:55:14.0
ALEXANDRE AUGUSTO BIZ
Avaliador Externo (UNIVERSIDADE FEDERAL DE SANTA CATARINA)

À minha amada família.

À Regiane, que me mostrou que a vida pode ser boa.

Ao Vicente, minha maior motivação para ser uma pessoa melhor.

AGRADECIMENTOS

Agradeço, inicialmente, a todos que colaboraram de alguma forma para o desenvolvimento desta tese.

Especialmente, gostaria de agradecer aos meus orientadores, Prof.ª Dr.ª Maria do Carmo Duarte Freitas e Prof. Ricardo Mendes Junior, pela oportunidade, pelo suporte, pela paciência, pela compreensão, pela empatia, mas sobretudo, pelo compartilhamento de conhecimento e experiência.

Aos professores Dr. Alexandre Augusto Biz, Dr. Cezar Karpinski, Dr.ª Helena de Fátima Nunes Silva e Dr. José Antonio Barata de Oliveira, pelo aceite em participar da banca de defesa desta tese, pelo respeito e, principalmente, pelas valiosas contribuições à pesquisa.

Aos professores do Programa de Pós-Graduação em Gestão da Informação, em especial, aos professores Dr. Cicero Aparecido Bezerra, Dr.ª Denise Fukumi Tsunoda, Dr.ª Paula Carina de Araújo, Dr. Rodrigo Eduardo Botelho Francisco e Dr.ª Taiane Ritta Coelho pelo aprendizado e pelo exemplo.

À Simone, cujas ações ultrapassam em muito as funções de secretariado, fazendo tudo para deixar nosso doutorado mais leve, fácil e acolhedor.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES pela concessão de bolsa de doutorado do período de abril de 2018 a setembro de 2022.

Aos colegas de pós-graduação, especialmente à turma de doutorado de 2018, pelo apoio, pelas contribuições ao trabalho, pelos momentos de descontração e compartilhamento do conhecimento. Nominalmente, agradeço aos amigos Flávia, Luana, Luiz e Rodrigo.

Aos irmãos que a vida me deu, Gino Marcomini (*in memoriam*), Heitor Magnani e Carlos Bittencourt, pelas conversas, por ouvir os desabafos e pela força que sempre me deram.

À Duda, à Mel e ao Nick, pelos passeios, pelos carinhos, pelas brincadeiras e pelas patas amigas.

Aos meus amados pais, Amélia e Antonio, por serem meu abrigo quando mais precisei.

Aos meus irmãos e sobrinhos, por serem razão de motivação para mim, por me apoiarem sempre e por compreenderem minhas constantes ausências.

Agradeço a Deus, por nossa reaproximação.

E agradeço à Regi, meu exemplo de competência, determinação, foco e, principalmente, de como a vida pode ser vivida de uma maneira simples e leve. Obrigado por partilhar a vida boa comigo. Obrigado pelo caminho de amor que temos pela frente.

RESUMO

Ciência de Dados é um campo interdisciplinar, em desenvolvimento, que surge na interseção de estatística, tecnologia da informação e conhecimento de domínio. A Ciência de dados busca extrair conhecimento de dados brutos, por meio de modelos estatísticos, para auxiliar a tomada de decisão organizacional, trazendo benefícios para todas as áreas da sociedade. As organizações precisam de profissionais habilitados nestas áreas, porém a contratação de um cientista de dados ainda é uma tarefa árdua e custosa, especialmente pela carência de pessoas qualificadas. Esta tese analisa as competências necessárias para a atuação de cientistas de dados no Brasil e propõe um modelo de competências elaborado com base em pesquisa realizada junto a profissionais da Ciência de Dados. O modelo proposto organiza as competências da Ciência de Dados em quatro dimensões: Tecnologia, Análise de Dados, Entendimento de Negócios e Competências Socioculturais, e fornece subsídios para a educação, contratação e autoavaliação de profissionais da área. O instrumento de coleta de dados apresenta 227 respostas completas e válidas que fundamentam a definição do perfil dos profissionais da Ciência de Dados que atuam no Brasil. A pesquisa de levantamento confirma a Ciência de Dados como um campo recente, formado por profissionais jovens e com relativo pouco tempo na área, cuja origem advém de disciplinas estabelecidas, como Estatística e Ciência da Computação. Em relação às competências, a Análise Fatorial Confirmatória resulta em um modelo de segunda ordem, composto por quatro fatores que apresenta um ótimo ajustamento (RMSEA = 0,035, GFI= 0,974, TLI = 0,992, SRMR = 0,081). O modelo estatístico em quatro dimensões demonstra boa confiabilidade, validade e credibilidade, mostrando-se uma ferramenta efetiva na avaliação das competências para a Ciência de Dados. Além da pesquisa de levantamento, a tese identifica os requisitos em anúncios de vagas de emprego para cientistas de dados e os tópicos abordados em cursos de nível superior e cursos livres para este profissional. Para analisar estes documentos, obtidos com raspagem de dados, utiliza métodos de mineração de texto: n-grama, modelagem de tópico e agrupamento. A mineração dos anúncios aponta uma concentração de vagas em São Paulo, mas revela que a modalidade remota é a segunda mais ofertada. Destaca que os salários no Brasil estão abaixo da média de outros países, mesmo que as organizações procurem por profissionais experientes e com alto nível educacional. Quanto aos requisitos, há o predomínio de habilidades técnicas como machine learning, modelos estatísticos, Python, banco de dados, dentre outras. A análise dos cursos superiores e cursos livres apresenta alinhamento com os anúncios e com as competências definidas na pesquisa de levantamento. Ainda assim, os três conjuntos de documentos analisados, anúncios, cursos superiores e cursos livres apresentam características próprias. Para as técnicas de mineração, a pesquisa demonstra que n-grama e o agrupamento são mais adequadas que a modelagem de tópicos. Por fim, ainda que se julgue que as características da área no Brasil sejam distintas de outros países, o modelo pode ser replicado em pesquisas futuras, dentro e fora do território brasileiro.

Palavras-chave: Ciência de dados. Cientista de dados. Mapeamento de competências. Modelo de competências. Competências em Ciência de Dados.

ABSTRACT

Data Science is a developing, interdisciplinary field that arises at the intersection of statistics, information technology, and domain knowledge. Data Science seeks to extract knowledge from raw data through statistical models, to aid organizational decision-making, bringing benefits to all areas of society. Organizations need professionals skilled in these areas but hiring a data scientist is still arduous and costly, especially due to the lack of qualified people. This thesis analyzes the competencies required for data scientists in Brazil and proposes a competency model based on a survey conducted with Data Science professionals. The proposed model organizes Data Science competencies into four dimensions: Technology, Data Analysis, Business Understanding, and Sociocultural Competencies, and provides subsidies for the education, hiring, and self-assessment of Data Science professionals. The data collection instrument presents 227 complete and valid responses that support the definition of the profile of Data Science professionals working in Brazil. The survey confirms Data Science as a recent field, formed by young professionals with relatively quick time in the area, whose origin comes from established disciplines such as Statistics and Computer Science. Regarding competencies, Confirmatory Factor Analysis results in a second-order model composed of four factors that presents a very good fit (RMSEA = 0.035, GFI= 0.974, TLI = 0.992, SRMR = 0.081). The statistical model in four dimensions shows good reliability, validity, and credibility, proving to be an effective tool in assessing Data Science competencies. In addition to the survey research, the thesis identifies the requirements in job advertisements for Data Scientists and the topics covered in college-level courses and free courses for this professional. The thesis uses text mining methods (n-gram, topic modeling, and clustering) to analyze these documents obtained with data scraping. The text mining of the ads points out a concentration of vacancies in São Paulo but reveals that the remote modality is the second most offered. It highlights that salaries in Brazil are below the average of other countries, even though organizations are looking for experienced and highly educated professionals. As for the requirements, there is a predominance of technical skills such as machine learning, statistical models, Python, and databases, among others. The analysis of the higher education and free courses shows alignment with the job posts and the competencies defined in the survey research. Even so, the three sets of documents analyzed - advertisements, graduate courses, and open courses - have their characteristics. Regarding the mining techniques, the research shows that n-gram and clustering are more appropriate than topic modeling. Finally, even though the characteristics of Brazil are distinct from other countries, the model can be replicated in future research inside and outside Brazil..

Keywords: Data science. Data scientist. Competency mapping. Competency model. Competencies in Data Science.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE SIGLAS

AFC	Análise Fatorial Confirmatória
API	*Application Programming Interface* / Interface de Programação de Aplicações
AVE	*Average Variance Extracted* / Variância média extraída
B2B	Business-to-business
BDTD	Biblioteca Digital Brasileira de Teses e Dissertações
BI	Business Intelligence
Brapci	Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação
BS	Teste de Esfericidade de Bartlett
CF-DS	Data Science Competences Framework
CFI	Índice Bentler's Comparative Fit
CR	*Composite Reliability* / Confiabilidade Composta
DS-BoK	Data Science Body of Knowledge
DSPP	Data Science Professional Profiles
DWLS	Mínimos Quadrados Ponderados Diagonalmente
EaD	Ensino à Distância
EDSF	EDISON Data Science Framework
EPC	*Expected parameter change* / Alteração de parâmetro esperada
ESCO	*European Skills, Competences, Qualifications and Occupations*
GFI	Índice Goodness-of-fit
gl	Graus de liberdade
HTMT	*Heterotrait-monotrait ratio*
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IES	Instituição de Ensino Superior
IoT	*Internet of Things* / Internet das Coisas
KMO	Teste de Kaiser Meyer Olkin
LGPD	Lei Geral de Proteção de Dados Pessoais
MC-DS	Data Science Model Curriculum
MEC	Ministério da Educação
ML	*Machine learning*

NDLTD	Networked Digital Library of Theses and Dissertations
NIST	Instituto Nacional de Padrões e Tecnologia
NLP	*Natural Language Processing* / Processamento de Linguagem Natural
PAP	Profissional de Análise Preditiva
PPGGI	Programa de Pós-Graduação em Gestão da Informação
PPP	Projeto Político Pedagógico
SRMR	*Standardized Root Mean Square Residuals* / Raiz Padronizada do Resíduo Médio
TI	Tecnologia da Informação
TLI	Índice Tucker Lewis
UFPR	Universidade Federal do Paraná
USP	Universidade de São Paulo

SUMÁRIO

1 INTRODUÇÃO

A quantidade de dados gerada diariamente tem transformado todas as áreas da sociedade, tanto no âmbito pessoal quanto organizacional (FINZER, 2013, p. 1; GROSSI *et al.*, 2021; STARK; HAWAMDEH, 2018, p. 142). Dentre as estimativas sobre esse crescimento de dados, está a afirmação de que 90% de todos os dados existentes no mundo em 2013 tinham sido gerados nos dois anos anteriores (DAVENPORT *et al.*, 2015, p. 3; MARR, 2018; SINTEF, 2013) e que um carro autônomo registra cerca de 1GB de dados por segundo (HALL; PHAN; WHITSON, 2016, p. 2). Na *web*, o volume de dados criados, capturados, copiados e consumidos teve uma média de crescimento de 22,95% entre os anos 2012 e 2019, atingindo o ápice em 2020, com crescimento de 36,14%, decorrente das demandas emergências da pandemia de COVID-19, onde hábitos de trabalho, de aprendizagem e de entretenimento foram alterados (STATISTA, 2022). Se por um lado essa proliferação de dados resulta em uma sociedade progressivamente conectada, ela também se configura em um desafio multifacetado.

A propagação de dispositivos conectados à internet integra pessoas, processos e dados, ampliando os recursos analíticos a ponto de determinar quais organizações são relevantes (CISCO IT INSIGHTS, 2016). Independentemente do tamanho, toda empresa está repleta de informação, seja ela operacional, como vendas, transações, pesquisas, seja mecânica, como dispositivos médicos, *smartphones*, dentre outros mecanismos de monitoramento, ou social, como redes sociais (DAVENPORT *et al.*, 2015, p. 4). Com esta vasta quantidade de dados disponíveis, a exploração de suas potencialidades passa a ser, nas próximas décadas, mais do que uma preocupação com fins comerciais e se transforma em um diferencial competitivo (DONOHO, 2017, p. 748; Data Science for Business: What you need to know about data mining and data-analytic thinkingPROVOST; FAWCETT, 2013, p. 1).

Até mesmo organizações com uma cultura analítica já estabelecida precisam repensar suas práticas. Os avanços tecnológicos, bem como dos próprios métodos de análise, impõem um olhar mais profundo sobre os dados, sempre com o objetivo de encontrar formas inovadoras de melhorar a eficiência e a competitividade (HALL; PHAN; WHITSON, 2016, p. 1). Por isso, empresas com uma visão orientada a dados

têm modificado seus processos para refletirem o importante papel desses recursos e transformem massas de informação em ações para o desenvolvimento de competências, melhores experiências e novas oportunidades (CISCO IT INSIGHTS, 2016; DEMCHENKO *et al.*, 2016, p. 621).

Neste cenário, emerge uma disciplina que busca extrair conhecimento de dados brutos, por meio de modelos estatísticos, para auxiliar a tomada de decisão organizacional (BAŠKARADA; KORONIOS, 2017, p. 65; Data Science for Business: What you need to know about data mining and data-analytic thinkingPROVOST; FAWCETT, 2013, p. 2). Esta nova disciplina, a ciência de dados ou *data science*, revela algo de valor a partir de uma grande quantidade de dados, ajudando na escolha das melhores decisões, tanto para gestores quanto para cidadãos "comuns" (HAYES, 2017). Assim, com decisões mais eficientes, rápidas e baratas, o real poder dos dados é evidenciado, garantindo a vantagem competitiva buscada pelas organizações (DAVENPORT *et al.*, 2015, p. 5).

Com o reconhecimento do potencial dos dados, que resulta em uma maior difusão e democratização destes recursos, as organizações precisam de profissionais habilitados nesta área em diversos níveis, executando diferentes papéis (HALL; PHAN; WHITSON, 2016, p. 8). Mais especificamente, o desenvolvimento da Ciência de Dados amplia a demanda por um novo tipo de profissional, com formação técnica e conhecimento em tecnologias voltadas à análise de dados (DEMCHENKO *et al.*, 2016, p. 621). Este profissional, comumente chamado de cientista de dados, embora não possua uma definição bem estabelecida devido ao grande número de habilidades que lhe são atribuídas (DEMCHENKO *et al.*, 2016, p. 621), pode ser entendido como alguém que utiliza métodos científicos para descobrir e criar significado a partir de um conjunto de dados (DONOHO, 2017, p. 746).

Entretanto, a contratação de um cientista de dados ainda é uma tarefa árdua e dispendiosa, cuja retenção de talentos é complicada pela dificuldade em encontrar pessoas com habilidade científica e analítica, além de pensamento computacional (DAVENPORT; PATIL, 2012; REIS; SÁ, 2020). No mercado, há carência de pessoas suficientemente qualificadas para desenvolver projetos analíticos que envolvam técnicas complexas como *machine learning* ou para coordenar mudanças em direção a uma cultura organizacional orientada a dados (CUNHA, 2018; HALL; PHAN; WHITSON, 2016, p. 7).

Para amenizar isso, empresas mais estruturadas estabelecem estratégias próprias para suprir esta carência. Diversas organizações e universidades estabelecem parcerias como forma de desenvolver mais rapidamente novos talentos (HALL; PHAN; WHITSON, 2016). Por exemplo, a Cisco, empresa transnacional de soluções para redes e comunicações, criou programas educacionais internos em Ciência de Dados, estabeleceu parcerias com universidades para treinar seus talentos e construiu laboratórios avançados para ampliar sua capacidade analítica organizacional (CISCO IT INSIGHTS, 2016). Contudo, os cientistas de dados são requeridos também pelas empresas de pequeno porte, especialmente por aquelas que procuram abordagens inovadoras e eficientes em relação à sua capacidade analítica (HARRIS; MURPHY; VAISMAN, 2013-).

Essa propagação da análise de dados é uma tendência global que transforma o ambiente de trabalho. Desta forma, a formação dos profissionais que pretendem satisfazer as demandas do mercado deve acompanhar estas mudanças (DAVENPORT *et al.*, 2015; FINZER, 2013, p. 1). Portanto, se o aumento na geração no volume de dados é impressionante, a discrepância entre a necessidade de profissionais "alfabetizados" em dados e a formação destes especialistas é igualmente significativa e preocupante (FINZER, 2013, p. 1).

Por ainda ser um campo emergente, muitos cientistas de dados advêm de cursos universitários já estabelecidos, como estatística, ciência da computação, economia, dentre outros (BAŠKARADA; KORONIOS, 2017). Como resultado, tem-se o campo da ciência de dados ocupado por profissionais não qualificados que se autodenominam "cientistas de dados" em busca de pretensos ganhos salariais e *status* (WALKER, 2015, p. 10). Assim, é evidenciada a necessidade de se repensar modelos tradicionais de educação, bem como dos próprios cursos existentes, para englobar os aspectos multidisciplinares da Ciência de Dados (DEMCHENKO *et al.*, 2016).

Desde que Cleveland (2001) definiu formalmente o conceito de *data science*, o autor já ressaltava a necessidade de mudança do ensino na área. Nos últimos anos, a análise de dados, as habilidades em gestão da informação e em linguagens de programação formam a base das competências para um cientista de dados, que está mais valorizado que nunca (DAVENPORT *et al.*, 2015, p. 19). Com essa valorização, muitas universidades, incluindo a Universidade da Califórnia, Universidade de Nova

Iorque, MIT e Universidade de Michigan, adotaram iniciativas em programas de Ciências de Dados (DONOHO, 2017, p. 745).

Ainda assim, as Instituições de Educação Superior (IES) não estão conseguindo atender satisfatoriamente as demandas dos graduandos e, consequentemente, do mercado (BONNELL; OGIHARA; YESHA, 2022, p. 64; DONOHO, 2017, p. 748). Uma vez que os currículos dos cursos de ciência de dados são constituídos a partir de cursos existentes, que não abrangem todo o conjunto de competências e conhecimentos relacionados àNovamente área, a formação dos profissionais apresenta lacunas que impedem uma perfeita integração ao ambiente real de trabalho (BONNELL; OGIHARA; YESHA, 2022, p. 63; DEMCHENKO *et al.*, 2016, p. 620).

Diante deste panorama, a presente pesquisa pretende explorar o contexto educacional e profissional da Ciência de Dados no Brasil, a fim de descobrir as competências necessárias para o cientista de dados que atua no país. Ao final, além da contribuição para a comunidade acadêmica, espera-se, a partir dos resultados obtidos por meio da efetivação dos procedimentos metodológicos adotados, proporcionar uma visão detalhada e múltipla acerca do cientista de dados no Brasil, com foco nos profissionais da área.

1.1 PROBLEMA DE PESQUISA

Diante do aumento no volume e na variedade de dados, que impossibilita a análise “manual”, a contratação de estatísticos, analistas e administradores de bancos de dados não é mais suficiente (Data Science for Business: What you need to know about data mining and data-analytic thinkingPROVOST; FAWCETT, 2013, p. 1). O avanço dos computadores, dos algoritmos e da ubiquidade das redes revela, além de possibilidades mais amplas e profundas de análises, um novo modelo de trabalho (BRANDT, 2016, p. 2; Data Science and its Relationship to Big Data and Data-Driven Decision MakingPROVOST; FAWCETT, 2013, p. 51). Assim, tem-se o surgimento de uma nova formação cujas responsabilidades estão majoritariamente orientadas ao trabalho com dados (CAO, 2019, p. 1). As características desse profissional, o cientista de dados, são marcadas por formação técnica e profundo

conhecimento em tecnologias focadas em dados, cujas habilidades são voltadas à melhoria dos processos organizacionais (DEMCHENKO *et al*., 2016, p. 621).

A profissionalização do cientista de dados é evidenciada pelo crescente número de cargos sistematicamente divididos, diversificados, predominantemente orientados a dados, e pelo profundo impacto causado por estes profissionais "nas pesquisas de dados, na inovação, na economia e na sociedade" (CAO, 2019, p. 1, tradução nossa). O cientista de dados é um especialista inerentemente interdisciplinar, capaz de lidar com todos os aspectos de um problema referente a dados, da coleta inicial às conclusões finais (LOUKIDES, 2012, p. 16). O advento da ciência de dados enfatiza a relevância deste profissional em muitas áreas sociais e econômicas (BRANDT, 2016, p. 2).

A demanda por profissionais nessa área leva à criação de cursos de formação, porém é necessário estabelecer as competências mínimas para a prática da profissão (WALKER, 2015, p. 10). Uma vez que a ciência de dados ainda está em uma fase inicial de desenvolvimento, onde, além de desafios e oportunidades, há muitas questões não esclarecidas e diversos pontos de vista (CAO, 2017, p. 2). Dentre estes temas obscuros, está a formação do cientista de dados, cujas competências carecem urgentemente de padronização e parametrização (CAO, 2019, p. 1).

Como agravante, a propagação de cursos livres de curta duração, em grande parte *online*, ou mesmo de instituições tradicionais, indica uma exploração da dita profissão "mais sexy do século XXI" (DAVENPORT; PATIL, 2012) e uma tentativa de capitalização do cargo "cientista de dados" a partir da confusão acerca da definição sobre ciência de dados (WALKER, 2015, p. 11). Como resultado, as organizações reclamam da disponibilidade limitada de cientistas de dados qualificados, capazes de colocar em prática suas estratégias em relação à exploração dos dados (CAO, 2019, p. 1). Ao passo que a ciência de dados se torna mais relevante, maior a pressão por mudanças e aperfeiçoamento nas instituições educacionais responsáveis pela formação dos profissionais da área (FINZER, 2013, p. 8).

Considerando as questões aqui levantadas, apresenta-se a questão que guiará a presente pesquisa:

QUAIS AS COMPETÊNCIAS NECESSÁRIAS PARA UM CIENTISTA DE DADOS ATUAR NO BRASIL?

Considera-se que ao responder à pergunta da pesquisa, será possível formular um modelo de competências em Ciência de Dados, colaborando com organizações, instituições educacionais, profissionais e estudantes da área de Ciência de Dados.

1.2 OBJETIVOS

O objetivo geral, que busca responder à questão de pesquisa, é analisar as competências necessárias para a atuação de cientistas de dados no Brasil. Para que esse objetivo seja atingido, definiram-se os seguintes objetivos específicos:

a) investigar o perfil e as competências dos cientistas de dados atuantes no Brasil;
b) identificar os requisitos em anúncios de vagas de emprego para cientistas de dados;
c) identificar os tópicos abordados em cursos de nível superior e cursos livres para cientistas de dados;
d) comparar os requisitos demandados pelo mercado brasileiro com o conteúdo abordado na educação formal e livre;
e) elaborar um modelo das competências necessárias à Ciência de Dados no Brasil.

1.3 JUSTIFICATIVA

A presente pesquisa surge durante a experiência pessoal do autor como aluno de doutorado do Programa Pós-Graduação em Gestão da Informação (PPGGI) da Universidade Federal do Paraná (UFPR). De disciplinas de inteligência artificial e análise quantitativa de dados a eventos acadêmicos, passando por grupos de estudos da comunidade local, a *buzzword* “data science” sempre fez parte das aulas e discussões com colegas. Mas o que realmente corresponde a uma nova área de conhecimento e o que é apenas um modismo? Essa foi a pergunta que deu origem a esta tese.

Muito tem se falado sobre a quantidade de dados gerados pelas organizações, consequência do que essas produzem, entregam ou fazem, sejam produtos ou serviços, ou mesmo nas interações com seus consumidores (CHONG; CHANG, 2018, p. 202; HAYES, 2017; MARR, 2018; Data Science for Business: What you need to know about data mining and data-analytic thinkingPROVOST; FAWCETT, 2013, p. 1). Independentemente do porte destas organizações, é esperado que estas informações guiem as decisões necessárias. Logo, é compreensível que o reconhecimento do valor da ciência de dados para aquisição de vantagens competitivas (BURTCH WORKS, 2019, p. 7; DEMCHENKO *et al.*, 2016, p. 621; HALL; PHAN; WHITSON, 2016, p. 1) leve à valorização do cientista de dados.

Para ilustrar a relevância desse profissional, Power (2016) realizou, em novembro de 2015, pesquisas relacionadas a esta área de atuação no mecanismo de busca Google. A partir de três expressões (“data science jobs”, “data science” e “data scientist”), o autor verificou os números absolutos de páginas que incluíam conteúdo correlato e, utilizando a ferramenta Google Trends, constatou a tendência de crescimento no interesse sobre a temática. Para reforçar esse achado, na presente pesquisa, realizou-se a mesma busca em 23 de junho de 2020 e em 29 de abril de 2022, cujos resultados são apresentados na Tabela 1:

TABELA 1 – PÁGINAS INDEXADAS PELO GOOGLE RELACIONADAS À CIÊNCIA DE DADOS

	Nov./2015	**Jun./2020**	**↑ 2015/2020**	**Abr./2022**	**↑ 2020/2022**
“data science”	9.650.000	96.000.000	894,82%	4.490.000.000	4577,08%
“data scientist”	5.500.000	29.300.000	432,73%	72.100.000	146,08%
“data science jobs”	57.900	644.000	1012,26%	1.470.000	128,26%
"ciência de dados"	-	4.890.000	-	5.840.000	19,43%
"cientista de dados"	-	510.000	-	1.470.000	188,24%
"ciência de dados" "vaga"	-	211.000	-	283.000	34,12%

FONTE: O autor (2022), com base em Power (2016).

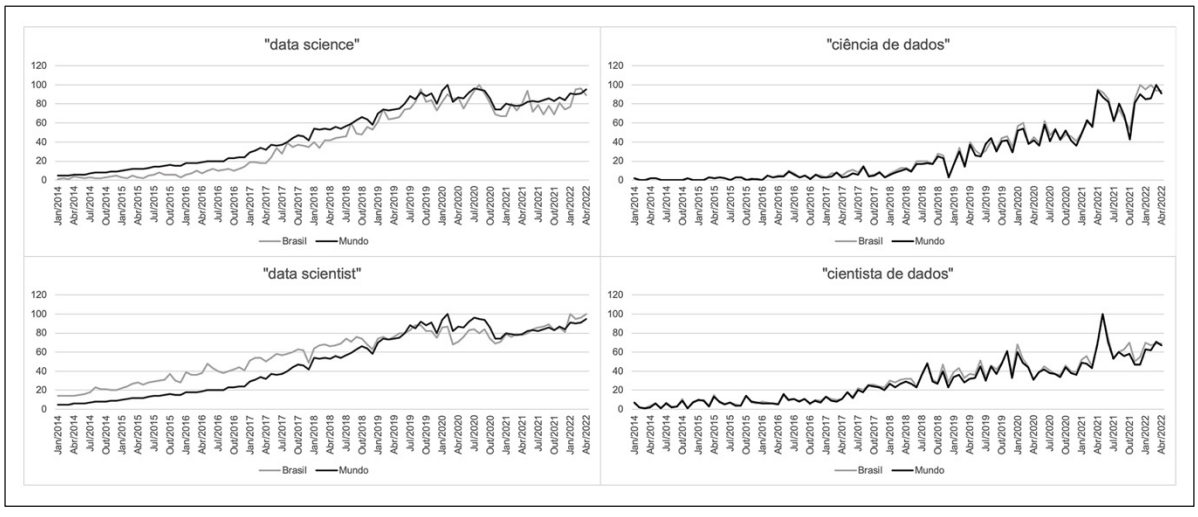
A busca realizada em 2020 confirmou o relatado crescimento, uma vez que os números absolutos referentes às três expressões aumentaram aproximadamente nove, cinco e onze vezes, respectivamente, entre 2015 e 2020. Nessa busca, também foram utilizadas as traduções das expressões originais que permitiram a mensuração dos termos para a língua portuguesa. Mesmo não sendo possível averiguar o crescimento relação aos dados de Power (2016), a relevância dos conceitos é

considerável, uma vez que a expressão com menos registros ("ciência de dados" "vaga") resultou em 211.000 páginas da web.

Na busca mais recente, mais uma vez, todas as expressões de busca demonstram crescimento durante o período de quase dois anos. A expressão principal, “data science”, apresentou uma quantidade de registros 46 vezes maior do que a encontrada em 2020, superando em muito o crescimento das demais expressões. Em contrapartida, a expressão equivalente em português, “ciência de dados”, foi a que demonstrou o menor crescimento (19,43%). A combinação de “ciência de dados” com o termo “vaga” apresentou o segundo menor crescimento (34,12%). Ainda assim, esse resultado indica que o mecanismo de busca utilizado indexou 72.000 novas páginas com conteúdo que relaciona a ciência de dados a vagas de emprego.

Replicando a utilização do Google Trends feita por Power (2016) para os termos “data science” e “data scientist”, bem como de suas respectivas traduções, confirma-se, para ambos, a tendência crescente de interesse. Os valores apresentados no Gráfico 1 foram coletados em 29 de abril de 2022 e demonstram o interesse pelas expressões durante os últimos sete anos, comparando os valores buscados no Brasil e por todo o mundo.

GRÁFICO 1 – INTERESSE POR CIÊNCIA DE DADOS NOS ÚLTIMOS CINCO ANOS



FONTE: O autor (2020), com base em Power (2016).

O Google Trends não apresenta números absolutos, sendo estipulado o valor referencial de 100 para a semana com o maior número de buscas para uma

determinada expressão. Assim, os demais meses recebem valores proporcionais a este ponto de maior interesse (GOOGLE INC., 2020). Mesmo o Google Trends não fornecendo valores absolutos, é possível identificar tendência de crescimento no interesse, em especial para os termos em inglês. A expressão “data science” atingiu seu ponto mais alto em relação ao buscador em fevereiro de 2020 para as buscas mundiais e agosto de 2020, para as buscas no Brasil. Já “data scientist” atingiu seu auge em janeiro de 2022 no mundo, resultado repetido em abril de 2022, e fevereiro de 2020 no Brasil. Dos termos em português, “ciência de dados” atingiu o pico em março de 2022 para as buscas mundiais e dezembro de 2021, valor repetido em fevereiro de 2022, para as nacionais. Por fim, “cientista de dados” atingiu em maio de 2021 o auge de buscas, para ambos os escopos de busca, nacional e internacional.

Esse interesse acerca da Ciência de Dados não é restrito ao público usuário do Google. Este tópico atrai o interesse de governos, empresários e acadêmicos e muitas iniciativas têm sido lançadas em diversos lugares, como Estados Unidos, China e União Europeia (CAO, 2019, p. 1). Muitas organizações constroem parcerias com universidades em busca de profissionais capacitados em manipulação e análise de dados (CISCO IT INSIGHTS, 2016; DAVENPORT *et al.*, 2015, p. 2; HALL; PHAN; WHITSON, 2016, p. 7). Universidades tradicionais, como a Universidade da Califórnia, Universidade de Nova Iorque e MIT têm se modificado para atender as demandas requeridas para a formação do profissional de Ciência de Dados (DONOHO, 2017, p. 745). Como exemplo, em setembro de 2015, a Universidade de Michigan anunciou um programa de 100 milhões de dólares, chamado *Data Science Initiative* que busca ampliar as oportunidades de pesquisas a estudantes e professores em toda a universidade (MICHIGAN INSTITUTE FOR DATA SCIENCE, 2015).

Esse fenômeno também vem se desenvolvendo no território brasileiro. Em pesquisa realizada no portal e-MEC[1], em 31 de maio de 2020, foram encontrados 26 cursos superiores cujo título continha a expressão “ciência de dados”. Porém, metade destes cursos estavam classificados como “Não iniciado”, ou seja, foram

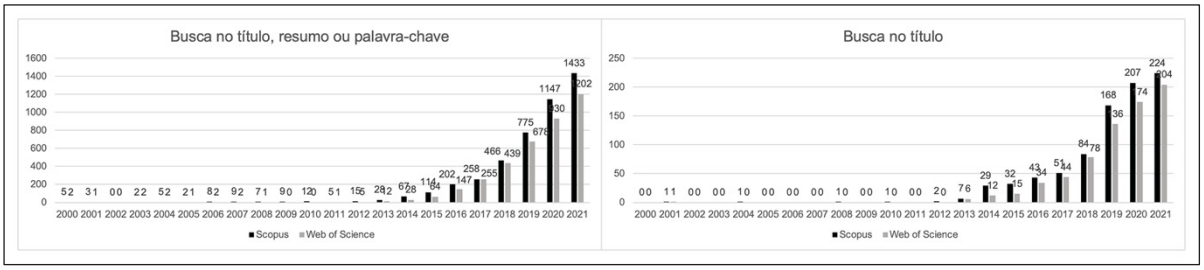
[1] Portal do Cadastro Nacional de Cursos e Instituições de Educação Superior que pode ser acessado pelo endereço eletrônico http://emec.mec.gov.br.

regularmente autorizados pelo MEC, mas não tiveram suas aulas iniciadas. Dos 13 cursos em atividade, oito eram de nível tecnológico, enquanto os outros cinco, bacharelado. A mesma pesquisa foi repetida em 29 de abril de 2022, tendo retornado 69 cursos, representando um aumento de 165% em quase dois anos. Os dados mais recentes apontam que 41 cursos estão em atividades, dos quais, 25 são de grau tecnológico (aumento de 212,50%) e 16, bacharelado (aumento de 220,00%). Ou seja, segundo dados do MEC, de 2020 a 2022 o número de cursos superiores voltados à Ciência de Dados foi triplicado tanto, seja tecnológico, seja bacharelado.

O registro há mais tempo em atividade se refere ao antigo curso de Estatística do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP) cujo nome foi alterado pelo projeto político pedagógico (PPP) iniciado em 2020. Esse curso é um dos quatro ofertados por instituições públicas. As outras três instituições são: Universidade Federal do Ceará (UFC), Universidade Federal da Paraíba (UFPB) e Faculdade de Tecnologia de Santana de Parnaíba (FATEC-SPB). Verifica-se, ainda, que 1/4 dos cursos estabelecem explícita relação entre Ciência de Dados e inteligência artificial e, praticamente, metade da oferta é para ensino à distância (21 cursos à distância, 20 cursos presenciais).

Para verificar se o interesse do mercado está refletido nas publicações científicas, foram realizadas, em 1º de maio de 2022, buscas nos repositórios Scopus e Web of Science. Para cada um, foram feitas buscas de artigos científicos publicados entre 2000 e 2021 (últimos 22 anos completos) que contivessem a expressão “data science” no título, no resumo ou palavra-chave. Posteriormente, essa busca foi restrita ao campo título, trazendo resultados mais específicos. Os números obtidos são demonstrados pelo Gráfico 2.

GRÁFICO 2 – NÚMERO DE ARTIGOS CIENTÍFICOS PUBLICADOS



FONTE: O autor (2020).

Em ambos os repositórios, as publicações relativas à Ciência de Dados se intensificaram nos últimos anos, tanto para os artigos com a expressão no título, quanto no resumo e palavras-chave. Para os documentos com “data science” no título, os anos 2020 e 2021, com 207 e 224 artigos respectivamente, representaram 50,65% do total do Scopus e, com 174 e 204 artigos, representaram 53,69% do Web of Science. O último quadriênio pesquisado representou 80,26% do Scopus e 84,09% do Web of Science.

Todavia, para publicações em português, esta tendência, mais uma vez, não se confirmou. Em busca pela expressão “ciência de dados”, no repositório Scielo, no dia 1º de maio de 2022, sete artigos foram retornados. Foram encontrados dois artigos duplicados e outros dois não continham a expressão buscada, resultando em três documentos distintos. Desses, dois foram publicados em 2021, outro em 2022, e apenas um deles continha o termo buscado no metadado título. Já para a Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (Brapci), a mesma busca retornou 25 artigos publicados entre 2016 e 2022, sendo que 11 deles continham a expressão buscada no título.

No meio acadêmico, buscou-se identificar dissertações e teses focadas em Ciência de Dados. Ainda em maio de 2022, na Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), identificaram-se três dissertações do ano de 2019 e três teses, sendo uma de 2018 e duas de 2019. Mesmo assim, nenhuma das pesquisas se relaciona com a presente pesquisa, uma vez que o tema central não é a Ciência de Dados em si e tampouco aborda o cientista de dados.

Efetuando-se a mesma busca no Catálogo de Teses e Dissertações da CAPES, foram encontrados 379 resultados, sendo 237 trabalhos de mestrados acadêmicos, 46 de mestrados profissionais e 96 teses. Este resultado é muito maior do que o encontrado em julho de 2019, quando se obteve 26 dissertações de mestrados acadêmicos, cinco de mestrados profissionais e cinco teses. Um dos fatores que explica essa diferença é que este sistema da CAPES não permite restringir a busca ao campo “título” e considera diversos metadados para trazer os resultados, incluindo a área de concentração na qual o programa de pós-graduação está inserido. Assim, após filtro manual, foram encontradas 13 dissertações e sete teses que atendiam esse critério (incluir “ciência de dados” no título), mas que pouco

estão relacionadas à presente pesquisa. Todas as pesquisas identificadas adotavam a Ciência de Dados como ferramenta e não como objeto de estudo.

Nova busca, utilizando o termo “cientista de dados”, foi realizada em ambas as bases de teses e dissertações. Outra vez, a BDTD não retornou nenhum registro e a base da CAPES trouxe nove pesquisas. Dentre as pesquisas retornadas, a dissertação de mestrado de Fabiano Castello de Campos Pereira, cujo título é “Cientistas de dados: proposta de um modelo conceitual considerando sua definição, sua formação, suas habilidades e as ferramentas que utilizam”, da Universidade de São Paulo, está relacionada à presente pesquisa. Porém, o texto completo da pesquisa não está registrado oficialmente no site da universidade, encontrado apenas no site pessoal do autor (CASTELLO, 2021).

Para buscar por teses e dissertações de língua estrangeira, foi utilizada a base de dados Networked Digital Library of Theses and Dissertations (NDLTD). No dia 25 de julho de 2020, a busca pela expressão “data science” nos títulos das pesquisas retornou 66 registros, dos quais 39 eram da língua inglesa. Foram encontrados três trabalhos escritos em português, mas, ao se refinar os resultados por este critério, foi possível verificar que dois registros se referiam à mesma tese. Essa duplicidade é decorrente que o trabalho está cadastrado em duas fontes indexadas pela NDLTD, a PUC-Rio e a BDTD. Em atualização da busca, realizada em 01 de maio de 2022, foram encontradas 111 teses, sendo 51 escritas em inglês. O número de teses em português se manteve em três.

Ao se buscar pela expressão “data scientist”, em maio de 2022, foram encontradas cinco pesquisas. Dentre essas, há um registro que se refere a uma dissertação escrita no idioma português cujo título utiliza a expressão inglês para designar o cientista de dados. Todavia, mesmo atendendo aos critérios de busca, essa pesquisa, cujo título é “Perspectivas e metodologias de pesquisa da Comunicação Social no contexto da internet com o Big Data e da especialização Data Scientist”, foi desconsiderada, pois seu enfoque está sobre a Comunicação Social. Assim, três pesquisas foram consideradas com potencial de contribuição para a presente tese, conforme apresentadas no Quadro 1.

QUADRO 1 – TESES RELACIONADAS

Identificação	Título	Objetivo	URL

Philipp Soeren Brandt Tese, 2016 *Columbia University*	The emergence of the data science profession	Estudar a formação de um novo especialista, o cientista de dados, e como este profissional se tornou relevante na sociedade.	https://doi.org/10.7916/D8BK1CKJ
Dana M. Dedge Parks Tese, 2017 *University of South Florida*	Defining Data Science and Data Scientist	Articular uma convergência sobre os conceitos de ciência e cientista de dados por meio de revisão da literatura.	http://pqdtopen.proquest.com/#viewpdf?dispub=10639701
Angel Krystina Washington Durr Tese, 2018 *University of North Texas*	A Text Analysis of Data Science Career Opportunities and U.S. iSchool Curriculum	Explorar a educação dos profissionais da iSchool dos EUA e a comparar com as funções presentes em anúncios de vagas para cientistas de dados.	https://digital.library.unt.edu/ark:/67531/metadc1404565/

FONTE: O autor (2020).

As três teses de doutorado são oriundas de universidades americanas e defendidas entres os anos de 2016 e 2018. Separadamente, todas abordam questões aqui citadas: a formação profissional do cientista de dados, as indefinições acerca dos conceitos de Ciência de Dados e cientista de dados e uma análise textual relacionada a anúncios de mercado para os especialistas da área.

Por fim, mesmo que nenhum trabalho de conclusão desenvolvido até o momento no PPGGI aborde a Ciência de Dados como objeto de estudo, esta pesquisa se mostra aderente à proposta do programa. Em síntese, a Ciência de Dados e, consequentemente, a atuação do cientista de dados é a extração de conhecimento útil a partir de quantidades massivas de dados. Assim, todos os elementos envolvidos nesse processo estão relacionados à gestão da informação e despertam interesse de diversos segmentos, sejam eles educacionais, governamentais, negócios, serviços ou indústrias. Ademais, o cientista de dados é essencialmente um profissional interdisciplinar cujos fundamentos estão na estatística, na computação e no conhecimento sobre o domínio, mas que depende de habilidades em comunicação, gestão e sociologia, por exemplo (CAO, 2017; DEMCHENKO *et al*., 2016; HALL; PHAN; WHITSON, 2016).

Diante deste panorama, é possível afirmar que a tese se justifica sob os mais distintos aspectos. A motivação pessoal, despertada pela experiência profissional e pelo interesse do mercado, vai ao encontro do aumento no número de publicações científicas sobre o tema, intensificado nos últimos cinco anos. Por outro lado, é possível identificar uma lacuna ainda não explorada pela comunidade acadêmica, especialmente nas pesquisas em língua portuguesa. Finalmente, a tese se mostra

alinhada aos objetivos e interesses do PPGGI, ainda que não haja trabalhos precedentes diretamente relacionados.

1.4 ASPECTOS DE NÃO-TRIVIALIDADE DA TESE

Como apresentado, a Ciência de Dados ainda é uma área em desenvolvimento e, consequentemente, possui uma série de características não esclarecidas. A falta de entendimento e confusões terminológicas proporcionam prejuízos a todos os envolvidos, empresas, instituições de ensino e indivíduos. Por isso, a formalização de suas competências é uma urgência para o desenvolvimento deste mercado, promissor e necessário (CAO, 2019; EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS)DEMCHENKO; BELLOUM; WIKTORSKI, 2017; SALTZ; GRADY, 2017).

Por outro lado, mesmo diante do interesse acerca do tema, verificou-se que as publicações em língua portuguesa não estão explorando essa lacuna de pesquisa. Este fato é refletido que apenas um artigo escrito em português foi utilizado nesta pesquisa. Ainda assim, a investigação de Curty e Serafim (2016) tem como objeto de estudo profissionais e instituições educacionais dos Estados Unidos. Considerando-se, ainda, a ausência de teses e dissertações realizadas em território brasileiro, fica evidenciado o ineditismo da presente pesquisa para o meio acadêmico nacional.

Sob a perspectiva metodológica, e pelas buscas realizadas, afirma-se que a pesquisa é inédita também em território internacional. A comparação entre os resultados de mineração de texto aplicada a anúncios de vagas de emprego e a currículos escolares foi identificada na tese de Durr (2018), que também adotou a técnica de análise de tópicos. Porém, a combinação deste tipo de mineração de texto, com técnicas de agrupamento e estendida aos resultados de uma pesquisa de levantamento não foi verificada em nenhum contexto.

Diante destes fatores, defende-se a tese de que as particularidades da Ciência de Dados requerem um modelo de competências específico para o cenário brasileiro. Desta maneira, nota-se que a presente investigação proporciona, além de benefícios práticos, contribuição para a literatura da área temática em questão.

1.5 DELIMITAÇÕES DA PESQUISA

Ainda que um dos objetivos específicos da tese seja a proposta de um modelo de competências para a Ciência de Dados, esta pesquisa não busca se aprofundar na discussão histórica acerca do conceito de competência, bem como de suas correntes teóricas. Sendo assim, a opção é por expor uma definição, até certo ponto consensual, sobre competência e adotá-la para identificar os conhecimentos, habilidades e atitudes dos profissionais analisados.

Outra delimitação se refere aos cursos de formação em Ciência de Dados que serão analisados. A educação à distância possibilita o acesso a inúmeros cursos estrangeiros, desde cursos gratuitos e de curta duração a mestrados em instituições renomadas. Todavia, uma vez que o foco da pesquisa é o contexto brasileiro, incluindo não só a atuação do cientista, mas também os cursos de formação deste profissional, considera-se que as análises sobre as iniciativas educacionais devam ser direcionadas às produções nacionais.

1.6 ESTRUTURA DA TESE

Em princípio, esta tese está estruturada em oito capítulos. O primeiro deles contém a introdução ao tema de estudo e apresenta o problema em questão, formalizando a pergunta de pesquisa. Na sequência, são definidos o objetivo geral e os cinco objetivos específicos, sucedidos pela seção de justificativa. Dentre as motivações da tese, estão questões pessoais, profissionais, acadêmicas, mas também são apresentadas justificativas para o Programa de Pós-Graduação em Gestão da Informação e para o mercado de Ciência de Dados. Fechando a primeira seção, são expostos os aspectos de não-trivialidade, remetendo ao ineditismo da pesquisa, e apresentadas as delimitações do estudo.

O segundo capítulo, denominado Ciência de Dados, engloba uma discussão conceitual sobre a área que começa com um levantamento histórico e termina com as abordagens contemporâneas dos teóricos da área. Neste capítulo, também são expostas representações visuais das diferentes visões da composição da Ciência de Dados. Ao final das representações, é apresentada a definição de Ciência de Dados adotada nesta tese, formulada com base na fundamentação teórica realizada. Em

seguida, há duas seções destinadas a relacionar a Ciência de Dados com o conceito de ciência e com o conceito de interdisciplinaridade. Por fim, a última parte do segundo capítulo traz metodologias e disciplinas relacionadas à Ciência de Dados, estabelecendo as diferenças entre termos que são comumente confundidos.

O terceiro capítulo apresenta um enfoque na profissão do cientista de dados. Inicialmente, são apresentadas definições e características deste profissional, bem como do mercado em que atua. Em seguida, após a apresentação do conceito de competência, são abordadas pesquisas que buscam investigar as competências e os papéis desempenhados pelo cientista de dados. O capítulo é concluído com a exposição de propostas e recomendações para programas educacionais na área de Ciência de Dados, provendo uma visão cronológica das discussões sobre a formação profissional do cientista de dados.

No capítulo seguinte, são apresentados os trabalhos relacionados à presente tese. Na primeira parte do capítulo, são descritas as três teses citadas na seção de justificativa. São expostos os objetivos, procedimentos metodológicos e principais resultados. Em seguida, são detalhadas seis pesquisas de levantamento que buscam uma maior compreensão sobre a Ciência de Dados e sobre os profissionais que atuam na área. Para cada pesquisa, são ressaltados os pontos que as tornam únicas e que servem de referência para a presente tese.

A seguir, o quinto capítulo apresenta os procedimentos metodológicos que serão adotados na pesquisa. Após a classificação do trabalho em relação à sua natureza, abordagem, objetivos e procedimentos técnicos, são descritas as unidades de análise da pesquisa, a amostra pretendida e os documentos que comporão o *corpus* de análise. Na sequência, após o detalhamento das variáveis do instrumento de coleta de dados, é apresentado o protocolo de análise, tanto da abordagem quantitativa quanto da abordagem qualitativa.

O sexto capítulo traz a apresentação e a análise dos resultados, iniciando pela pesquisa de levantamento. Nesta seção, são exploradas as características do perfil dos respondentes e, na sequência, os níveis de proficiência em relação às competências para a Ciência de Dados. Em seguida, são expostos os resultados das análises dos anúncios de vagas de emprego por meio das técnicas de mineração definidas na metodologia. O mesmo protocolo é seguido para os cursos superiores e livres, correspondendo às duas seções que fecham o capítulo.

Depois, há um capítulo destinado ao cruzamento entre os resultados obtidos por todos os procedimentos de análise definidos para a tese. Aqui, são retomados os principais achados, identificando pontos em comum e as divergências entre as unidades de análise. Por fim, neste capítulo, é apresentada a proposta de modelo de competências para a Ciência de Dados, consequência dos processos prévios de análise. Por fim, o último capítulo traz as conclusões, contribuições e limitações da pesquisa, assim como sugestões para pesquisas futuras.

2 CIÊNCIA DE DADOS

Os benefícios e a popularização da Ciência de Dados só se concretizaram por conta de dois fenômenos ocorridos, principalmente, a partir dos anos 2000. O primeiro é o montante de dados e informações que se expande progressivamente (GROSSI *et al.*, 2021). As pessoas estão mais conectadas, deixando rastros de dados por onde passam, visto que suas aplicações móveis fazem uso de informações de geolocalização, texto, vídeo e áudios, todos dados passíveis de serem minerados (LOUKIDES, 2012, p. 5–6). Os dados são gerados a cada ação tomada em todo dispositivo conectado à internet (DAVENPORT *et al.*, 2015, p. 3), situação que tende a se intensificar com o advento da Internet das Coisas (IoT) e com a entrada dos *Millennials*, indivíduos nativos digitais, no mercado de trabalho. Grandes montantes de dados, complexos e desordenados, contendo sujeiras, ruídos, além de desestruturação, estão disponíveis em uma infinita rede computacional (ANTONS *et al.*, 2020; HALL; PHAN; WHITSON, 2016, p. 2).

Por outro lado, as tecnologias de processamentos avançaram de um modo nunca visto. Poder de processamento, redes em todos os lugares, computação em nuvem, novos algoritmos de análise de dados, inteligência artificial, bem como a própria proliferação de dispositivos móveis compõem o segundo fenômeno que, por meio da conexão de múltiplas fontes de dados, possibilita a expansão dos princípios da Ciência de Dados e de técnicas de mineração (DAVENPORT *et al.*, 2015, p. 3; Data Science and its Relationship to Big Data and Data-Driven Decision MakingPROVOST; FAWCETT, 2013, p. 52). Ao mesmo tempo, essas tecnologias se tornaram menos caras e, consequentemente, mais acessíveis. Como resultado, encontram-se soluções tecnológicas robustas, com recurso de processamento paralelo, aptas a processar grandes bancos de dados com capacidade de gerar soluções mais sofisticadas, como problemas de *deep learning*, por exemplo (BRANDT, 2016, p. 85; HALL; PHAN; WHITSON, 2016, p. 6; PARKS, 2017, p. 32).

Para uma melhor compreensão deste fenômeno, as próximas seções são dedicadas à exploração do conceito de Ciência de Dados. Primeiramente, há uma discussão conceitual que busca identificar pontos de convergência e distinções entre as definições propostas pelos autores consultados. Na sequência, são apresentadas representações visuais da Ciência de Dados que, além dos fins didáticos, são

adotadas para demonstrar as interações e as disciplinas envolvidas pela área. Em seguida, há uma reflexão sobre a relação entre ciência e Ciência de Dados que é sucedida por uma seção destinada a explanar conceitos correlatos de acordo com a perspectiva adotada na presente pesquisa.

2.1 DISCUSSÃO CONCEITUAL

Em sua tese, Brandt (2016) utiliza a Wikipedia para apresentar o conceito de Ciência de Dados. Esta estratégia se baseia nos argumentos de Giles (2005) que afirma que, para termos estáveis, a Wikipedia apresenta acurácia equivalente a outras enciclopédias, com a vantagem de ter seus erros corrigidos mais rapidamente, uma vez que é construída coletiva e continuamente. Por outro lado, justamente por esta característica, a definição encontrada para Ciência de Dados passou por alterações desde a pesquisa de Brand (2016) e na versão de 2020, de língua inglesa, o conceito é apresentado da seguinte maneira:

> A Ciência de Dados é um campo interdisciplinar que emprega métodos científicos, processos, algoritmos e sistemas para extrair conhecimento e *insights* de grandes quantidades de dados, estruturados e não-estruturados. A Ciência de Dados está relacionada à mineração de dados, *machine learning* e *big data* (Data ScienceWIKIPEDIA.ORG, 2020, tradução nossa).

Adicionalmente, para fins de comparação, buscou-se pela definição em português e o conteúdo encontrado foi:

> A Ciência de Dados é uma área interdisciplinar voltada para o estudo e a análise de dados econômicos, financeiros e sociais, estruturados e não-estruturados, que visa a extração de conhecimento, detecção de padrões e/ou obtenção de variáveis para possíveis tomadas de decisão (Ciência de DadosWIKIPEDIA.ORG, 2020).

Inicialmente, destaca-se que ambas as definições contêm conceitos correlatos que carecem de maior aprofundamento, como mineração de dados, *machine learning*, reconhecimento de padrões, dentre outros que serão explorados adiante (Seção 2.5.3). Entre os pontos comuns, sobressai-se o caráter interdisciplinar da Ciência de Dados e a descoberta de conhecimento a partir dos dados, sejam eles estruturados ou não-estruturados. Todavia, enquanto a versão inglesa enfatiza o

caráter científico e técnico, destacando termos como “método científico”, “processos”, “algoritmos”, “mineração de dados”, “*machine learning*” e “*big data*”, a versão em português se volta a uma abordagem gerencial, frisando a natureza dos dados (econômicos, financeiros e sociais), culminando no conceito de tomada de decisão.

A primeira ocorrência encontrada na literatura da expressão *Data Science* data de 1974, utilizada por Naur em seu livro *Concise Survey of Computer Methods*. Após definir dado como “uma representação formal de fatos ou ideias com capacidade de ser comunicada ou manipulada por algum processamento” (NAUR, 1974, p. 30, tradução nossa), o autor define a Ciência de Dados como a ciência de “lidar” com os dados estabelecidos, porém a relação destes com aquilo que representam é delegada às suas áreas específicas. Uma abordagem ainda muito ligada à ideia de processamento de dados.

Mais próxima à concepção adotada na segunda década do século XX, a Ciência de Dados vem sendo discutida desde o final dos anos 1990 (INTERNATIONAL FEDERATION OF CLASSIFICATION SOCIETIES, 1996). Sob a alegação de que a estatística descritiva é apenas uma pequena parte do trabalho estatístico e que a educação da área deveria ser mais equilibrada, direcionada à ciência, focada em dados maiores e mais complexos, ligada a outras disciplinas, Wu (1997, p. 12) sugere que a estatística seja chamada de Ciência de Dados. Neste sentido, uma conceituação introdutória é elaborada por Hayashi (1998, p. 41) que afirma que a Ciência de Dados não é apenas um conceito que sintetiza a estatística e a análise de dados, envolvendo seus métodos relacionados. Mais especificamente, o autor alega que a Ciência de Dados está diretamente comprometida com os resultados, uma vez que procura entender um determinado fenômeno por meio da identificação de informações ocultas nos dados (HAYASHI, 1998, p. 41).

Três anos depois, Cleveland (2001) apresenta um plano para educação superior, focado na análise de dados, cujo objetivo é expandir as áreas técnicas da estatística. Com tantas e substanciais mudanças propostas, necessárias para a efetivação de potenciais benefícios das citadas áreas, o autor sugere a conformação de um novo campo, denominado *Data Science*. Finzer (2013, p. 2) aponta o trabalho de Cleveland (2001) como um dos primeiros usos do termo “ciência de dados” e destaca dentre seus pontos inovadores, a inclusão do ensino computacional focado em dados, além do aprofundamento em Pedagogia. Uma busca pela expressão “data

science", realizada na base de dados Scopus, no dia 24 de julho de 2020, confirma o ano de publicação do referido artigo como o mais antigo dentre os resultados obtidos.

Após quase duas décadas, apesar de sua popularidade, a Ciência de Dados ainda é um campo em estágio de desenvolvimento, onde os princípios ainda estão emergindo (POWER, 2016, p. 346; Data Science for Business: What you need to know about data mining and data-analytic thinkingPROVOST; FAWCETT, 2013, p. 16). Naturalmente, ainda há muita confusão e debate sobre uma definição que se aproxime de um consenso, tanto da Ciência de Dados, quanto do profissional chamado de cientista de dados (WALKER, 2015, p. 7). Se por um lado não há dúvidas sobre o potencial da área para uma cultura de valorização dos dados e para o desenvolvimento da economia, há questões não esclarecidas que ultrapassam as disciplinas da computação, informática e estatística, adentrando em campos como negócios, ciências sociais e até das ciências da saúde (CAO, 2017, p. 2).

Como parte de sua visão geral sobre o conceito de Ciência de Dados, Cao (2017) apresenta definições que sintetizam seu trabalho de revisão teórica. A primeira, apresentada em uma sentença de alto nível de abstração, define a Ciência de Dados como a "ciência dos dados" ou "ciência que estuda os dados" (ibidem, 2017, p. 8, tradução nossa). Já, sob uma perspectiva disciplinar, o autor define a Ciência de Dados como:

> um novo campo interdisciplinar que sintetiza e se baseia em estatística, informática, computação, comunicação, gestão e sociologia para estudar dados e seus ambientes (incluindo domínios e outros aspectos contextuais, como ambientes organizacionais e sociais) a fim de transformar dados em *insights* e decisões, seguindo um pensamento e metodologia do tipo "dado-para-conhecimento-para-sabedoria" (CAO, 2017, p. 8, tradução nossa).

Na sequência do trabalho, Cao (2017) apresenta essa definição pela fórmula:

$$\begin{aligned} Ciência\ de\ Dados \\ = Estatística + Informática + Computação + Comunicação \\ + Sociologia + Gestão|Dados + Ambiente + Mentalidade \end{aligned} \quad (1)$$

Onde "|" significa "condicionado a", ou seja, gestão concerne às questões relativas aos dados.

A definição e a equação elaboradas de Cao (2017) ilustram a dificuldade em estabelecer um consenso sobre esse conceito tão abrangente. A Ciência de Dados é um campo inter e multidisciplinar que engloba teorias e tecnologias de outras disciplinas, como matemática, estatística, ciência da computação e sistemas de informação (CHEN *et al.*, 2018, p. 171). Naturalmente, o desenvolvimento da Ciência de Dados impacta estas outras disciplinas já estabelecidas a ponto de gerar questionamentos sobre suas distinções e, até mesmo, sobre a necessidade de criação dessa nova área (DHAR, 2013; LOUKIDES, 2012; WU, 1997).

Com base no trabalho de Chen et al. (2018) e para expandir a percepção sobre o conceito de Ciência de Dados, exemplificando a diversidade que o cerca, no Quadro 2, são apresentadas as visões dos autores consultados durante a presente pesquisa:

QUADRO 2 – CONCEITOS DE CIÊNCIA DE DADOS

(continua)

Conceito	**Autores**
Ciência de dados não é apenas um conceito sintético para unificar estatística, análise de dados e métodos relacionados, mas também compreende seus resultados, procurando analisar e compreender um fenômeno com dados.	Hayashi (1998)
Ciência de dados é uma expansão substancial de áreas técnicas da estatística.	Cleveland (2001, p. 21)
A Ciência de Dados permite a criação de produtos de dados, isto é, aplicativos cujo valor advém de dados e, como resultado, produz mais dados.	Loukides (2012, p. 1)
Ciência de dados é envolvimento de princípios, processos e técnicas para compreender fenômenos por meio da análise (automatizada) de dados.	Provost e Fawcet (Data Science for Business: What you need to know about data mining and data-analytic thinking2013, p. 4)
Ciência de dados é o estudo da extração generalizável de conhecimento a partir de dados.	Dhar (2013, p. 64)
Ciência de dados é um campo inerentemente colaborativo e criativo, onde o profissional de sucesso pode trabalhar com administradores de banco de dados, empresários e outros com conjuntos de habilidades sobrepostos para concluir projetos de dados de maneiras inovadoras.	Harris, Murphy e Vaisman (2013-, p. 19)
De maneira geral, a Ciência de Dados é a aplicação de métodos quantitativos e qualitativos para resolver problemas relevantes e prever resultados.	Waller e Fawcett (2013, p. 78)
Ciência de dados é o estudo de onde as informações vêm, o que representam e como podem ser transformadas em um recurso de valor para a criação de estratégias de negócios e de tecnologia da informação.	Banafa (2014)

Ciência de dados é a extração de conhecimento que proporcione ações diretamente a partir dos dados por meio de um processo de descoberta ou formulação e teste de hipótese.	NIST Big Data Public Working Group (2015, p. 7)
Ciência de dados é o estudo científico da criação, validação e transformação de dados para criar significado.	Walker (2015, p. 8)

(conclusão)

Ciência de dados é um campo emergente da ciência, que requer uma abordagem multidisciplinar com uma forte ligação com *Big Data* e tecnologias orientadas a dados que criaram um efeito transformador para todos os domínios da pesquisa e da indústria.	Demchenko et al. (2016, p. 620)
Ciência de dados não é a ciência de dados. Em vez disso, o termo se refere a esforços para desenvolver uma análise de dados mais sofisticada e sistemática. [...] A Ciência de dados se posiciona como uma ciência aplicada que desenvolve soluções práticas de estatísticas, sistemas de informação e gerenciamento de dados.	Power (2016, p. 351)
Uma combinação entre descoberta científica e prática que envolve a coleta, gerenciamento, processamento, análise, visualização e interpretação de grandes quantidades de dados heterogêneos associados a uma ampla variedade de aplicações científicas, translacionais e interdisciplinares.	Donoho (2017, p. 745)
Ciência de dados é a extração de conhecimento, que proporciona ações, diretamente dos dados por meio de um processo de descoberta ou formulação e teste de hipótese.	PwC (2017, p. 3)
É um campo interdisciplinar que explora a metodologia científica e a tecnologia computacional sobre dados, incluindo gerenciamento de dados, acesso, análise e avaliação para benefício dos seres humanos.	Chen et al. (2018).
Um campo emergente que integra as definições de problemas, algoritmos e processos que podem ser usados para analisar dados de forma a extrair *insights* de (grandes) conjuntos de dados. Intimamente relacionado ao campo de mineração de dados, mas mais amplo em escopo e preocupação. Lida com (*big*) dados estruturados e não estruturados e abrange princípios de um conjunto de disciplinas, incluindo *machine learning*, estatística, ética e regulamentação de dados e computação de alto desempenho.	Kelleher e Tierney (2018)
De um ponto de vista cognitivo e metodológico, a Ciência de dados é um campo científico de alto nível, uma ciência transdisciplinar, um sistema complexo e um processo cognitivo e de descoberta abrangente.	Cao (2019, p. 3)
A ciência de dados é um conjunto interdisciplinar de habilidades encontradas na interseção de estatística, programação de computadores e conhecimento de domínio. Compreende três áreas distintas e sobrepostas: 1) Estatística, para modelar e resumir conjuntos de dados; 2) Informática, para projetar e usar algoritmos para armazenar, processar e visualizar dados; 3) Conhecimento de domínio, necessário para formular as perguntas certas e contextualizar as respostas.	IBM (2020, p. 3)
Ciência de dados é usar dados para fazer coisas úteis [...] envolve desde estatística e reconhecimento de padrões a análise de negócios e comunicação. Requer pensamento criativo tanto quanto pensamento analítico.	Kampakis (2020, p. 2–3)

FONTE: Desenvolvido com base em Chen et al. (2018).

Logicamente, a referência a dados está presente em praticamente todas as definições apresentadas. Pelo próprio nome, o propósito da Ciência de Dados é obtenção de informação ou conhecimento a partir de dados, fazendo desse recurso seu insumo essencial (CHEN *et al.*, 2018, p. 171). Essa extração de conhecimento de dados brutos é ressaltada pelas definições de Dhar (2013) e PwC (2017), mas está presente em outras pesquisas (BRANDT, 2016; DONOHO, 2017; PARKS, 2017).

Outra particularidade observada na maioria das definições é o caráter técnico e prático da Ciência de Dados. Começando pelo compromisso com o resultado, apontado por Hayashi (1998), passando pela criação de produtos de dados de Loukides (2012) e chegando à expressão "benefícios dos seres humanos" utilizada por Chen et al. (2018), a Ciência de Dados é vista como forma de auxiliar na tomada de decisões melhores, que podem ajudar as pessoas, as organizações e o poder público. Sua utilização permite o desenvolvimento de novas estratégias e políticas. Possibilita a exploração de novos modos e paradigmas, sejam eles organizacionais, educacionais, éticos, sociais, culturais, econômicos ou políticos (CAO, 2017). Como resultado, há uma melhor compreensão do ambiente para o desenvolvimento da sociedade.

Como mencionado, a Ciência de Dados utiliza recursos e conceitos de disciplinas tradicionais como matemática, estatística, ciência da computação e de domínios específicos (GRANVILLE, 2014, p. 84). Com o aumento da demanda e com o déficit de IES específicas, os profissionais costumam vir de campos tradicionais e estabilizados como banco de dados, pesquisas operacionais, *business intelligence*, ciências físicas e sociais, além das citadas anteriormente (HARRIS; MURPHY; VAISMAN, 2013-, p. 3).

Por fim, outro elemento recorrente no Quadro 2 é a característica interdisciplinar da Ciência de Dados. Essa necessidade de o cientista de dados atuar em distintas disciplinas é refletida na preocupação da formação desse profissional, desde Cleveland (2001), que sugeria que um quarto da carga horária da formação fosse destinado a investigações multidisciplinares, onde ocorreriam colaborações entre as áreas de estudo. Mais recentemente, pesquisas confirmam que a natureza interdisciplinar da Ciência de Dados continua sendo uma questão crucial nos cursos

superiores para os futuros profissionais da área (RAWLINGS-GOSS, 2019; TOSIC; BEESTON, 2018). Todavia, embora haja expectativas acerca disso, a interdisciplinaridade da Ciência de Dados não obriga os profissionais a serem especialistas em todas as disciplinas, mas que saibam se integrar a equipes interdisciplinares (BAŠKARADA; KORONIOS, 2017, p. 67; HARRIS; MURPHY; VAISMAN, 2013-).

Para decifrar estes conceitos, teóricos e praticantes propõem representações visuais da Ciência de Dados, onde, por meio de diagramas, ilustram as relações entre estas múltiplas disciplinas. Na próxima seção, modelos visuais são apresentados para um aprofundamento sobre a ideia do que é a Ciência de Dados.

2.2 REPRESENTAÇÃO VISUAL DA CIÊNCIA DE DADOS

Como uma maneira de estender as aplicações dos Círculos de Euler, Venn (1880) propõe novas formas de representação de conceitos. Por meio da composição de círculos, ou outras formas geométricas simples, Venn demonstra graficamente relacionamentos de inclusão, exclusão e interseção entre elementos, chamados de classes, que compõem um modelo visual das relações entre conceitos. Esse tipo de representação ficou conhecido como Diagrama de Venn e é uma das maneiras mais aceitas para se demonstrar graficamente um conceito (HAREL, 1988),

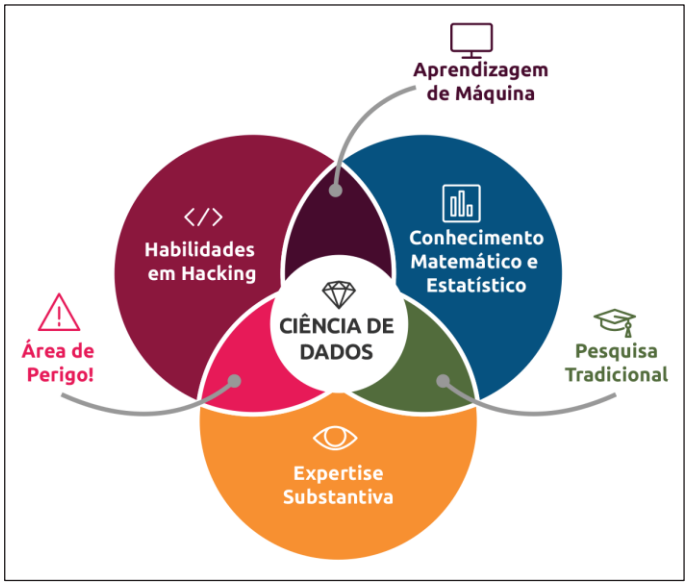
Em 2016, David Taylor, biotecnólogo e autor no portal KDnuggets[2], classifica a Ciência de Dados como um campo que ainda procura definição, embora muitos já tenham tentado realizar esta tarefa. Para Taylor (2016), em uma área repleta de profissionais da visualização, não poderiam faltar Diagramas de Venn que procurassem esclarecer o que é Ciência de Dados e realiza, então, um levantamento de representações gráficas que se encaixam nesse critério. Essa coleção de Taylor (2016) é utilizada como base para os diagramas que são aqui apresentados.

O primeiro diagrama foi proposto por Conway (2010) e deriva de sua dificuldade em lecionar Ciência de Dados, especialmente, em definir quais

[2] O KDnuggets (https://www.kdnuggets.com) é site líder e referência em Inteligência Artificial, *Big Data*, Mineração de Dados, Ciência de Dados e *Machine Learning*, tendo recebido numerosos prêmios e menções em diversos tipos de publicação (https://www.kdnuggets.com/about/index.html).

competências são necessárias para um cientista de dados “completo”. Para Conway (2010), o nome ”ciência de dados” acaba se tornando impróprio e contribui para não haver um acordo acerca de um currículo apropriado para a área. O autor argumenta ainda que é difícil distinguir entre *hackers*, estatísticos e especialistas da área, uma vez que a Ciência de Dados está onde as competências desses profissionais se sobrepõem. O modelo de Conway (2010) é apresentado na Figura 1.

FIGURA 1 – O DIAGRAMA DE VENN DA CIÊNCIA DE DADOS



FONTE: Adaptado de Conway (2010).

Composto por três disciplinas, o diagrama apresenta uma estrutura simples da formação da Ciência de Dados, sem ordem de hierarquia. A primeira apresentada pelo autor se refere às Habilidades em *Hacking*. Como os dados são recursos digitais, capacidade em manipular arquivos de texto por linha de comando, compreensão de estruturas de dados, pensamento lógico (algoritmos) são habilidades requeridas para um *data hacker*, mesmo não sendo necessária a formação em Ciência da Computação. Com a obtenção e limpeza dos dados, é necessário lançar mão de abordagens quantitativas para extrair conhecimento dos recursos adquiridos. O emprego apropriado de métodos matemáticos e estatísticos aos dados compõe a segunda disciplina do arranjo. Porém, a escolha mais adequada destes métodos só possível diante de um sólido conhecimento da área e do contexto aos quais os dados se referem. A este substancial conhecimento, Conway (2010) definiu como Expertise Substantiva.

As interseções do modelo também são descritas pelo autor. Segundo seu critério, o trabalho dos dados e dos métodos quantitativos, sem a devida expertise, criam um cenário de *machine learning*, mas a criação do conhecimento não é parte do arranjo. Em contrapartida, habilidades quantitativas e conhecimento da área se configuram em pesquisa tradicional, onde muitos pesquisadores dedicam, proporcionalmente, muito pouco tempo à aprendizagem sobre tecnologia. A terceira interseção, chamada de Área de Perigo, representa os profissionais que possuem conhecimento da área e técnica suficiente para trabalharem com dados, mas suas habilidades quantitativas são limitadas, podendo levar a análises equivocadas, com consequências negativas. Embora o diagrama de Conway, segundo palavras do próprio autor, não preze pela especificidade e tendo até negligenciado aspectos importantes do contexto, o mesmo é amplamente utilizado (BRANDT, 2016; FINZER, 2013; HARRIS; MURPHY; VAISMAN, 2013-), tornando-se referência a propostas posteriores.

Dois anos depois, Brendan Tierney (2012), autor de livros sobre Ciência de Dados e temas relacionados, propõe um diagrama para demonstrar a multidisciplinaridade da Ciência de Dados e o amplo conjunto de habilidades necessárias a um cientista de dados. Sua proposta é apresentada na Figura 2:

FIGURA 2 – A MULTIDISCIPLINARIDADE DA CIÊNCIA DE DADOS

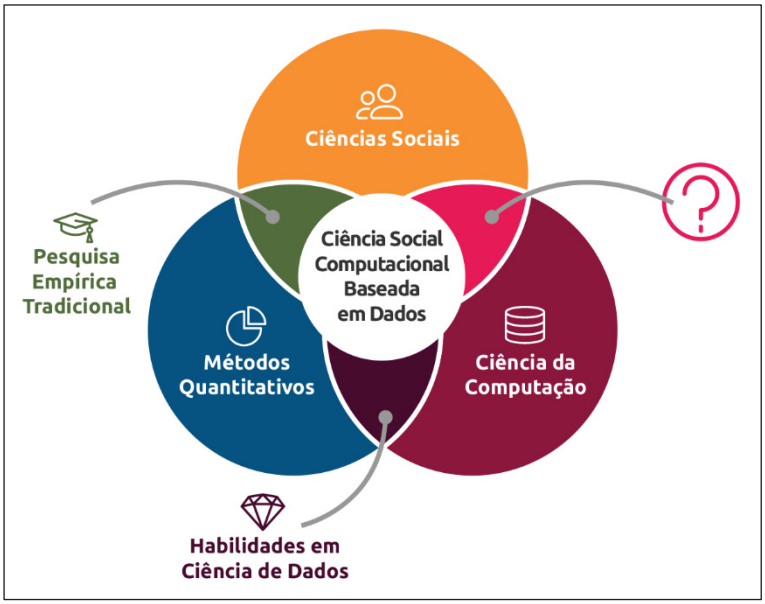
FONTE: Adaptado de Tierney (2012).

O círculo externo, disposto em um fluxo contínuo, representa as competências fundamentais para o profissional que deseja se tornar um cientista de dados. Na região interna, estão as habilidades nas quais a maioria dos profissionais são experientes em uma ou duas. Dentre as competências fundamentais, todas são abordadas por outros autores, como Conhecimento da área (KIM; LEE, 2016; LINDEN *et al.*, 2015, p. 4), Comunicação (CAO, 2017, p. 31; KIM; LEE, 2016, p. 167), Apresentação (DAVENPORT *et al.*, 2015, p. 18; SCHOENHERR; SPEIER-PERO, 2015, p. 129), Curiosidade (LINDEN *et al.*, 2018, p. 6; POWER, 2016, p. 353), Solução de problema (GRANVILLE, 2014, p. 87; RAWLINGS-GOSS, 2019, p. 7), Análise de negócio (DEMCHENKO *et al.*, 2019, p. 10; Data Science for Business: What you need to know about data mining and data-analytic thinkingPROVOST; FAWCETT, 2013, p. 20) e Estratégia de negócio (HAWAMDEH; CHANG, 2018, p. 12). O diagrama de Tierney (2012) não se sobressai pela simplicidade e destaca, além das sete competências fundamentais, nove habilidades mais direcionadas, incluindo Neurocomputação, de caráter bastante específico. De qualquer maneira, é uma demonstração da confusão característica da Ciência de Dados.

Voltando à simplicidade formal, Ulrich Matter (2013), pesquisador das áreas de Políticas Econômicas, Econometria e Ciência Social Computacional, faz uma releitura

do diagrama de Conway (2010) para representar o espaço da Ciência de Dados no ambiente puramente acadêmico e científico. Sua intenção é contrapor a perspectiva de que o desenvolvimento da Ciência de Dados se deve unicamente às demandas do mercado. O diagrama de Matter (2013), que parte da abordagem de que a Ciência de Dados não é uma única ciência, mas a aplicação de várias outras conforme apresentado na Figura 3.

FIGURA 3 – CIÊNCIA SOCIAL COMPUTACIONAL BASEADA EM DADOS

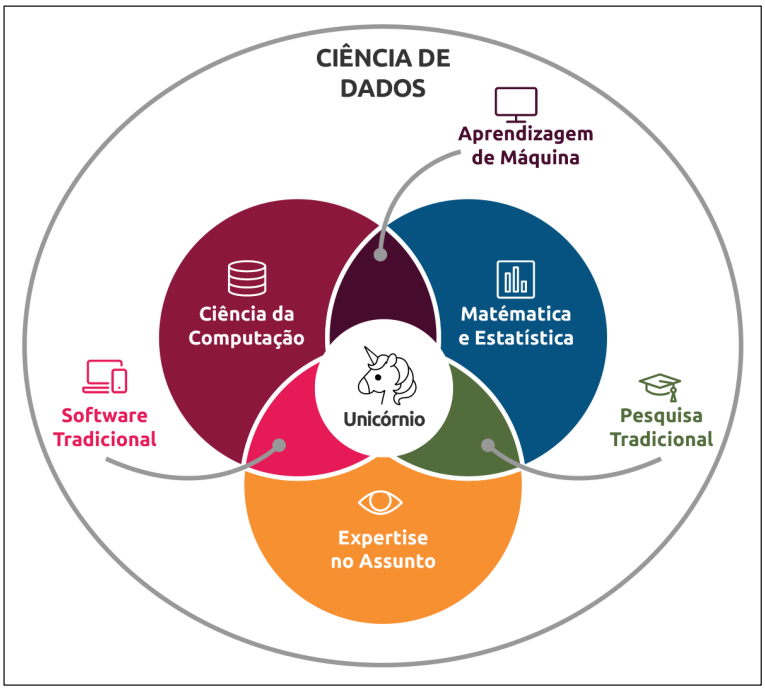


FONTE: Adaptado de Matter (2013).

Percebe-se que o diagrama original de Conway (2010) foi rotacionado e que as três disciplinas rebatizadas, adotando termos mais formais. A disciplina "Habilidades em *Hacking*" foi alterada para "Ciência da Computação", "Conhecimento Matemático e Estatístico", para "Métodos Quantitativos" e "Expertise Substantiva" ficou com o nome "Ciências Sociais". Matter (2013) chama de "Ciência Social Computacional" a aplicação de métodos numéricos ou simulações, baseadas em modelos, para pesquisar questões complexas das Ciências Sociais. Assim, ao centro do diagrama, o autor propõe uma Ciência Social Computacional voltada aos dados, formada pela interseção das habilidades em Ciência de Dados e Pesquisa Empírica Tradicional. A junção das Ciências Sociais à Ciência da Computação, que anteriormente correspondiam à Área de Perigo, forma uma área indefinida pelo autor.

Após quatro anos, o Diagrama de Venn da Ciência de Dados ganhou sua versão 2.0, proposta por Steven Geringer (2014), cientista de dados na IBM. O que motivou a segunda versão do diagrama foram as discussões que se iniciavam à época sobre o quão difícil era se encontrar um cientista de dados. Essa dificuldade era tamanha, que a tarefa de encontrar esse profissional passou, anedoticamente, a ser comparada à missão de se encontrar um unicórnio. A Figura 4 apresenta a estrutura de Geringer (2014):

FIGURA 4 – O DIAGRAMA DE VENN DA CIÊNCIA DE DADOS V2.0



FONTE: Adaptado de Geringer (2014).

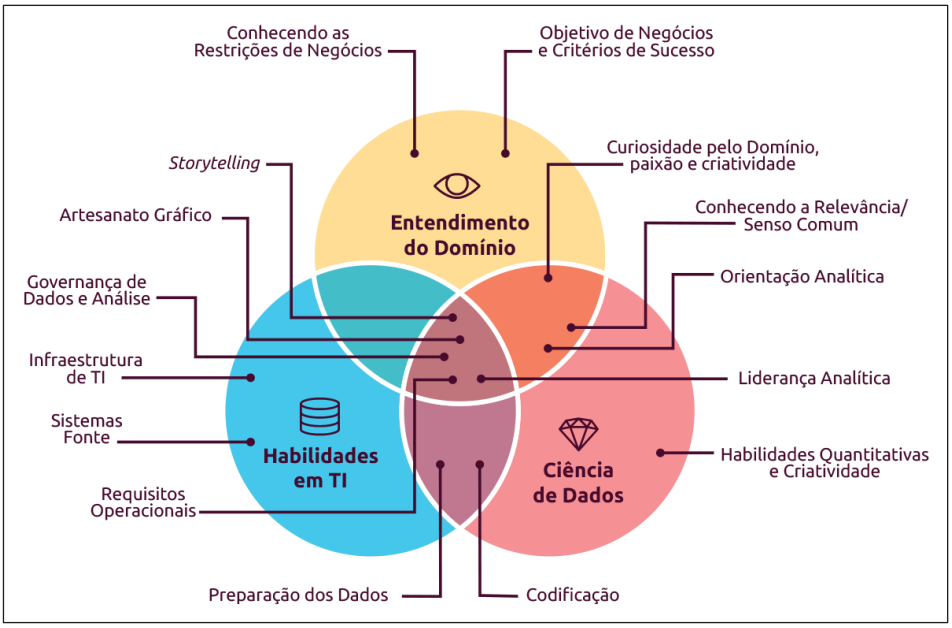
A primeira diferença em relação à primeira versão, talvez a menos significativa, é a formalização da disciplina “Ciência da Computação”, anteriormente chamada de “Habilidades em *Hacking*”. A interseção dessa disciplina à “Expertise no Assunto” se refere à área de desenvolvimento de software tradicional, extinguindo a “Área de Perigo” do diagrama original. Em segundo lugar, a Ciência de Dados deixa de ser a interseção das três disciplinas principais para ser representada como um elemento que engloba os demais. Por fim, a alteração mais relevante é a inserção do “Unicórnio” no centro do diagrama.

Geringer (2014) reconhece que as organizações orientadas a dados estão mudando suas práticas para a formação de equipes de dados com competências complementares no lugar de encontrarem indivíduos que sejam especialistas em

ciência da computação, estatística e, ainda, conheçam a área de atuação. Mesmo assim, pesquisas posteriores ao seu diagrama apontam a insistência na busca por indivíduos com conhecimentos avançados em todas as áreas da Ciência de Dados, reforçando a "síndrome do unicórnio" (BAŠKARADA; KORONIOS, 2017; HAYES, 2017; LINDEN *et al.*, 2018; STODDER, 2015).

Em 2015, uma das publicações da Gartner, empresa de atuação global em pesquisa e consultoria em TI, apresentou, por meio de Linden et al. (2015), um diagrama mapeando as tarefas mais amplas referentes às habilidades necessárias na montagem de equipes em Ciência de Dados. O resultado desse diagrama está na Figura 5 a seguir:

FIGURA 5 – MAPEAMENTO DE HABILIDADES NAS EQUIPES DE CIÊNCIA DE DADOS



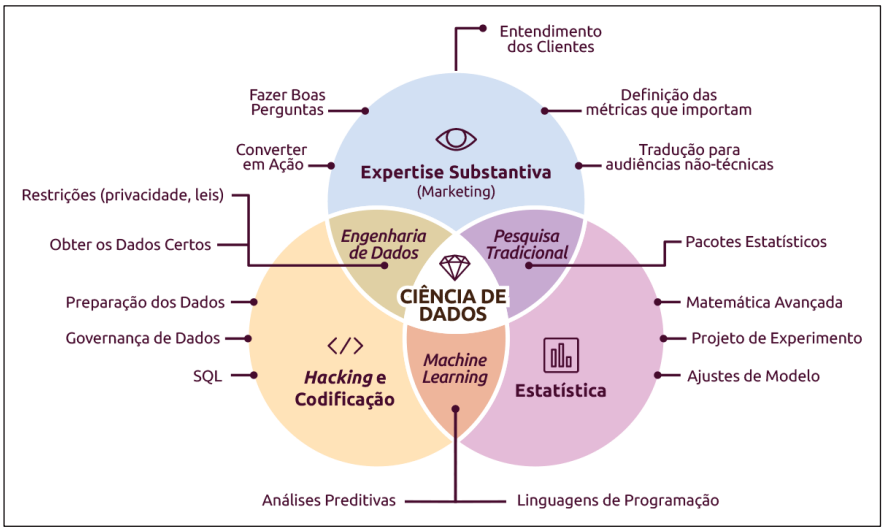
FONTE: Adaptado de Linden et al. (2015).

Novamente a estrutura é formada por três disciplinas principais. Porém, a Ciência de Dados deixa de ser representada pela interseção de duas ou mais áreas, tornando-se uma das disciplinas formadoras do tríplice arranjo. A disciplina Ciência de Dados contém as habilidades quantitativas e a criatividade para solucionar problemas. As outras duas disciplinas se referem às Habilidades em TI, que englobam a infraestrutura e os sistemas e; ao Entendimento do Domínio, que atende às restrições do negócio e aos objetivos e critérios de sucesso da organização.

Pelo modelo, a interseção da Ciência de Dados e do Entendimento do Domínio contempla a paixão e a curiosidade pela área de trabalho, a orientação de análise e o senso comum, conhecimento relevante. A Ciência de Dados e as Habilidades em Tecnologia da Informação englobam juntas a preparação dos dados e a codificação. Por fim, a união das três disciplinas contempla os requisitos operacionais, a liderança analítica, a governança analítica e de dados, o trabalho gráfico e o *storytelling*. O diagrama apresentado por Linden et al. (2015) tem como objetivo ser componente de um guia prático para montagem de equipe em Ciência de Dados e, por isso, diminui o grau de abstração, adentrando na descrição de habilidades e papéis dos profissionais de dados.

Neste sentido, Christi Eubanks (2016), vice-presidente de prática e diretora de dados na própria Gartner, proporciona outro mapeamento de habilidades relativas à Ciência de Dados, apresentado na Figura 6:

FIGURA 6 – HABILIDADES EM CIÊNCIA DE DADOS

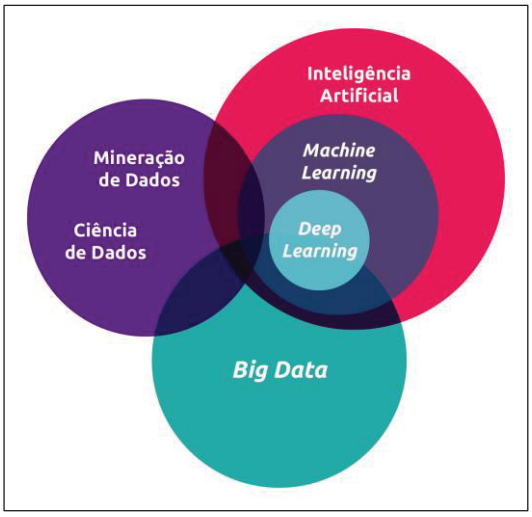


FONTE: Adaptado de Eubanks (2016).

Diferentemente do diagrama anterior, a proposta de Eubanks (2016) vai ao encontro do modelo de Conway (2010) e coloca a Ciência de Dados na interseção de três disciplinas: Expertise Substantiva, Estatística e *Hacking* e Codificação. Sem a área de perigo, as interseções são compostas pela Pesquisa Tradicional e *Machine Learning*, também presentes no primeiro diagrama, além da Engenharia de Dados, originada da junção entre habilidades em TI e conhecimento da área. Dentre as habilidades específicas citadas no modelo, destacam-se o cuidado com a privacidade e o atendimento à legislação, a busca pela compreensão dos consumidores e a preocupação em traduzir os resultados para um público não-técnico.

Outra representação relativa à Ciência de Dados é elaborada por Mayo (2016). Porém, nesta abordagem o foco não é o conceito de Ciência de Dados em si, mas sim sua relação com outras disciplinas que estão igualmente envolvidas na adoção de tecnologias a massivos volumes de dados. A estrutura levantada por Mayo (2016) é exposta na Figura 7:

FIGURA 7 – O QUEBRA-CABEÇA DA CIÊNCIA DE DADOS



FONTE: Adaptado de Mayo (2016)

Conforme Mayo (2016), a Ciência de Dados é um “quebra-cabeça” formado por seis conceitos-chave, um deles a própria Ciência de Dados. O autor afirma que, mesmo com o vasto conteúdo na *web* comparando e contrastando os itens da terminologia da Ciência de Dados, novos estudos são válidos, uma vez que terminologia se trata de um conceito fluido. Além disso, como neste caso não há unanimidade, expor seu posicionamento a terceiros é uma maneira de aprimorá-lo. Os conceitos apresentados na estrutura de Mayo (2016) são explorados nesta pesquisa durante a seção 2.5.4.

Logicamente, há outras propostas em relação à concepção da Ciência de Dados que não são aqui apresentadas. Com diferentes abordagens e níveis de profundidade, alguns desses diagramas não apresentados podem ser consultados na própria publicação de Taylor (2016). De qualquer forma, mesmo diante das distinções entre cada modelo, é observado um padrão entre eles. A maioria se baseia em três disciplinas principais que, embora possam adotar nomes diferentes, referem-se a Métodos Quantitativos, a Tecnologia da Informação e ao conhecimento do domínio.

Alinhados a Conway (2010), Harris, Murphy e Vaisman (2013-, p. 24) afirmam que experiência no desenvolvimento de software, pensamento estatístico e conhecimento da área são fundamentais para a efetividade da Ciência de Dados. Como demonstração dessa afirmação, instituições de educação superior têm modificado seus currículos para atenderem essa demanda, como cursos de estatística aplicada que incorporam aspectos da área de consultoria para proporcionarem aos seus estudantes, maior expertise do domínio de atuação. Nesta

direção, nos Estados Unidos, o Instituto Nacional de Padrões e Tecnologia (NIST) publicou um *framework* cujo tema principal era *big data* e, nele, as habilidades fundamentais para a Ciência de Dados eram Conhecimento de Domínio, conhecimento específico à área de atuação, estatística (e *Machine Learning*), métodos quantitativos e técnicas de análise, e engenharia, conhecimentos em TI (NIST BIG DATA PUBLIC WORKING GROUP, 2015, p. 9).

Esses diagramas da Ciência de Dados procuram apresentar uma visão geral sobre os elementos, disciplinas e habilidades, componentes da Ciência de Dados. Ainda que não haja consenso, ao menos formalizam ideias de investigação e reflexão acerca da área estudada. Nota-se, porém, que o diagrama original, proposto por Conway (2010), continua sendo relevante e, mesmo diante das propostas de melhorias, as versões posteriores reconhecem seu valor. Adicionalmente, é notório a recorrência de termos como habilidades em TI, *machine learning*, estatística, matemática, conhecimento do domínio e pesquisa. Este último conceito ora é denominado "pesquisa tradicional", ora "pesquisa empírica", mas sempre aludindo ao método científico.

A presente tese adota a perspectiva de que é das interações destes termos, potencializados pelas habilidades sociais, que se dá a Ciência de Dados. Assim, define-se a Ciência de Dados como a **interação interdisciplinar de competências socioculturais, métodos quantitativos, ciência da computação e conhecimento de domínio para a extração metodológica de conhecimento a partir de quantidades massivas de dados**.

Espera-se com esta definição evidenciar que elementos externos à tríplice Tecnologia, Estatística e Conhecimento de Domínio também são fundamentais para a Ciência de Dados. O processo de descoberta do conhecimento a partir dos dados é explicitado junto à expressão metodológica para formalizar que o método (científico) é parte essencial da estrutura estudada. Além disso, as competências socioculturais são fundamentais aos indivíduos e às organizações na promoção da comunicação, da colaboração, da liderança, do incentivo à criatividade e do trabalho em equipe. Todos estes elementos são especialmente vitais quando os envolvidos possuem formações e experiências advindas de diferentes disciplinas.

Durante o decorrer desta tese, espera-se fundamentar esta definição pela literatura pesquisada, pelos procedimentos metodológicos definidos e,

consequentemente, pelos resultados futuramente encontrados. A seguir, é retomada a fundamentação teórica pelo questionamento da relação entre a Ciência de Dados e a Ciência.

2.3 CIÊNCIA DE DADOS É CIÊNCIA?

É comum associar a ciência a conceitos como verdade e conhecimento, à busca por regras que possam explicar ou prever fenômenos e fatos (HEPBURN; ANDERSEN, 2021). Desde a antiguidade, povos como babilônios e egípcios produziam, movidos pela curiosidade, conhecimentos úteis pelos séculos seguintes (BAZZO, 2014, p. 141). Francis Bacon, tido como um dos precursores da ciência moderna, descreveu que a função primordial da ciência é melhorar a vida humana na Terra, o que só seria possível mediante coleta e análise de dados que pudessem fundamentar teorias (CHALMERS, 1993, p. 20).

Hepburn e Andersen (2021) afirmam que historicamente o principal produto da ciência é o conhecimento e, consequentemente, o objetivo da metodologia científica é descobrir os métodos mais adequados pelos quais este conhecimento é gerado. A maneira mais apropriada de buscar o conhecimento científico, ou a questão do "método científico", é tópico de reflexões há milênios (HEPBURN; ANDERSEN, 2021). Mas foi a partir do século XVII, quando a ciência adentra em uma nova era, que estas discussões sobre o tema se intensificaram, dando origem a um novo ramo filosófico: a filosofia da ciência (CHIBENI, 2004).

Chalmers (1993), em seu livro "O que é ciência afinal?", dedica-se a levantar e descrever as principais correntes ou abordagens de como o conhecimento científico é estabelecido, segundo a ciência moderna. Inicialmente, o autor apresenta afirmações que formam um senso comum acerca da ciência e tornam o conhecimento científico amplamente aceito. "Conhecimento científico é conhecimento provado", "As teorias científicas são derivadas de maneira rigorosa da obtenção dos dados da experiência adquiridos por observação e experimento", "Opiniões ou preferências pessoais e suposições especulativas não têm lugar na ciência", "A ciência é objetiva" são algumas das afirmações que resumem a concepção popular de conhecimento científico.

Em seguida, Chalmers (1993) apresenta duas explicações da ciência que considera equivocadas: indutivismo e falsificacionismo. No indutivismo, contando que determinadas condições sejam atendidas, é válido generalizar uma lei universal a partir de observações singulares. Ou seja, conforme descreve Vieira (2008, p. 112–113), no método indutivo, "você observa um conjunto de objetos, fatos, pessoas, e, com base no que observou, conclui para objetos, fatos, pessoas que não observou". Segundo este método, a ciência sempre se inicia com a observação que, por sua vez, proporciona a fundamentação necessária para que o conhecimento científico possa ser elaborado (CHALMERS, 1993). Por isso, os fenômenos precisam ser examinados repetidas vezes, em diferentes condições, sem qualquer interferência da pessoa do cientista, do tempo ou do espaço (VIEIRA, 2008). A ciência, segundo esta abordagem, avança pela somatória de conhecimentos adquiridos pelas experiências realizadas.

Contrariando estes preceitos do método indutivo, surge o falsificacionismo com Karl Popper. Este filósofo austríaco foi um crítico do papel atribuído à observação no empirismo e argumentava que qualquer observação é influenciada por expectativas, teorias ou mesmo pontos de vista (VIEIRA, 2008). Para Popper, a ciência começa com teorias, que podem ser suposições ou conjecturas especulativas, a fim de explicar aspectos do mundo, resolvendo problemas presentes em teorias anteriores. Chalmers (1993) descreve da seguinte maneira a evolução da ciência conforme o falsificacionismo:

> Uma vez propostas, as teorias especulativas devem ser rigorosa e inexoravelmente testadas por observação e experimento. Teorias que não resistem a testes de observação e experimentais devem ser eliminadas e substituídas por conjecturas especulativas ulteriores. A ciência progride por tentativa e erro, por conjecturas e refutações. Apenas as teorias mais adaptadas sobrevivem. Embora nunca se possa dizer legitimamente de uma teoria que ela é verdadeira, pode-se confiantemente dizer que ela é a melhor disponível, que é melhor do que qualquer coisa que veio antes (CHALMERS, 1993, p. 64)

Vieira (2008) salienta que nesta perspectiva, a ciência não acontece pelo método indutivo. Contrariamente, esse progresso ocorre de forma dedutiva, onde não se objetiva provar teorias, mas sim contestá-las (CLELAND, 2001). Assim, mesmo que jamais uma teoria seja devidamente comprovada como verdadeira, pode-se adotá-la como a melhor disponível, uma vez que nunca foi refutada.

Dentre outras perspectivas apresentadas por Chalmers (1993), está a abordagem das teorias como estruturas. A primeira proposta deste grupo que o autor apresenta corresponde aos programas de pesquisa de Imre Lakatos, estruturas que orientam as pesquisas em uma determinada área do conhecimento. Nestas estruturas, há um núcleo duro com as suposições basilares do programa, protegido de falsificação por hipóteses auxiliares. Em caso de refutação do programa, haverá um desenvolvimento suplementar ao "núcleo irredutível com suposições adicionais numa tentativa de explicar fenômenos previamente conhecidos e prever fenômenos novos" (CHALMERS, 1993, p. 113). Neste sentido, esta abordagem é superior às abordagens anteriores, pois uma teoria não é descartada mediante um resultado contraditório. Refutações equivocadas podem ocorrer por diversas razões, incluindo a limitação de tecnologia, com a simples falta de instrumentos adequados para mensurar o objeto de estudo (VIEIRA, 2008).

Outra abordagem da teoria como estrutura exposta por Chalmers (1993) é a ideia de paradigma de Thomas Kuhn. Nesta perspectiva, a progressão da ciência acontece pelas seguintes etapas:

pré-ciência → ciência normal → crise → revolução → nova ciência normal → nova crise

Assim, aquilo que precede a formação da ciência é visto como uma atividade diversa, dispersa e desorganizada que, quando dirigida e estruturada, constitui um paradigma destinado a uma comunidade científica. Deste modo, um paradigma é "composto de suposições teóricas gerais e de leis e técnicas para a sua aplicação adotadas por uma comunidade científica específica" (CHALMERS, 1993, p. 124). Para Kuhn, em um determinado campo da ciência, é o paradigma que define os padrões e a forma de se pesquisar, sendo que uma ciência é considerada madura quando é regida por um único paradigma. Além disso, a revolução da ciência ocorre quando um paradigma é abandonado pela comunidade científica e substituído por um novo.

Depois de apresentar o histórico das metodologias científicas, que inclui até mesmo a teoria anarquista do conhecimento de Feyerabend, Chalmers (1993) encara a pergunta que intitula seu livro: O que é ciência afinal? O autor assume que sua pergunta é "enganosa e arrogante" (ibid., p. 211), pois considera que não haja um

conceito universal e atemporal sobre o que é a ciência ou mesmo o método científico. Adicionalmente, é impossível determinar uma única concepção do que é ciência, principalmente se considerada a diversidade de áreas como Física, Biologia, História, Sociologia, dentre outras (CHALMERS, 1993).

Mesmo assim, conforme relatado anteriormente, este tema é tópico de diversos autores, tanto filósofos quanto pesquisadores de outras áreas do conhecimento. Dentre estes filósofos, estão Bachelard (1996) e Morin (2002, 2005) que tratam de uma questão até então ignorada por Popper, Kuhn, Feyerabend, Lakatos, dentre outros, que é a complexidade científica. Para Morin (2002, 2005, p. 13), o paradigma da complexidade, que não se resume ao conceito de complicação, envolve "acontecimentos, ações, interações, retroações, determinações, acasos, que constituem nosso mundo fenomênico", tornando-se uma exigência social e política do século XXI.

Nesta perspectiva, uma ciência exclusivamente objetiva implica em um pensamento mutilante que, para Morin (2005, p. 14), é "o pensamento que se engana, não porque não tem informação suficiente, mas porque não é capaz de ordenar as informações e os saberes, é um pensamento que conduz a ações mutilantes". A hiper simplificação, característica da ciência moderna, não permite que o cientista veja a complexidade do real (MORIN, 2005). O conhecimento é fragmentado em disciplinas que, frequentemente, compõem campos do conhecimento que não se conversam.

Morin (2002) afirma que o problema da complexidade, ou seja, a dificuldade em conceber a ciência por meio de conceitos claros, distintos e fáceis, dá-se pelo envolvimento da produção do conhecimento científico com três conjuntos de condições subjetivas: bio-antropológicas, socioculturais e noológicas. Segundo sua perspectiva, é necessário "conceber os limites biológicos, os limites cerebrais, os limites antropológicos, os limites sociológicos, os limites culturais de todo o conhecimento, o que nos permitirá ao mesmo tempo conhecer o nosso conhecimento" (MORIN, 2002, p. 32). Somente desta forma, é possível expandir os territórios do conhecimento, confrontando com as características do real.

Em uma perspectiva mais pragmática, Umberto Eco (2016, p. 27–31), ao discorrer sobre a cientificidade, define da seguinte maneira as características de um estudo científico, aplicadas às mais amplas áreas de pesquisa:

1) O objeto de estudo deve ser reconhecido por outros pesquisadores. A formalização das condições de investigação, que pode ocorrer com base em regras estabelecidas previamente ou pela própria pesquisa, é fundamental para caracterizar o objeto, que não se refere necessariamente a algo físico.
2) Como resultado, o estudo deve apresentar algo novo em relação ao objeto de pesquisa. O critério de novidade pode se referir ao conteúdo apresentado ou mesmo à abordagem do problema.
3) O estudo precisa ser útil. Deve-se apresentar uma contribuição benéfica aos demais pesquisadores da área.
4) O estudo deve apresentar subsídios para verificação, e possíveis contestações, das hipóteses testadas. Desta forma, é possível haver, de forma pública, a continuidade da pesquisa.

Com diversos fatores em comum, Gil (2008, p. 8) afirma que para um conhecimento ser considerado científico, é necessário identificar as operações, sejam mentais ou técnicas, que permitam sua verificação. Ou seja, o método científico (explorado como processo de Ciência de Dados ao final da seção 2.5.1) é o "conjunto de procedimentos intelectuais e técnicos adotados para se atingir o conhecimento" (GIL, 2008, p. 8). Desta forma, para Ciência de Dados, o termo "ciência" implica em conhecimento adquirido por meio de um estudo sistemático, é a construção e organização do conhecimento por meio de explicações e predições realizadas de forma testada e verificável (DHAR, 2013, p. 64).

Marconi e Lakatos (2003, p. 80) entendem ciência como "uma sistematização de conhecimentos, um conjunto de proposições logicamente correlacionadas sobre o comportamento de certos fenômenos que se deseja estudar". A ciência é um processo, em que o conhecimento é constantemente revisitado e modificado conforme as evidências são reveladas. Diante de novas evidências, que revelem lacunas no entendimento atual, este pode ser modificado ou até mesmo rejeitado, gerando uma nova compreensão sobre um determinado fenômeno (MCCRACKIN, 2017).

Longe da esperada rigidez acadêmica, termos como "científico" ou "cientificamente comprovado" são adotados como forma de adquirir mérito ou um

atestado de confiabilidade. Para Chalmers (1993), quem utiliza este tipo de jargão busca demonstrar fundamentação e se colocar acima de qualquer contestação. Chibeni (2004) justifica este contexto pela confiança generalizada que é depositada na ciência, pois se acredita que o conhecimento científico está acima das demais formas de conhecimento. Para o autor, as indústrias evocam teorias, métodos e técnicas da ciência para legitimar a qualidade e o processo de fabricação de seus produtos. Seria este o caso da Ciência de Dados?

Diante destas reflexões, questiona-se: a Ciência de Dados é realmente ciência? Para responder esta pergunta, Wolfgang Pietsch (2017) a contrapõe diante de outra questão: ou a Ciência de Dados é meramente uma prática inferior que não consegue se sustentar sozinha e pode, no máximo, contribuir com outros empreendimentos científicos? Em seguida, o autor responde não haver consenso na literatura sobre estas perguntas, sendo que a discordância é a principal característica das discussões.

Dentre os entusiastas da Ciência de Dados, destaca-se Jim Gray (2007) que apresenta o conceito de eSience como um método científico totalmente transformado, um novo empirismo. Para o autor, este novo paradigma da ciência é o encontro dos cientistas com a tecnologia da informação, unificando teoria, experimentação e simulação. A quantidade de dados, seja capturada por instrumentos, seja gerada por simuladores, exige novas soluções na coleta, no processamento, na gestão e, logicamente, na análise de dados.

Por outro lado, Gray (2007) reforça haver carência de novos algoritmos que possam lidar com o volume e a diversidade de dados disponíveis. Nesta perspectiva, Dhar (2013) reforça a importância da Ciência de Dados em lidar com quantidades de dados crescentes, heterogêneos e desestruturados. Além disso, Granville (2014) destaca que a Ciência de Dados ultrapassa os limites da estatística quando envolve, além da análise de dados, a implementação de algoritmos que, automaticamente, processam dados e provêm predições e ações automatizadas. Nesta mesma perspectiva, Pietsch (2017) afirma que a Ciência de Dados proporcionou, com seus “poderosos algoritmos”, resultados inovadores em diversos campos científicos. O autor defende o caráter indutivista da Ciência de Dados, em especial pela possibilidade de extrapolar análises tradicionais, como correlações, estabelecendo relações de causa e efeito.

Na prática, assim como cientistas sabem como dividir grandes problemas em questões menores, os cientistas de dados precisam adotar estratégias de problemas auxiliares menores para solucionar uma questão maior que pode ser intratável, dependendo do volume de dados (LOUKIDES, 2012, p. 15). Por isso, Leek (2013) considera que o termo chave de Ciência de Dados é “ciência”, e não “dados”, pois de nada adianta grandes volumes de dados sem a devida formulação de perguntas que possam ser respondidas por métodos que gerem valor às organizações. Analogamente, Hayes (2017) classifica o termo “ciência de dados” como redundante, pois a ciência depende de dados, uma vez que é a maneira com que os cientistas testam suas ideias.

Mas como é a relação da Ciência de Dados com outro item importante da ciência, a reprodutibilidade? Para refletir sobre esta questão, são trazidos os itens apontados por Fiona Fidler e John Wilcox (2021) como centrais na crise da reprodutibilidade científica, tema que ganhou relevância nas últimas décadas. Dentre os tópicos destacados pelos autores estão: a) virtual ausência de estudos de replicação publicados na literatura de diversos campos científicos; b) falha em reproduzir resultados de estudos publicados; c) evidências de viés de publicação; d) alta ocorrência de práticas de pesquisa questionáveis que, em geral, inflacionam as taxas de falso positivo e; e) a falta de transparência e completude no relato de métodos, dados e análise na publicação científica.

Para combater esta crise, destacam-se as características da ciência aberta. As iniciativas para lidar com os tópicos apresentados envolvem o compartilhamento de dados, o registro prévio público de estudos e a defesa de políticas editorias mais rígidas “em torno de relatórios estatísticos, incluindo a publicação de estudos de replicação e resultados estatisticamente não significativos” (FIDLER; WILCOX, 2021, n.p., tradução nossa). Estas práticas estão alinhadas à visão de Gray (2007), quando o autor defende que tudo na ciência está mudando pelos impactos da tecnologia da informação. Segundo o pesquisador, tanto a ciência experimental, quanto teórica e computacional está sendo afetada pelo “dilúvio de dados”, resultando na emergência de um paradigma científico intensivo em dados. A tecnologia permite que a literatura esteja online, que os dados das pesquisas, independentemente do tamanho, e os códigos de análise estejam online, possibilitando que todo esse conteúdo seja interoperado por outros pesquisadores.

É razoável concluir que a relação entre Ciência de Dados e a ciência vai além de uma pretensa chancela de qualidade que o termo “ciência” traz à área pesquisada. Conforme apresentado, os algoritmos, os métodos, as técnicas, bem como a própria tecnologia voltada à área de dados, são cada vez mais difundidos no meio científico. De acordo com Vieira (2008), o conhecimento científico se deve a instituições ou centros de pesquisa, compostos por equipes multidisciplinares. Por isso, geralmente, conhecimento em uma única área do conhecimento não é mais suficiente para se produzir ciência, vide a complexidade e sofisticação do contexto atual. Neste sentido, para categorizar a Ciência de Dados como ciência, tem-se como mais adequado o paradigma da complexidade, explorado por Bachelard (1996) e Morin (2002, 2005).

Se a prática da ciência é a extração de conhecimento dos dados, pode-se afirmar que Ciência de Dados é, sim, ciência. Sobretudo, se for ponderado que sem os métodos da Ciência de Dados, muitos conhecimentos estariam indisponíveis ou inacessíveis. Por outro lado, se a ciência é vista como forma de explicar coletiva e publicamente qualquer tipo de fenômeno, a Ciência de Dados deixar de cumprir requisitos, como a reprodutibilidade, especialmente, se considerada a prática de organizações privadas.

Para concluir esta discussão, recorre-se à descrição de Chalmers (1993, p. 62) sobre a ciência, na qual o autor afirma ser “essencial compreender a ciência como um corpo de conhecimento historicamente em expansão”. Para se validar uma teoria, bem como qualquer conclusão científica, é mandatório considerar o contexto histórico. Se o contexto em questão é a era do big data, a forma de produção da ciência é inevitavelmente transformada por ele. A ciência não é apenas Ciência de Dados. Assim como a Ciência de Dados nem sempre é ciência, embora também o seja.

2.4 CIÊNCIA DE DADOS E A INTERDISCIPLINARIDADE

A interdisciplinaridade adquire relevância em pesquisas acadêmicas do Brasil na década 1970, impulsionada principalmente pelos estudos de Ivani Fazenda, precursora da área (FAZENDA; TAVARES; GODOY, 2018). Para a autora, a

interdisciplinaridade, que ainda enfrenta os mesmos dilemas de décadas atrás, não pode ser contida em uma abordagem teórica única, absoluta e geral (FAZENDA, 2017). As questões relativas à interdisciplinaridade estão diretamente relacionadas ao percurso teórico pessoal de cada indivíduo que se arrisque a pesquisar e praticar o tema.

Por outro lado, a própria Ivani Fazenda (2017) destaca que a primeira obra brasileira com significância para as discussões sobre a interdisciplinaridade é o livro "Interdisciplinaridade e patologia do saber", publicado por Hilton Japiassu em 1976. Neste livro, Japiassu transcende a interdisciplinaridade para além de um conceito teórico. Para o autor, interdisciplinaridade é intrinsecamente uma atitude e uma prática individual, feita de "curiosidade, de abertura, de sentido da descoberta, de desejo de enriquecer-se com novos enfoques, de gosto pelas combinações de perspectivas e de convicção levando ao desejo de superar os caminhos já batidos" (JAPIASSU, 1976, p. 82). Por isso, é algo que não pode ser aprendido, apenas exercido.

Dentre os avanços proporcionados pela interdisciplinaridade e exaltados por Japiassu (1976), estão a superação do dualismo entre pesquisa pura e aplicada, e o consequente surgimento de um novo tipo de investigação, simultaneamente teórica e prática, chamada de pesquisa orientada. Fazenda, Tavares e Godoy (2018) reforçam que, uma vez que a interdisciplinaridade é fundamentada na disciplinaridade, é necessário que se conheça todos os passos de uma metodologia tradicional de pesquisa. Por outro lado, a pesquisa dita interdisciplinar envolve distintos métodos de investigação que proporcionam benefícios incomensuráveis, sem causar prejuízos de cientificidade. Desta forma, a interdisciplinaridade é classificada como caminho para uma "não robotização" dos indivíduos que traz a pesquisadores, conhecimento aprofundado sobre suas práticas e sobre si mesmos (FAZENDA; TAVARES; GODOY, 2018).

Neste sentido, Alvarenga et al. (2011) apresentam a interdisciplinaridade como alternativa às fronteiras impostas pela simplificação, dicotomia e segmentação características da ciência moderna ou clássica. Para estes autores, pesquisas interdisciplinares representam a inovação na produção do conhecimento científico, apresentando-se como alternativa e complemento à visão disciplinar do conhecimento. Desta forma, a interdisciplinaridade tem como "finalidade última dar

conta de fenômenos complexos, de diferentes naturezas", trabalhando na conexão ou reconexão de saberes que, quando separados, apresentam uma visão fragmentada e simplista da realidade (ALVARENGA *et al.*, 2011, p. 20).

Para Pacheco, Tosta e Freire (2010, p. 137), a interdisciplinaridade nada mais é do que a integração sistemática de distintas disciplinas, de maneira que suas características, verdades e diferenças sejam respeitadas, reintegrando-as a um todo outrora naturalmente unido. Neste sentido, Balbino (2021, p. 92) reforça que a produção interdisciplinar não despreza ou invalida qualquer conhecimento oriundo das disciplinas elementares. Pelo contrário, ao respeitar os paradigmas das áreas basilares, o novo conhecimento colabora com o desenvolvimento destas disciplinas.

É com a interdisciplinaridade que os horizontes do conhecimento podem ser expandidos de maneira mais audaciosa e independente, sem abrir mão da cautela e do rigor científico (FAZENDA, 2017). E é neste ponto que a interdisciplinaridade e a Ciência de Dados se encontram. A Ciência de Dados é por definição um campo interdisciplinar (LEY; BORDAS, 2018, p. 170) e foi desta maneira classificada desde seu surgimento (CAO, 2017). A Ciência de Dados é formada a partir de disciplinas tradicionais como matemática, estatística e ciências da computação (GRANVILLE, 2014, p. 84; HALL; PHAN; WHITSON, 2016, p. 7), mas também se relacionada à ciência da informação, gestão do conhecimento, sistemas de gestão da informação e ciência da decisão (CHEN *et al.*, 2018, p. 172).

Para Cao (2017), ao lidar com problemas de dados de maneira inovadora, a Ciência de Dados precisa de abordagens e metodologias interdisciplinares e sistemáticas. O autor resume as estratégias características da Ciência de Dados da seguinte maneira:

> Envolve o desenvolvimento de uma sinergia de várias disciplinas e áreas de pesquisa, incluindo representação de dados, pré-processamento e preparação, processamento de informações, processamento paralelo, sistemas distribuídos, computação de alto desempenho, gestão de dados, armazenamento de dados, computação em nuvem, computação evolutiva, redes neurais, sistemas difusos, infraestrutura empresarial, arquitetura de software, comunicação e rede, integração e interoperação, aprendizado de máquina, modelagem de dados, análise e mineração, computação de serviço, simulação de sistema, projeto experimental e avaliação. Também pode envolver aspectos de negócios e sociais, incluindo transformação da indústria, sistemas de informações empresariais, inteligência de negócios, gestão de processos de negócios, gestão de projetos, segurança da informação, confiança e reputação, processamento de privacidade, modelagem de impacto

> nos negócios, valor de negócios e avaliação de utilidade (CAO, 2017, p. 24–25, tradução nossa).

Logicamente, para dar conta de todos estes elementos, são necessários conhecimentos advindos de múltiplas disciplinas. Neste sentido, Chen et al. (2018, p. 172) afirmam que a Ciência de Dados está mais próxima de algumas disciplinas do que outras, como é comum encontrar na literatura acerca do tema. Mesmo assim, os autores, que consideram que a finalidade da Ciência de Dados é beneficiar os seres humanos, evitam enumerar as disciplinas relacionadas à área, pois acreditam que todo campo científico ou mercadológico é impactado e pode se aproveitar das oportunidades geradas neste contexto. Visão parecida é compartilhada por Rawlings-Goss (2019, p. 41) que afirma que a interação da Ciência de Dados com outras disciplinas podem geram inúmeros benefícios a áreas como astronomia, biologia, negócios, química, ciências ambientais, dados médicos, ciências políticas, física, ciências sociais, ciências comportamentais, artes e humanidades, dentre outras.

A interdisciplinaridade, por meio de seu exercício, proporciona ao pesquisador, ou praticante, um aprofundamento em conhecimento, seja pessoal, seja das próprias práticas interdisciplinares, evitando escolhas já percorridas e proporcionando um "saber transcendental" (FAZENDA; TAVARES; GODOY, 2018, p. irreg.). Desde a década de 1980, conforme os atributos compilados por Fazenda (2017), a interdisciplinaridade trazia elementos característicos da prática em Ciência de Dados. Segundo a descrição da autora, a interdisciplinaridade já estava relacionada a uma atitude relacionada à ação, ao perguntar e ao duvidar, ao processo que traz resultados criativos e ousados. Desta forma, por mais que a interdisciplinaridade não seja uma competência em si, a atitude interdisciplinar do cientista de dados lhe proporciona gosto pela pesquisa, pelo conhecimento e comprometimento com seu ofício.

2.5 METODOLOGIAS E DISCIPLINAS RELACIONADAS

Nesta seção, busca-se distinguir a Ciência de Dados de conceitos relacionados, sejam metodologias, disciplinas ou outros termos da terminologia da área. Inicia-se pelo processo de Ciência de Dados que compreende as etapas necessárias para que os dados brutos proporcionem insumos para a tomada de

decisão (*insights*), abordando a adoção do método científico para esta finalidade. Em seguida, são exploradas as metodologias e práticas seguidas pelo mercado para balizar os projetos de dados. Por fim, são explorados conceitos e disciplinas cuja delimitação se faz necessária à presente pesquisa.

2.5.1 Do dado ao *insight*

Como apresentado, Ciência de Dados está relacionada à extração de conhecimento a partir de grande quantidade de dados, subsidiando ações, especialmente em relação à tomada de decisão (BRANDT, 2016; DHAR, 2013; DONOHO, 2017; PARKS, 2017; PWC, 2017). O produto dessa extração é comumente chamado de *insight* e é entendido como o conteúdo descoberto e aprendido, geralmente a partir de padrões, sendo utilizado para melhorar a gestão dos negócios (HAYES, 2017). Contudo, o caminho que parte da aglomeração de dados brutos e chega a novos recursos de valor para a organização possui vários estágios.

Desde as primeiras aparições do termo Ciência de Dados, as etapas do processo já eram uma preocupação. Para Hayashi (1998, p. 41), a Ciência de Dados, sucintamente, é composta por três fases: *design* de dados, coleta de dados e análise de dados. Sob sua perspectiva, um fenômeno geralmente é multifacetado, por isso as delimitações do experimento importam para definir qual abordagem será adotada (fase de *design* de dados). Os dados devem ser claros mesmo expressando múltiplas dimensões e apresentando, muitas vezes, representações de fenômenos temporais. Por isso, a etapa de coleta deve ser criteriosa para representar tais propriedades. Finalmente, por meio de métodos de classificação e análises, além de modelos matemáticos e estatísticos, a conceituação e a simplificação dos dados são reveladas, concluindo a terceira fase.

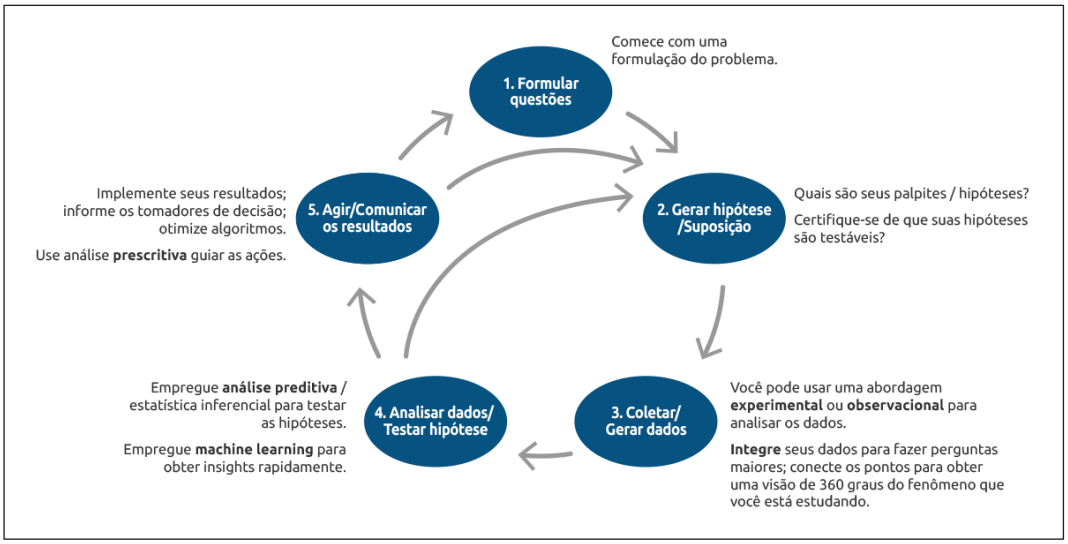
Duas décadas depois, Kampakis (2020, p. 52) define cinco etapas para o ciclo de vida da Ciência de Dados: 1) Coleta de dados: obtenção de dados de fontes internas e externas à empresa; 2) Organização dos dados: garantia que os dados estão devidamente formatados, sem inconsistências ou ruídos; 3) Análise de dados: engloba desde a análise exploratória dos dados à definição dos métodos mais apropriados; 4) Interpretação dos dados: verificação dos indicadores, com

participação do especialista do domínio, para extrair *insights* que possam ser convertidos em ações e; 5) Comunicação das descobertas dos dados: sumarização dos resultados e apresentação às partes interessadas de forma simplificada.

Kampakis (2020, p. 52) também define o Processo da Ciência de Dados, que engloba as etapas onde a participação do cientista de dado é essencial. O processo proposto é composto por quatro passos: 1) Definição do problema: em conjunto com o especialista do domínio; 2) Escolha dos dados corretos: o cientista de dados precisa conhecer o domínio para selecionar os elementos mais relevantes ao problema; 3) Solução do problema: etapa fundamental, onde define os métodos mais adequados para resolver o problema definido e; 4) Criação de valor por meio de *insights* que geram ação: etapa na qual o cientista de dados demonstra sua contribuição para os negócios e a importância da Ciência de Dados, contribuindo para a criação de uma cultura de dados.

De forma mais abrangente, Hayes (2017) aponta como maneira genérica de se obter *insights* a partir de grandes quantidades de dados, um modelo que é utilizado há muitos séculos: o método científico. Segundo a definição adotada pelo autor, o método científico é um conjunto de técnicas utilizadas para investigar um fenômeno, adquirir novo conhecimento ou contradizer e integrar conhecimento prévio, sendo composto pelas fases apresentadas na Figura 8.

FIGURA 8 – OBTENDO INSIGHTS DOS DADOS PELO MÉTODO CIÊNTÍFICO



FONTE: Adaptado de Hayes (2017).

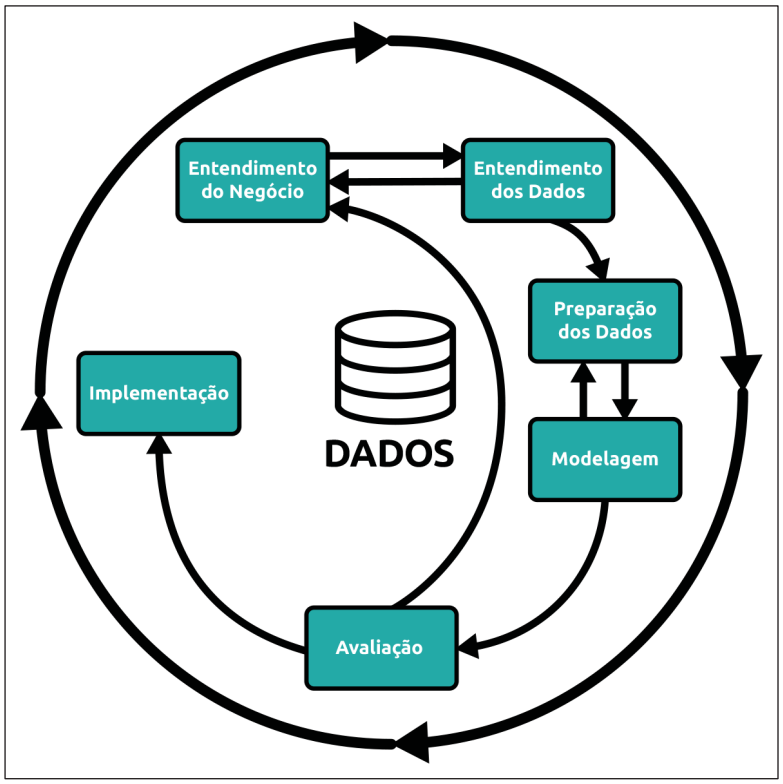
Em linhas gerais, as etapas do método científico, definição da questão de pesquisa, formulação de hipótese, coleta/geração de dados, análise de dados e comunicação, encaixam-se nas etapas da Ciência de Dados. Adicionalmente, os resultados do método científico, pelo menos em teoria, devem fornecer subsídios para pesquisas futuras, ajudando na elaboração de novas ou refinadas questões de pesquisa. De qualquer maneira, há metodologias específicas para se extrair conhecimento dos dados. A seguir, as práticas aqui identificadas são detalhadas e comparadas entre si.

2.5.2 Metodologias e práticas do mercado

Em relação às práticas do mercado, o site KDNuggets (PIATETSKY-SHAPIRO, 2014) publicou o resultado de uma enquete com a seguinte pergunta: “Qual a principal metodologia que você utiliza em seus projetos de análise, mineração de dados ou ciência de dados?”. Os resultados, que muito se mostraram estáveis em relação à mesma pesquisa realizada em 2007, indicaram que a metodologia mais utilizada é a CRISP-DM, adotada por 43,0% dos 200 respondentes. Na segunda posição, 27,5% dos participantes utilizam sua própria metodologia de projeto, seguidos pelos adotantes da metodologia SEMMA (8,5%), outra não-especificada (8,0%), processo KDD (7,0%), processo próprio da organização contratante (3,0%) e, finalmente, aqueles que utilizam metodologias para domínios específicos (2,0%).

A metodologia CRISP-DM, acrônimo *de Cross Industry Standard Process for Data Mining*, ou Processo Padrão de Mineração de Dados Intersetorial (tradução nossa), é um modelo de processos, iniciado em 1996, com um conjunto de tarefas voltadas a projetos em Mineração de Dados (CHAPMAN *et al.*, 2000). Dentre as principais vantagens do modelo estão sua popularidade, uma vez que é amplamente adotado, e o fato de ter sido desenvolvido de maneira independente de qualquer software, setor, aplicação ou técnica de análise (KELLEHER; TIERNEY, 2018). O ciclo de vida de um projeto de dados proposto pelo CRISP-DM é composto por seis estágios: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implementação, conforme demonstrado na Figura 9.

FIGURA 9 – CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)



FONTE: Adaptado de Chapman et al. (2000).

Como principal recurso da estrutura, os dados são localizados no centro do modelo. As setas indicam as mais importantes e frequentes dependências entre as fases. Porém, a ordem das fases não é rígida e o fluxo do processo avança ou retrocede entre elas, sempre que necessário. Esse movimento é definido pelo

resultado de cada fase ou de uma tarefa em particular, que determina qual o procedimento seguinte. O círculo externo simboliza a natureza cíclica inerente ao processo de mineração de dados, estendido a projetos de Ciência de Dados, onde a implementação de produtos de dados gera resultados para novos questionamentos e, consequentemente, novos projetos da mesma natureza (CHAPMAN *et al.*, 2000).

Neste sentido, cada fase do CRISP-DM é composta, em um nível mais alto, por tarefas genéricas cujos resultados definem o prosseguimento do projeto. A Figura 10 sumariza os diferentes estágios de um projeto de Ciência de Dados, listando as tarefas mais frequentes em cada um.

FIGURA 10 – FASES E TAREFAS GENÉRICAS DO CRISP-DM

Entendimento do Negócio	Entendimento dos Dados	Preparação dos Dados	Modelagem	Avaliação	Implementação
Determinar os objetivos do negócio Background Objetivos do negócio Critérios de sucesso do negócio **Avaliar a situação** Inventário de recursos Requisitos, pressupostos e restrições Riscos e contingências Terminologia Custos e Benefícios **Determinar as metas de mineração de dados** Metas de mineração de dados Critérios de sucesso de mineração de dados **Produzir plano de projeto** Plano de projeto Avaliação inicial de ferramentas e técnicas	**Coletar dados iniciais** Relatório da coleta de dados iniciais **Descrever os dados** Relatório da descrição dos dados **Explorar os dados** Relatório da exploração dos dados **Verificar a qualidade dos dados** Relatório da qualidade dos dados	**Selecionar os dados** Critérios de inclusão / exclusão **Limpar os dados** Relatório de limpeza dos dados **Construir os dados** Atributos derivados Registros gerados **Integrar os dados** Dados mesclados **Formatar os dados** Dados reformatados **Conjunto de dados** Descrição do conjunto de dados	**Selecionar técnicas de modelagem** Técnica de modelagem Suposições de modelagem **Gerar projeto de teste** Projeto de teste **Construir o modelo** Modelos de configurações de parâmetros Descrições do modelo **Avaliar o modelo** Avaliação do modelo Configurações de parâmetro revisadas	**Avaliar resultados** Avaliação dos resultados da mineração de dados Critérios de sucesso do negócio Modelos aprovados **Revisar processos** Revisão dos processos **Determinar os próximos passos** Lista de possíveis ações de decisão	**Planejar a implementação** Plano de implementação **Planejar o monitoramento e manutenção** Plano de monitoramento e manutenção **Produzir o relatório final** Relatório final Apresentação final **Revisar o projeto** Documentação da experiência

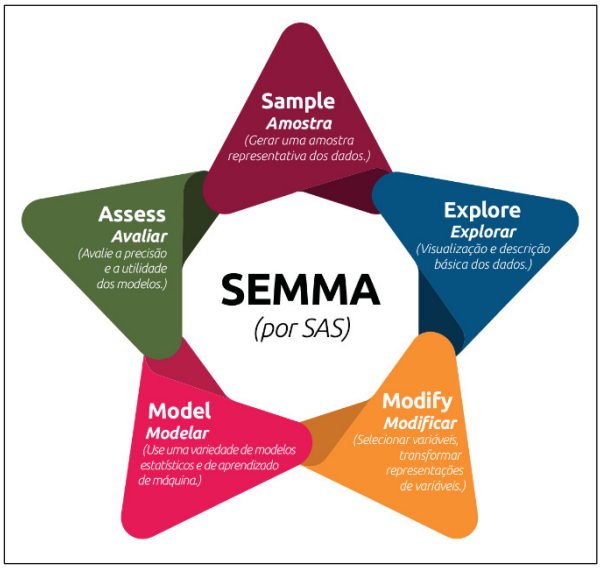
FONTE: Adaptado de Champman et al. (2000).

Assim como o fluxo apresentado anteriormente, as tarefas de cada fase são uma referência e sua efetivação depende da natureza de cada projeto. Para Kelleher e Tierney (2018), um dos principais equívocos de cientistas de dados inexperientes é priorizar a fase de modelagem, negligenciando as demais. Ao acreditarem que a principal entrega de um projeto de dados é o modelo quantitativo, dedicam muito tempo à construção e refinamento estatístico. Por outro lado, profissionais mais experientes se dedicam a garantir que o projeto possua os dados corretos para a solução do problema. Devido à sua relevância, as tarefas de coletar, limpar e

organizar os dados consomem mais da metade do tempo de trabalho dos cientistas de dados (CROWDFLOWER, 2016, 2017).

A segunda metodologia mais recorrente na enquete realizada por Piatetsky (2014) ao site KDNuggets foi a SEMMA, de *Sample, Explore, Modify, Model and Assess*, criada em 2009 pelo Instituto SAS, uma das empresas pioneiras no desenvolvimento de softwares para estatística e BI. Na metodologia proposta pelo SAS, o ciclo de um projeto de dados é composto por cinco passos apresentados na Figura 11.

FIGURA 11 – MODELO SEMMA PARA PROJETOS DE DADOS



FONTE: Adaptado de Piatetsky-Shapiro (2014).

Em seu manual de referência (SAS INSTITUTE INC., 2017), os passos da SEMMA são assim descritos:

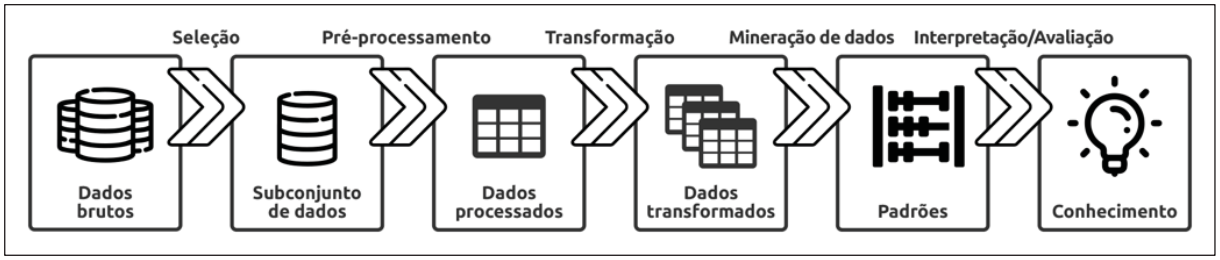
- **Amostra:** criação da amostragem de dados, por uma ou mais tabelas, que seja grande suficiente para apresentar informações significativas, mas pequena o bastante para ser processada.
- **Explorar:** análise exploratória dos dados para identificar relações, tendências ou mesmo anomalias, para maior compreensão e novas ideias.
- **Modificar:** alteração dos dados pela criação, seleção ou transformação das variáveis com foco no processo de seleção do modelo.

- **Modelar:** modelagem dos dados com a adoção de métodos estatísticos e ferramentas analíticas para se obter os resultados pretendidos, garantindo a segurança.
- **Avaliar:** validação da utilidade e confiabilidade das descobertas do processo.

Diferentemente da abordagem do CRISP-DM, onde a primeira etapa corresponde ao entendimento do negócio, percebe-se que a SEMMA é concentrada exclusivamente nas tarefas de modelagem do processo de mineração de dados, estendido para projetos de Ciência de Dados. Além disso, embora possa se adaptar a qualquer situação, a metodologia foi desenvolvida e direcionada aos profissionais que utilizam as ferramentas SAS. Mesmo assim, a SEMMA proporciona uma fácil compreensão do processo em projeto de dados, permitindo o desenvolvimento e a manutenção dos projetos, de maneira organizada e adequada (AZEVEDO; SANTOS, 2008).

A terceira metodologia pública mais recorrente na referida enquete foi o KDD, *Knowledge Discovery in Databases*, ou Descoberta de Conhecimento em Bases de Dados (tradução nossa), sendo iniciada na década de 1980 e, portanto, a mais antiga dentre os modelos citados. Diante da necessidade de novas teorias e ferramentas para auxiliar na extração de conhecimento de volumes de dados cada vez maiores, Fayyad, Piatetsky-Shapiro e Smyth (1996) propõe um modelo que atenda às necessidades desse processo. Os autores definem o KDD como uma atividade multidisciplinar que engloba técnicas que não se restringem aos escopos de uma disciplina específica, como *machine learning*. A proposta foca no processo de descoberta do conhecimento de maneira geral, incluindo a armazenagem e o acesso aos dados, além dos algoritmos utilizados.

FIGURA 12 – VISÃO GERAL DOS PASSOS DO KDD



FONTE: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Na Figura 12, percebe-se que os dados são o elemento inicial do processo que é concluído na obtenção do conhecimento. Contudo, o processo não é linear. Assim como explicitado para a metodologia CRISP-DM, o KDD é um processo interativo e iterativo, que "envolve numerosas etapas determinadas pelas decisões do usuário" (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 42, tradução nossa). Essa metodologia, que também não cita questões de negócios, é composta por cinco fases principais: Seleção, Pré-processamento, Transformação, Mineração de Dados e Interpretação/Validação.

Todas as três metodologias, em especial a CRISP-DM e KDD, embora apresentem boa descrição do processo analítico, foram desenvolvidas há muitos anos e não abordam questões como *big data*, computação em nuvem, IoT, dentre outras (SALTZ; GRADY, 2017). Conforme destacado por Piatetsky (2014), a ausência de metodologias posteriores que englobem estes elementos está relacionada ao aumento de profissionais que não utilizam nenhuma das três citadas, desenvolvendo suas próprias metodologias ou adotando outras não especificadas. Esse público que na pesquisa de 2007 era equivalente a 23,0% dos respondentes passou para 35,5% na pesquisa de 2014.

Pela exposição dos diagramas de Venn e dos processos da Ciência de Dados, percebe-se que a área engloba inúmeros termos chaves que estão conectados e podem, facilmente, gerar confusão entre si (CAO, 2017). Termos como *big data*, análise de dados, engenharia de dados, mineração de dados são alguns exemplos. A seguir, procura-se estabelecer um breve entendimento, especificamente em relação à Ciência de Dados, sobre termos relevantes para a presente pesquisa e que se enquadram neste critério.

2.5.3 Conceitos e disciplinas correlatas

O emaranhado de disciplinas que formam a Ciência de Dados, exposto pelos diagramas delineados na seção 2.2, envolve conceitos e relações que nem sempre são distinguidos com clareza. Cao (2017, p. 3–4) sumariza os termos chaves da terminologia da Ciência de Dados que frequentemente geram mal entendidos, como *data analysis*, *data analytics*, *big data*, *data science*, *predictive analytics*, dentre outros. O objetivo dessa seção é estabelecer a definição de alguns desses termos perante a Ciência de Dados.

Além dos termos elencados por Mayo (2016), expostos na sequência, foram selecionados outros conceitos que carecem de maior aprofundamento. Os termos abordados são Estatística e Análise de Dados, áreas já estabelecidas quando a Ciência de Dados ganha relevância, seguidos por Engenharia de Dados, uma disciplina contemporânea ao tema pesquisado.

2.5.3.1 Estatística

A Estatística está diretamente relacionada ao método científico. Em uma definição clássica, Kendall (1945, p. 2, tradução nossa) define a estatística como "o ramo do método científico que lida com os dados obtidos, contabilizando ou mensurando as propriedades de populações de fenômenos naturais". Para exemplificar a evolução do conceito, Smith (2014, p. 3) seleciona uma definição de 2009 concebida pelo presidente da Royal Statistical Society, David Hand, para o qual a Estatística é

> "a diversão em encontrar padrões em dados, o prazer em realizar descobertas, aprofunda-se em questões filosóficas, o poder em lançar luz sobre decisões importantes e a capacidade de orientar decisões em negócios, ciência, governo, medicina, indústria" (2014, p. 3, tradução nossa)

Nota-se, pelas duas definições, a importância e relevância que a Estatística adquiriu nas duas primeiras décadas do século XXI, extrapolando os limites da área científica e sendo adotada pelos mais diversos setores da sociedade. De fato, a Estatística, cuja origem vem da Matemática do século XVIII, foi a primeira tentativa de analisar dados sistematicamente e produziu os fundamentos de toda as técnicas

modernas de análise de dado, como análise descritiva e exploratória, regressão, classificação, teste de significância, previsões, dentre outros (KAMPAKIS, 2020, p. 14).

Porém, para Donoho (2017), a profissão de estatístico se encontra em um momento confuso: as atividades que foram objeto de estudo durante tantos anos estão agora sendo reivindicadas como novas e revolucionárias, e sendo largamente praticadas por iniciantes e estranhos a esta área do conhecimento. Segundo o levantamento realizado pelo autor supracitado, esse contexto gera reação dos estatísticos que questionam a relação da Ciência de Dados com a Estatística. Os títulos dessas publicações sugerem o teor da discussão: “Nós não somos Ciência de Dados?”, “Um grande debate: a Ciência de Dados é apenas uma “*rebranding*” da Estatística?”, “Deixem a Ciência de Dados para nós”. Outra vertente questiona o real valor da Ciência de Dados: “Por que precisamos da Ciência de Dados quando temos a Estatística há séculos?”, “Ciência de dados **é** Estatística”. Por outro lado, há publicações que diminuem o valor da Estatística para a Ciência de Dados: “A Ciência de Dados sem Estatística é possível, até mesmo desejável”, “Estatística é a parte menos importante da Ciência de Dados”.

O que diferencia os cientistas de dados dos estatísticos é a abordagem holística da Ciência de Dados, que engloba as ações de obter os dados na forma bruta, tratar este material para uma forma manipulável, extrair uma história de valor para o domínio em questão e comunicar esta história para os interessados (LOUKIDES, 2012, p. 4). A própria natureza desse material bruto da Ciência de Dados, que está cada vez mais heterogêneo e desestruturado, composto por textos, imagens, vídeos, muitas vezes obtido de profundas redes de relacionamentos, configura-se em outra divergência para a abordagem meramente estatística (DHAR, 2013, p. 64).

A aplicação de métodos estatísticos tradicionais a dados dessa natureza, em volumes progressivamente maiores, pode levar a conclusões equivocadas (GRANVILLE, 2014, p. 65). Uma vez que os métodos estatísticos tipicamente exigem dados com determinadas características e, usualmente, utilizam poucos atributos para produzir resultados, métodos mais modernos de *machine learning* podem utilizar milhões de parâmetros para encontrar similaridades e padrões nos dados (HALL; PHAN; WHITSON, 2016, p. 2). Assim, por mais que a Estatística seja o mais discutido

tópico da Ciência de Dados, esta não pode se basear exclusivamente nos preceitos estatísticos (BRANDT, 2016, p. 22).

2.5.3.2 Análise de dados

A análise de dados, um dos termos precursores à Ciência de Dados, refere-se à aplicação de teorias, tecnologias e ferramentas tradicionais voltadas aos dados (estatística clássica, matemática e lógica) para obtenção de informações, com fins práticos (CAO, 2017, p. 4). Por mais que os dois termos possam gerar certa confusão, sob a concepção da Ciência de Dados, a análise de dados é a subcategoria correspondente ao processo de dar significado aos dados por meio de técnicas de análise (EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS)DEMCHENKO; BELLOUM; WIKTORSKI, 2017, p. 8; WASHINGTON DURR, 2018, p. 2).

Para Parks (2017, p. 9), as funções em análise de dados são um grupo de competências do cientista de dados que correspondem às habilidades em Matemática, em métodos de análises e em modelagem estatística necessárias para solucionar um problema relacionado a dados. Já os analistas de dados são indivíduos cuja principal característica são seus sólidos conhecimentos em análise estatística. Sob esta perspectiva, torna-se difícil distinguir entre o cientista de dados e o analista de dados. Porém, este segundo profissional tende a estar focado em áreas específicas da organização, sendo, frequentemente, lotado na unidade de negócios (LINDEN *et al*., 2018, p. 12).

Saltaz e Grady (2017, p. 2359) apresentam um cenário no qual o analista de dados está localizado na interseção das disciplinas de métodos quantitativos (Matemática, Estatística e algoritmos) e de comunicação em dados. Assim, para estes autores, a função do analista de dados é fornecer relatórios e visualizações que expliquem *insights* ocultos nos dados. O analista de dados funciona como uma ponte entre os cientistas de dados e os analistas de negócio, traduzindo análises técnicas em itens qualitativos de ações para que haja uma comunicação eficiente das descobertas aos *stakeholders* (BERKELEY SCHOOL OF INFORMATION, 2020).

Ainda sob a proposta de Saltaz e Grady (2017, p. 2359), o cientista de dados se dá na união de três disciplinas. Além das duas apresentadas, Métodos

Quantitativos e Comunicação em Dados, a Engenharia de Software é a outra área integrante da Ciência de Dados. Neste arranjo, a interseção da Engenharia de Software com métodos quantitativos, configura a especialidade do engenheiro de dados, apresentada a seguir.

2.5.3.3 Engenharia de Dados

A Engenharia de Dados é uma área crítica para a Ciência de Dados e mais um termo que gera confusão dentro da terminologia da área (Data Science for Business: What you need to know about data mining and data-analytic thinkingPROVOST; FAWCETT, 2013, p. 7). A principal diferença entre estes profissionais destas duas áreas é que enquanto a maior responsabilidade do cientista de dados se concentra na exploração e descoberta de conhecimento nos dados, os engenheiros de dados focam na preparação e no processamento de dados, fornecendo os recursos tecnológicos necessários para que essa referida descoberta ocorra (CAO, 2019, p. 36).

O engenheiro de dados lança mão de sua experiência da engenharia de software para manipular grandes volumes de dados em escala, dedicando-se, normalmente, às tarefas de codificação, limpeza de bancos de dados e implementando requisições advindas dos cientistas de dados (BERKELEY SCHOOL OF INFORMATION, 2020; SALTZ; GRADY, 2017, p. 2358). Os profissionais da Engenharia de Dados, mais do que fornecerem acesso apropriado aos dados, são responsáveis por ganhos de produtividade, uma vez que garantem o alinhamento entre os dados de produção e dados de treinamento, além de fornecerem os subsídios operacionais para implementação de resultados da Ciência de Dados, como modelos de *machine learning* (LINDEN *et al.*, 2018, p. 12). De modo geral, é pela Engenharia de Dados que se projeta, desenvolve e gerencia a infraestrutura da informação para projetos em *big data* (KAMPAKIS, 2020, p. 19), um dos conceitos explorados a seguir.

2.5.4 Métodos, Técnicas e Tecnologias

Desde o início do século XXI, salienta-se que uma utilização eficiente dos dados na solução de problemas depende da migração de uma dependência exclusiva de modelos estatísticos para a adoção de um conjunto mais diversificado de ferramentas de análise (BREIMAN, 2001, p. 1). Esse posicionamento era decorrente do que acontecia no mercado, onde, desde meados dos anos 1980, a tomada de decisão utilizando grande quantidade de dados já era prática comum, tendo evoluído para a mineração de dados na década seguinte e tornado processos cada vez mais automatizados. Nessa época, métodos de *machine learning* eram empregados em problemas de negócio, levando a uma explosão de ferramentas e softwares que, a partir do uso dos dados, resultavam em aplicações explanatórias e preditivas (DHAR, 2013, p. 67).

As aplicações derivadas desses modelos são na retenção de clientes, especialmente nas áreas de telecomunicações e financeira, na análise de sentimentos, recomendações, detecção de fraudes, análise de crédito, propaganda *online*, reconhecimento de padrões e imagens, prevenção de falhas de equipamentos, sistemas de busca, filtros de *spam*, detecção de invasões em redes, dentre outros casos (HALL; PHAN; WHITSON, 2016, p. 3; Data Science for Business: What you need to know about data mining and data-analytic thinkingPROVOST; FAWCETT, 2013, p. 4). Os benefícios das tecnologias voltadas aos dados e da adoção de métodos científicos nas análises de negócios e soluções de problema são reconhecidos pelas empresas, que podem se manter inovativas e competitivas, fornecendo serviços avançados e centrados nos consumidores (DEMCHENKO *et al.*, 2016, p. 621; HALL; PHAN; WHITSON, 2016, p. 1). Desde então, a Ciência de Dados e tópicos relacionados se tornaram tema central de discussões em conferências de estatística, mineração de dados, *machine learning* e, mais recentemente, *big data* (CAO, 2017, p. 2), atraindo crescentemente o interesse de órgãos governamentais, do mercado e da própria academia (CAO, 2019, p. 1). A seguir, são explorados os componentes deste arranjo de métodos, técnicas e tecnologias que tornam a Ciência de Dados viável.

2.5.4.1 Big Data

Se há um consenso entre as definições sobre *big data*, ele se localiza no quesito tamanho. Os autores se referem ao *big data* como uma quantidade tão grande de dados que impossibilita que a captura, a gestão ou o processamento sejam realizados de forma eficiente por meio das teorias, tecnologias e ferramentas tradicionais (CAO, 2017, p. 4; KAMPAKIS, 2020, p. 20; MAYO, 2016). Porém, embora o volume de dados seja um item importante na concepção do *big data*, Russom (2011, p. 6) ressalta outros dois atributos fundamentais para uma definição compreensiva de *big data*: Variedade, referente à diversidade de dados, e Velocidade, remetente à capacidade de processamento.

Juntas, estas propriedades formam o conhecido conceito dos três Vs do *big data.* Posteriormente, este conjunto foi complementado com outros fatores, como Veracidade, que é o grau de confiabilidade dos dados, Variabilidade, referente à consistência dos dados ao longo do tempo e, Valor, potencial benefício proporcionado pelos dados (ANDREU-PEREZ *et al.*, 2015). O *big data* é consequência de um mundo conectado, digitalizado, inundado de sensores e orientado à informação, que gera vastos recursos de dados com potencial de responder perguntas cujas respostas estavam, anteriormente, fora de alcance (NIST BIG DATA PUBLIC WORKING GROUP, 2015). De qualquer forma, o real valor do *big data* não está nos dados em si, mas no conhecimento que a organização consegue extrair por meio da Ciência de Dados e ferramentas de análise de dados (HAWAMDEH; CHANG, 2018).

Com recursos para capturar, armazenar e processar grandes quantidades de dados, aproveitar as possibilidades proporcionadas pela combinação do fluxo e da análise de dados se torna um desafio importante que levaram à emergência da Ciência de Dados (POWER, 2016). Dentre outras designações, a maioria dos profissionais que trabalha com *big data* é reconhecida como cientista de dados (PARKS, 2017). Uma vez que o cerne da Ciência de Dados são os dados, o *big data* se torna o principal fornecedor de matéria-prima, o objeto a ser investigado (HAWAMDEH; CHANG, 2018).

2.5.4.2 Machine Learning

Como abordado no Diagrama de Venn da Ciência de Dados, de acordo com a perspectiva de Conway (2010), o *machine learning* (ML), ou aprendizado de máquina,

é a junção da Matemática e da Estatística à tecnologia aplicada aos dados. Em linhas gerais, *machine learning* corresponde a métodos computacionais capazes de "aprender" com a experiência para melhorar o desempenho ou aprimorar a acurácia das predições (HALL; PHAN; WHITSON, 2016, p. 3; KAMPAKIS, 2020, p. 4). Há três grupos para os métodos de *machine learning*, supervisionados, não-supervisionados e aprendizagem por reforço, dos quais os algoritmos mais populares são regressão, árvores de decisão, *random forest*, redes neurais artificiais e máquinas de vetores de suporte (SVM).

Machine learning é uma ferramenta essencial à Ciência de Dados, pois enquanto a Estatística e a Análise de Dados correspondem às atividades manuais do cientista de dados, é por meio dos métodos de aprendizagem de máquina que este profissional manipula grandes quantidades de dados (KIM; LEE, 2016, p. 169; LOUKIDES, 2012, p. 9). Por meio da aprendizagem de máquina, a organização tem a possibilidade de tomar decisões mais acuradas do que aquelas baseadas unicamente nos métodos analíticos tradicionais, pois esta metodologia envolve novas fontes de dados não-estruturados como imagens, sons, vídeos, dentre outros recursos (HALL; PHAN; WHITSON, 2016, p. 3). Neste cenário, o objetivo do *machine learning* é auxiliar os profissionais de dados a darem sentido a quantidades massivas de dados (LANTZ, 2017, p. 1), predizendo o que acontecerá no futuro a partir de registros passados (RAWLINGS-GOSS, 2019, p. 11).

Dentre os algoritmos de *machine learning*, há um subgrupo de métodos, chamado de *deep learning*, que vem ganhando popularidade por conseguir avançar na solução de problemas, por muito anos considerados como insolúveis pela comunidade de inteligência artificial. Enquanto os métodos tradicionais de aprendizado de máquina eram limitados em sua capacidade de processar grandes quantidades de dados em seu estado bruto, o *deep learning* emprega modelos computacionais compostos por várias camadas de processamento que aprendem a representar os dados com vários níveis de abstração (LECUN; BENGIO; HINTON, 2015, p. 436, tradução nossa). O *deep learning* é, com sua capacidade aprimorada em classificar, reconhecer, detectar e descrever, um componente fundamental às aplicações mais avançadas, confirmando-se como o estado da arte do *machine learning* (HALL; PHAN; WHITSON, 2016, p. 6).

2.5.4.3 Mineração de dados

A mineração de dados é a aplicação de algoritmos específicos para extração de padrões a partir de dados. Com essa definição, Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 39) fazem a distinção do conceito para o processo de descoberta de conhecimento em bases de dados (KDD), apresentado anteriormente. Para os autores, enquanto o KDD é todo o processo para identificar padrões válidos, novos, potencialmente úteis e legíveis nos dados, a mineração de dados é uma etapa desse processo. Neste sentido, Sumathi e Sivanandam (2006, p. 16) definem a mineração de dados, ou descoberta do conhecimento, como o processo assistido por computador de "cavar" e analisar enormes quantidades de dados para extrair informação, predizendo comportamentos e tendências futuras e, possibilitando um comportamento proativo, com base no conhecimento, por parte das organizações.

Embora, frequentemente, os termos Ciência de Dados e Mineração de Dados sejam trocados, o primeiro se refere a um conjunto de princípios fundamentais que orientam a extração de conhecimento a partir dos dados, enquanto a mineração é a extração propriamente dita, realizada por meio da aplicação de tecnologias que incorporam esses princípios (Data Science for Business: What you need to know about data mining and data-analytic thinkingPROVOST; FAWCETT, 2013, p. 2). Sob essa perspectiva, a Ciência de Dados é conceitualmente mais ampla, enquanto a Mineração de Dados corresponde às técnicas que efetivam esse conceito. Mayo (2016) também corrobora com essa abordagem ao dizer que Ciência de Dados é tanto sinônimo de Mineração de Dados quanto é um conjunto de conceitos que engloba a Mineração de Dados. Em comum, ambos conceitos compartilham o *insight* como resultado de um processo, porém a Ciência de Dados engloba outros aspectos como a aquisição e armazenamento, etapas anteriores ao processo de KDD.

Em comparação à *machine learning*, Lantz (2015, p. 3) afirma que toda mineração de dados envolve o uso de algoritmos de aprendizado de máquinas, mas o contrário não é verdadeiro. Como exemplos, o autor cita a aplicação de *machine learning* para minerar os dados de trânsito para identificação de padrões no número de acidentes e o processo de aprendizagem de um carro autônomo. No primeiro caso, *machine learning* e mineração de dados são a mesma coisa. Porém, no segundo exemplo, há a aprendizagem de máquina, mas não a mineração de dados.

2.5.4.4 Inteligência artificial

Para simplificar o conceito de Inteligência Artificial (IA), pode-se defini-la como a capacidade de um software em se aprimorar por conta própria, sem intervenções adicionais dos desenvolvedores, utilizando apenas as interações com dados do mundo real para se tornarem “mais inteligentes” (GOOGLE CLOUD, 2017). Em termos mais gerais, a IA é qualquer coisa que auxilie máquinas na realização de tarefas características da inteligência humana, como aprender com eventos passados, compreender linguagens (naturais), enxergar e reconhecer objetos e imagens. Com esta abordagem, Rawlings-Goos (2019, p. 10) compara tarefas comumente alcançadas por métodos de IA a tarefas realizadas por seres humanos, identificando quatro grupos:

- ***Machine Learning*:** capacidade de aprender com fatos passados, reconhecendo padrões;
- **Visão computacional:** aptidão dos computadores em processar imagens de maneira próxima às funções do cérebro humano.
- **Processamento de linguagem natural:** reconhecimento de texto escrito e falado que permite a extração de *insights* de dados não estruturados.
- **Internet das Coisas:** uso de sensores e dos dados por eles produzidos é uma realidade que possibilita uma infinidade de aplicações de IA.

Em linhas gerais, a IA é tudo aquilo que tenta replicar o funcionamento do cérebro humano em uma máquina, começando pelo pensamento lógico, passando para autocorreção e, enfim, aprendizagem (KAMPAKIS, 2020, p. 4). Para a Ciência de Dados, as aplicações de IA aqui rapidamente apresentadas fornecem inúmeras possibilidades de extração de conhecimento para as organizações além de possibilitarem o acesso a um volume de dados que, até pouco tempo, permanecia inexplorado e inacessível. Todavia, conforme salientado por Mayo (2016), é difícil definir com precisão a relação desse conceito “guarda-chuva” perante a Ciência de

Dados. Mesmo sendo benéfica para as duas áreas, essa relação não é direta, é sempre intermediada por outros elementos, como o próprio *machine learning*.

Assim, procurou-se compreender os conceitos presentes na Figura 7, explorando o quebra-cabeça da Ciência de Dados por Mayo (2016). Mais que isso, esta seção apresenta uma perspectiva histórica e conceitual sobre o tema, apresentando também representações visuais dos elementos da Ciência de Dados. Adicionalmente, após dissertação sobre os motivos da Ciência de Dados ser considerada Ciência, esta seção expõe metodologias e práticas do mercado, distinguindo conceitos e práticas correlatas que causam confusão na terminologia da área.

Na próxima seção, a discussão se concentra no profissional da Ciência de Dados. Quais as competências e papéis que fazem um profissional ser considerado um cientista de dados eficiente. Além do conceito de competência, conhecimentos e habilidades, as propostas educacionais da área são exploradas, assim como são apresentadas recomendações para futuros programas educacionais.

2.6 SÍNTESE DO CAPÍTULO

Este capítulo se destina, fundamentalmente, à exploração do conceito de Ciência de Dados. Inicialmente, é retomado o recurso utilizado por Brandt (2016) e amparado por Giles (2005), de iniciar a investigação acerca do conceito de Ciência de Dados a partir da Wikipedia. Esse procedimento tem o objetivo de identificar padrões e conceitos relacionados ao objetivo principal, mas adicionalmente possui papel introdutório à investigação mais formal que se segue. Desta maneira, parte-se para um levantamento histórico de como o termo “data science” começou a ser difundido e como evoluiu em interesse público, chegando à definição proposta por Cao (2017).

Com base em Chen et al. (2018), apresenta-se uma lista (Quadro 2 – Conceitos de Ciência de Dados) de definições acerca de Ciência de Dados, com posterior exposição de convergências e divergências encontradas entre elas. Dentre as definições apresentadas, encontram-se autores basilares e contemporâneos a esta tese, bem como organizações privadas e públicas. Em seguida, há uma seção dedicada a representações visuais da Ciência de Dados, onde, por meio de

diagramas do tipo Venn, autores e organizações buscam ilustrar conceitos, disciplinas e relações características da área.

Nesta seção, apresenta-se, ainda, a definição de Ciência de Dados adotada por esta tese. Com base em pesquisa realizada na literatura da área, defende-se que o conceito de Ciência de Dados é "a interação interdisciplinar de competências socioculturais, métodos quantitativos, ciência da computação e conhecimento de domínio para a extração metodológica de conhecimento a partir de quantidades massivas de dados".

As duas seções seguintes trazem reflexões sobre a relação da Ciência de Dados com a ciência e a com a interdisciplinaridade. Assim, faz-se um breve levantamento sobre o que é ciência e o papel da Ciência de Dados neste contexto. Além disso, o conceito de interdisciplinaridade é apresentado juntamente com sua relevância para os cientistas de dados. Por fim, o capítulo é concluído com a seção que aborda metodologias para extração de conhecimento a partir de massas de dados, além de apresentar e diferenciar conceitos correlatos à Ciência de Dados.

3 AS COMPETÊNCIAS E A PROFISSÃO DO CIENTISTA DE DADOS

O mais simples conceito de cientista de dados é defini-lo como o praticante da Ciência de Dados (KIM; LEE, 2016, p. 162). No entanto, assim como ocorre para a própria área de atuação, não há uma definição bem estabelecida e consensual desse profissional, especialmente devido à diversidade de competências e habilidades que lhe são atribuídas (DEMCHENKO *et al.*, 2016, p. 621). Ademais, conforme apontado por Kim e Lee (2016, p. 162), o surgimento e evolução dos dois termos não ocorreram de maneira síncrona. Enquanto a Ciência de Dados começa a ganhar relevância na década de 90, despertando reflexões como a de Wu (1997) apresentada na seção 2.1, o uso do termo "cientista de dados" é impulsionado pela publicação de Davenport e Patil (DAVENPORT; PATIL, 2012).

Dentre as definições para cientista de dados, Donoho (2015, p. 5) preza pela simplicidade ao significá-lo como o profissional que emprega métodos científicos para relevar e criar significado a partir de dados brutos. Outra abordagem, focada nas disciplinas da Ciência de Dados, é apresentada pelo Instituto Nacional (Americano) de Padrões e Tecnologia, que define um cientista de dados como

> um profissional que tem conhecimento suficiente para atuar nas disciplinas (e suas interseções): necessidades de negócios, conhecimento de domínio, habilidades analíticas, engenharia de software e sistemas para gerenciar os processos de dados, de ponta a ponta no ciclo de vida dos dados (NIST BIG DATA PUBLIC WORKING GROUP, 2015, p. 8, tradução nossa)

Saltz e Grady (2017, p. 2358) adota uma perspectiva mais pragmática, focada nas funções realizadas por este profissional. Para estes autores, o cientista de dados tem o papel de "traduzir" um problema de negócio para uma questão relacionada a dados, criar modelos preditivos para responder à pergunta gerada e comunicar os resultados por meio de *storytelling*. Ao enfatizar a principal característica do cientista de dados, Linden et al. (2018, p. 7) definem esse profissional como um mestre em habilidades quantitativas, transformando dados em conhecimento que pode solucionar problemas organizacionais que não podem ser atendidos pela engenharia de software tradicional.

Além das competências quantitativas, pode-se destacar a proximidade com tecnologia: o cientista de dados é um profissional especializado em análises

preditivas que, além de possuir habilidades analíticas para entregar informação útil à organização, apresenta proficiência em programação para tornar acessíveis grandes volumes de dados desestruturados e diversificados (BURTCH WORKS, 2019, p. 48). Já Kampakis (2020, p. 19) aborda a formação profissional e enxerga o cientista de dados como alguém que, além de conhecer e trabalhar com análise de dados, possui formação avançada na área, com títulos como doutorado em *machine learning* ou mestrado em Estatística.

Se inicialmente o termo “cientista de dados” era empregado para redefinir o estatístico, um novo perfil profissional surgiu à medida que as organizações reconheciam que, para explorar os benefícios dos dados que vinham se acumulando, era necessário conhecimento especializado e habilidades advindas de outras disciplinas (KIM; LEE, 2016, p. 162). Os profissionais de dados começavam a ser envolvidos em uma série de problemas relevantes, sejam decisões comerciais, direcionamentos políticos e participação no desenvolvimento de políticas públicas (BRANDT, 2016, p. 1). Em busca de status e ganhos salariais, profissionais da área de análise de dados alteravam seus cargos para “cientista de dados”, aproveitando-se da confusão que dominava o mercado (WALKER, 2015, p. 10).

Como demonstração do aumento da relevância e da demanda por profissionais de dados, cita-se o estudo realizado pela Comissão Europeia para mensurar e identificar tendências no mercado de dados europeu. No relatório publicado em 2017, eram estimados 6,1 milhões de profissionais de dados atuando no continente europeu e, com um crescimento médio anual de 14,1%, a previsão para 2020 foi de 10,43 milhões (EUROPEAN COMMISSION, 2017). Por outro lado, se em 2016 havia 420.000 vagas direcionadas aos profissionais de dados que não estavam preenchidas, a previsão para 2020 era que este déficit chegasse a 769.000. Como agravante, nem todos os profissionais de dados possuem as habilidades necessárias à prática da Ciência de Dados. Assim, como a formação de novos profissionais não acompanha o ritmo da geração e do acúmulo de dados, há uma evidente lacuna de profissionais na área (HAYES, 2017).

Naturalmente, as instituições educacionais têm demonstrado interesse na formação deste profissional emergente e numerosos cursos têm sido desenvolvidos para o treinamento e pesquisa em Ciência de Dados (BRANDT, 2016, p. 5). Por outro lado, a formalização educacional da Ciência de Dados e entrada dos profissionais

formados por estas iniciativas no mercado de trabalho são fenômenos relativamente recentes (CURTY; SERAFIM, 2016, p. 308). De qualquer forma, o crescente número de papéis profissionais e de cursos relacionados à Ciência de Dados, sendo muitos oferecidos por instituições tradicionais, demonstra claramente a solidificação de uma nova profissão (CAO, 2019, p. 1).

Mesmo com o aspecto positivo desse fenômeno, faz-se necessário definir diretrizes organizacionais e padrões de desempenho que estabeleçam uma manifesta distinção entre profissionais qualificados e não-qualificados (WALKER, 2015, p. 10). Sobretudo, pelo fato de que más condutas na seleção, coleta, estruturação ou análise de dados trazerem riscos a organizações e indivíduos. Assim, como uma profissão ainda em construção, há a necessidade de se especificar quais habilidades são demandadas e que devem ser desenvolvidas na formação e no decorrer da carreira de cientistas de dados (KIM; LEE, 2016, p. 171).

Como uma nova área, os primeiros profissionais a atuarem como cientistas de dados têm origem em campos tradicionais e estabelecidos. A Física, por exemplo, é considerada uma das áreas que mais fornecem profissionais à Ciência de Dados, principalmente pela fundamentação matemática, habilidades computacionais e forte dependência dos dados (HARRIS; MURPHY; VAISMAN, 2013-, p. 3; LOUKIDES, 2012, p. 14). Por outro lado, profissionais com outros perfis, como advindos da área de Ciências Sociais, também formam eficientes cientistas de dados (HARRIS; MURPHY; VAISMAN, 2013-, p. 3).

De qualquer maneira, Walker (2015, p. 7, tradução nossa) argumenta que a Ciência de Dados deve ser formalizada como profissão pelos mesmos motivos que a Medicina ou o Direito se tornaram profissões: “cada um requer educação, treinamento e habilidade específicos; cada um requer um código de conduta próprio e uma determinação única do que é má prática no domínio específico”. Para definir a Ciência de Dados como profissão, é necessário estabelecer quais competências garantem que o profissional de dados conduza pesquisas e projetos relacionados aos dados , de maneira eficiente e ética (CAO, 2019, p. 2). Na próxima seção, são apresentadas as competências identificadas na literatura que vão nesta direção.

3.1 COMPETÊNCIAS

Antes de abordar especificamente as competências dos profissionais da Ciência de Dados, faz-se necessário abordar o conceito de competência, ainda que não seja o cerne desta pesquisa. Assim, tendo em vista a definição utilizada, pode-se identificar os elementos relevantes na literatura sobre as competências do cientista de dados.

3.1.1 O conceito de Competência

Mesmo sem definir claramente o conceito de competência, McClelland (1973), por meio do seu artigo *Testing for Competence rather than Intelligence*, incitou as discussões sobre o tema ao questionar os testes de inteligência vigentes à época. A visão do autor defendia que, para avaliar a competência de um profissional, os testes deveriam ser direcionados à função exercida e considerar critérios como habilidades comunicativas, paciência, capacidade em definir metas de forma "moderada" e desenvolvimento do ego (MCCLELLAND, 1973, p. 10).

A publicação de McClelland (1973), além de iniciar o debate sobre o tema entre administradores e psicólogos dos Estados Unidos, é precursora da corrente de pensamento americana. Nessa perspectiva, os autores abordam a competência como um conjunto de capacidades humanas, onde o desempenho está relacionado à inteligência e à personalidade dos indivíduos (FLEURY; FLEURY, 1991, p. 185). Ou seja, a competência é o conjunto de recursos que o sujeito possui para realizar seu trabalho. A visão americana apresenta uma abordagem comportamentalista, onde a competência se refere à "capacidade que o indivíduo traz para o trabalho", resultante de treinamento ou da experiência (CAMARGO, 2013, p. 24).

Questionando esse vínculo da competência à qualificação profissional, emerge na década de 1970, na França, o debate que buscou relacionar a competência à capacidade de saber agir, extrapolando a questão meramente técnica (FLEURY; FLEURY, 1991, p. 186). A perspectiva francesa, que se difundiu entre autores europeus de outros países, baseia-se em valores construtivistas e classifica um indivíduo como competente se este "consegue apresentar resultados" em determinado contexto (CAMARGO, 2013, p. 32).

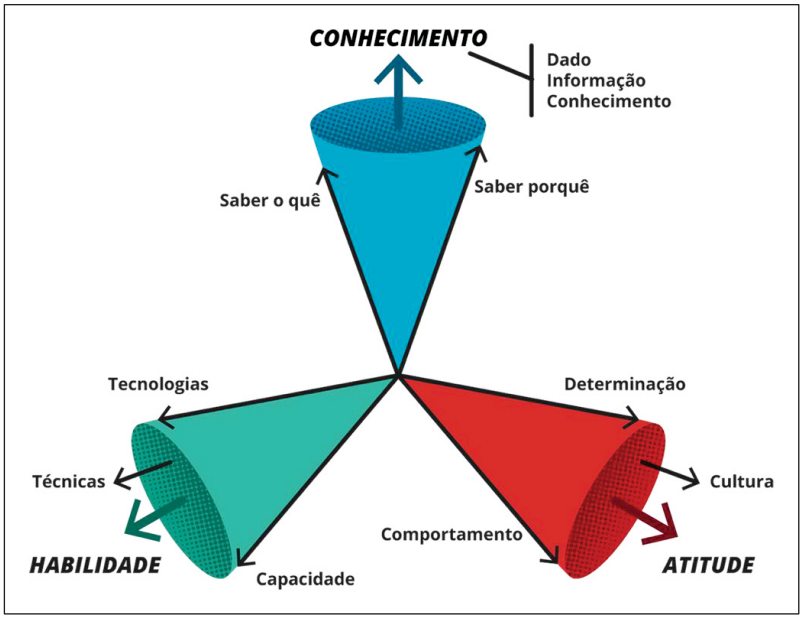
Buscando construir o conceito de competência, Fleury e Fleury (1991) fazem um levantamento cronológico do debate teórico acerca do tema, no cenário acadêmico e no contexto empresarial, ao nível dos indivíduos, das organizações e dos países. Após identificar os principais estudiosos das duas correntes teóricas, americana e europeia, os autores formulam ao seguinte entendimento de competência:

> um saber agir responsável e reconhecido, que implica mobilizar, integrar, transferir conhecimentos, recursos e habilidades, que agreguem valor econômico à organização e valor social ao indivíduo (FLEURY; FLEURY, 1991, p. 188).

Nesta visão, destaca-se que a competência é algo benéfico tanto para a organização quanto para o indivíduo. Ou seja, ao passo que as pessoas desenvolvem competências essenciais para o trabalho, também investem em si mesmas, tornando-se melhores para as organizações, para seus países e, por consequência, para o mundo.

Baseando-se nos princípios da área de Educação, em especial da aprendizagem individual, Durand (2000) define o conceito de competência por três dimensões: Conhecimento, Habilidade e Atitude. Este modelo, que se tornou conhecido como “CHA” é apresentado na Figura 13:

FIGURA 13 – AS TRÊS DIMENSÕES DA COMPETÊNCIA



FONTE: Adaptado de Durand (2000).

Sucintamente, as três dimensões da competência são assim descritas por Durand (2000, p. 8):

- **Conhecimento:** é o conjunto de informações já assimiladas que viabiliza a compreensão do mundo. Abrange o acesso aos dados e a capacidade de reconhecê-los como informação, integrando-os a sistemas existentes que são melhorados a partir destas interações;
- **Habilidade:** capacidade em agir de maneira efetiva em relação aos objetivos e processos definidos. Relaciona-se aos aspectos empíricos e tácitos;
- **Atitude:** refere-se às características comportamentais, até então negligenciadas pelos pesquisadores da área, que são determinantes para que um indivíduo ou uma organização atinjam o que foi estabelecido.

Adicionalmente, Durand (2000, p. 20, tradução nossa) reforça que estas três dimensões são interdependentes e se reforçam “à medida que a aprendizagem ocorre simultaneamente em todas as direções, por meio de informação, ação e interação”. Assim, a construção do conhecimento só ocorre quando na presença da ação (habilidade). Por outro lado, habilidade sem conhecimento é vulnerável, suscetível a riscos. Já o conhecimento sem atitude é infrutífero, bem como atitude sem conhecimento não tem sentido. Finalmente, habilidade sem atitude é inútil, assim como a situação oposta: atitude sem capacidade de realizar é improfícua.

3.1.2 Competências do Cientista de Dados

Em relação às principais competências necessárias para a prática da Ciência de Dados, os pesquisadores apresentam distintas perspectivas. Loukides (2012, p. 16) destaca habilidades interpessoais como paciência, motivação para construir produtos de dados, vontade em explorar e gerar soluções de maneira contínua e incremental. O autor ressalta que a atuação do cientista de dados é inerentemente interdisciplinar, uma vez que este profissional busca constantemente novas

respostas, sempre abordando os problemas de maneira abrangente. Seguindo no âmbito das habilidades interpessoais, classifica a Ciência de Dados como um campo colaborativo e criativo que engloba profissionais de diferentes especialidades, trabalhando conjuntamente para concluir projetos de dados de maneira inovadora.

Porém, sob uma visão mais técnica, Harris, Murphy e Vaisman (2013-, p. 1) declaram que o senso comum sobre as competências primordiais ao cientista de dados indica conhecimentos em Estatística, programação e visualização de dados. Mas os autores afirmam também que as competências necessárias à profissão são muito mais abrangentes: os cientistas de dados devem possuir conhecimento de todo o processo, da extração e integração dos dados, do desenvolvimento de modelos, de análises avançadas para criarem ferramentas escaláveis que auxiliem no processo decisório.

De uma forma mais genérica, Dhar (2013, p. 64) afirma que para exercer a profissão de cientista de dados é necessário um conjunto integrado de competências que envolvem conhecimento matemático, inteligência artificial, *machine learning*, estatística, banco de dados e otimização. Ademais, é necessária experiência na formulação de problemas que possam ser efetivamente sanados. Extrapolando as funções do cientista de dados e incluindo qualquer profissional ligado à área de dados, Davenport et al. (2015, p. 19) ressaltam que competências em Análise de Dados, Gestão da Informação e linguagens de programação, aliadas às habilidades em liderança, promovem a eficiência em liderar, comunicar e colaborar.

Para Hall, Phan e Whitson (2016, p. 7), os cientistas de dados são os profissionais com maior capacidade analítica cujas competências mesclam conhecimento em Ciência da Computação, Matemática e expertise no domínio em questão. Neste sentido, Baškarada e Koronios (2017, p. 66) definem as competências centrais do cientista de dados: 1) conhecimento de Matemática e Estatística, e habilidade para mineração de dados, testes de hipótese e análises preditivas; 2) Ciência da Computação com os conceitos de estruturas de dados e algoritmos e; 3) Conhecimento de Domínio, como contabilidade, economia, gestão, marketing, logística e engenharia de produção. Os autores destacam como habilidades adicionais o carregamento, a integração e a transformação de dados, assim como a prática de visualização de dados.

Sob uma perspectiva histórica, Kim e Lee (2016, p. 169) confirmam a Estatística como a principal competência na Ciência de Dados. O profissional com conhecimentos estatísticos avançados combinados à Ciência da Computação possui o perfil ideal para um candidato a cientista de dados. Anteriormente, o rigor das análises estatísticas já havia sido destacado por Serouss (2014) salientando, porém, que apenas esta competência não basta para a profissão. Sob sua visão, as competências necessárias a um cientista de dados demandam tempo para serem desenvolvidas, visto que este deve compreender o que é necessário no desenvolvimento de sistemas de dados, estando apto a lidar com linguagens de programação, seja para trabalhar nos dados, seja para implementar funcionalidades.

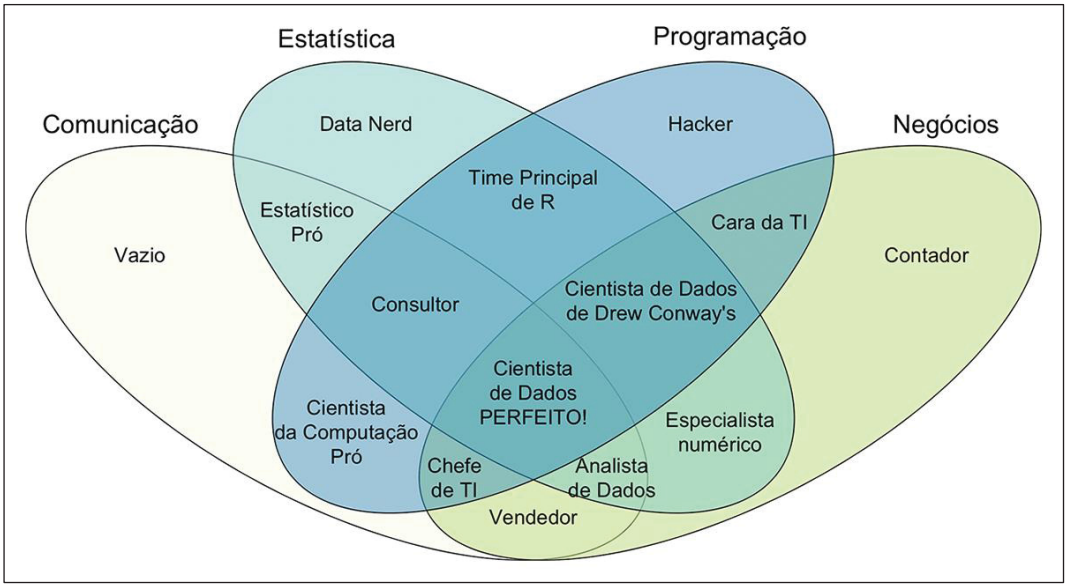
Esta combinação de habilidades em programação e conhecimentos estatísticos fornece uma ampla gama de atuação ao cientista de dados, distinguindo-o de outros profissionais próximos, como estatísticos, analistas de dados ou analistas de BI (KIM; LEE, 2016, p. 170). Já para Demchenko, Belloum e Wiktorski (EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS)2017, p. 15) o que diferencia o cientista de dados de outros profissionais é seu conhecimento em métodos e técnicas de pesquisa. Por mais que nem todos os cientistas de dados sejam cientistas primorosos, estes devem conhecer os métodos gerais de pesquisa como a formulação de hipóteses, elaboração e aplicação de métodos científicos e, validação de hipóteses.

Mas distinguir o cientista de dados de outras profissões não é uma tarefa simples, principalmente porque muitos papéis possuem competências que se sobrepõem (CONWAY, 2010). Essa indefinição, juntamente com a variedade de papéis inerentes à Ciência de Dados, afeta até mesmo os anúncios de vagas de empregos na área, onde frequentemente o termo cientista de dados é associado a qualquer profissional que realize uma atividade envolvendo dados (SALTZ; GRADY, 2017, p. 2355; STODDER, 2015). Não importando se a vaga é para gestão de dados, sistemas de processamento de dados, análise de dados.

Por isso, apesar de toda a popularidade em torno da Ciência de Dados, ainda há uma falta de entendimento sobre a área, seja em relação aos papéis (cientista de dados, analista de dados, engenheiro de dados, especialista de negócio ou engenheiro de software), seja sobre as habilidades (expertise do domínio, tecnologia da informação e conhecimentos quantitativos) (BAŠKARADA; KORONIOS, 2017, p.

65). Para tentar facilitar essa compreensão, Stephan Kolassa (2014), doutor e especialista em Ciência de Dados, propôs um diagrama com as competências necessárias à área e os papéis resultantes das combinações das mesmas. O resultado da proposta é apresentado na Figura 14:

FIGURA 14 – AS COMPETÊNCIAS E OS PAPÉIS NA CIÊNCIA DE DADOS



FONTE: Adaptado de Kolassa (2014).

Propositalmente, a proposta de Kolassa (2014) é mais elaborada e mais detalhista que sua referência, o primeiro diagrama da Ciência de Dados proposto por Conway (2010). Se esta proposta é baseada na experiência pessoal e profissional do autor, há outras perspectivas que buscam explorar e formalizar as competências e os papéis do cientista de dados. A seguir, procura-se apresentar estas abordagens para ampliar os aspectos conceituais e metodológicos desta seção.

3.1.3 Conhecimentos e habilidades

Inicialmente, para abordar os conhecimentos e habilidades do cientista de dados, apresenta-se o conceito deste profissional contido no *European Skills, Competences, Qualifications and Occupations* (ESCO) ou, em português, “Uma taxonomia das qualificações, competências e profissões europeias”. Este tesauro, proposto pela Comissão Europeia, identifica, descreve e classifica ocupações profissionais, habilidades e qualificações relevantes ao território europeu, com

benefícios para o mercado de trabalho e para o sistema educacional (What is ESCOESCO, 2020). Pela atualização de agosto de 2020, o ESCO apresenta as descrições de 13.485 habilidades vinculadas a 2.942 ocupações, traduzidas em 27 idiomas. A descrição em português da ocupação de cientista de dados é

> Os cientistas de dados pesquisam e interpretam fontes de dados enriquecidos, geram grandes quantidades de dados, procedem à fusão de fontes de dados, asseguram a consistência dos conjuntos de dados e criam visualizações para ajudar na compreensão dos dados. Constroem modelos matemáticos utilizando os dados, apresentam e comunicam informações e conclusões relativas aos dados aos especialistas e cientistas da sua equipa e, se necessário, a um público não especializado, e recomendam formas de aplicar os dados (Cientista de DadosESCO, 2020).

A ocupação de cientista de dados está assim precedida na hierarquia do tesauro:

2 - Especialistas das atividades intelectuais e científicas

25 - Especialistas em tecnologias da informação e comunicação

251 - Analistas e programadores de software e aplicações

2511 - Analistas de sistemas

2511.3 – Cientista de dados

E, por fim, para realizarem as tarefas que lhes são atribuídas, as seguintes habilidades e conhecimentos são considerados essenciais:

QUADRO 3 – HABILIDADES E CONHECIMENTOS ESSENCIAIS AO CIENTISTA DE DADOS

Habilidades essenciais	**Conhecimentos essenciais**
• Construir sistemas de recomendação • Coletar dados • Entregar apresentação visual dos dados • Projetar esquemas de banco de dados • Desenvolver aplicações de processamento de dados • Estabelecer processos de dados • Executar cálculos analíticos • Manipular amostras de dados • Implementar processos de qualidade de dados • Interpretar os dados • Gerenciar sistemas de coleta de dados • Normalizar dados • Executar limpeza dos dados • Reportar os resultados de análises	• Mineração de dados • Modelagem de dados • Categorização da informação • Extração da informação • Processamento de análises *online* • Linguagens Query • Linguagem RDF (*resource description framework*) • Estatística • Técnicas de apresentação visual

FONTE: Desenvolvido com base em ESCO (Cientista de Dados2020).

Mesmo diante de seu formalismo, característico de um tesauro, e de sua amplitude, visto que a taxonomia engloba mais de 1.200 profissões, o ESCO é considerado uma fonte de referência adequada, principalmente, pelos conhecimentos essenciais apresentados estarem em consonância com as disciplinas apontadas pela literatura da área. Há de se considerar, no entanto, que é destinada à realidade europeia e sua aderência ao mercado brasileiro, logicamente, carece de verificação.

A abordagem de Dhar (2013) se concentra na relação do cientista de dados com as aplicações de análises preditivas. Para o autor, a construção de sistemas de tomada de decisão automática depende da acurácia dessas aplicações, fazendo a aprendizagem de máquina uma das competências mais requisitadas pelo mercado de Ciência de Dados. Neste contexto, as habilidades essenciais ao cientista de dados estão divididas em três classes:

a) **Estatística:** principalmente a estatística Bayesiana, que requer um conhecimento prático de probabilidade, mas também distribuições, teste de hipótese e análise multivariada;
b) **Ciência da computação:** representação e manipulação dos dados pelos computadores. Nesta classe, estão as habilidades em estruturas de dados, algoritmos, sistemas, bancos de dados, computação distribuída, computação paralela e computação tolerante a falha;
c) **Correlação e causalidade:** envolve o conhecimento do domínio para estabelecer relações causais. Mesmo que os dados observados, geralmente, limitem as análises à correlação, em algumas situações, diante de uma certa quantidade de dados e com o devido conhecimento do domínio, é possível calcular probabilidades condicionais para estabelecer um modelo com estrutura causal.

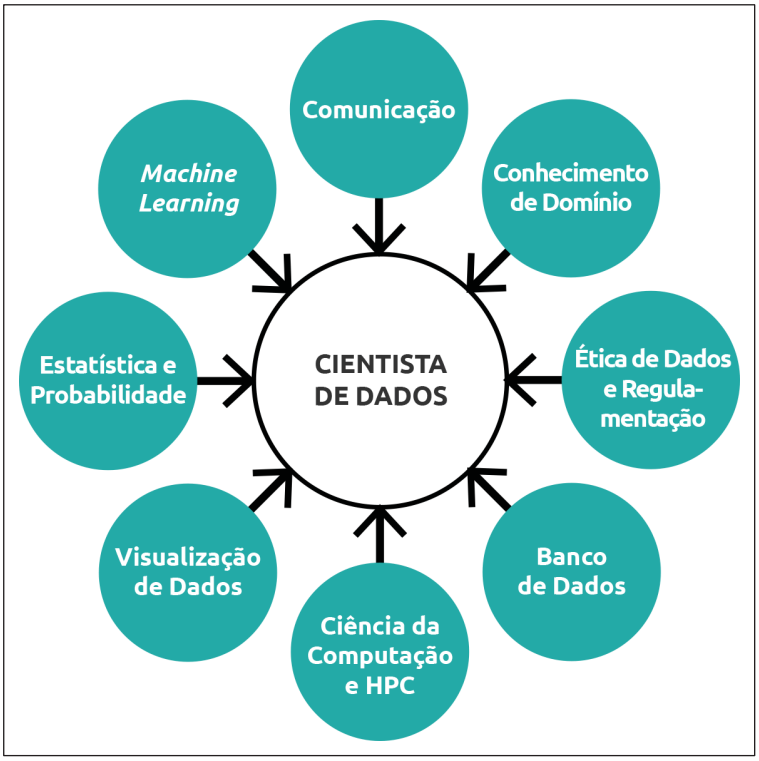
Em seus apontamentos, Cao (2017, p. 31) define os conhecimentos, habilidades e atitudes que fazem do cientista de dados um bom profissional:

- Pensar analítica, criativa e criticamente. A inquietação perante as possibilidades é um atributo necessário;
- Metodologias e conhecimento em sistemas e abordagens complexas para conduzir a resolução de problemas, quer seja do tipo *top-down*, quer seja do tipo *bottom-up*;
- Mestrado ou doutorado em Ciência da Computação, Estatística, Matemática, Análise de Dados, Ciência de Dados, Informática, Engenharia, Física, Pesquisa Operacional, reconhecimento de padrão, IA, Visualização, Recuperação da Informação ou áreas afins;
- Um profundo entendimento de metodologias e modelos em estatística, mineração de dados e *machine learning*;
- Habilidade em implementar, manter e solucionar problemas na infraestrutura relativa aos dados, como computação em nuvem, infraestrutura de alto desempenho, paradigmas de processamento distribuído, fluxo de processamento e bancos de dados;
- Conhecimento de interação humano-computador, visualização, representação e gestão;
- Fundamentos em engenharia e qualidade de software;
- Experiência em lidar com grandes conjuntos de dados que mesclem diferentes tipos e fontes de dados, em um ambiente de rede e distribuído;
- Experiência em extração e processamento de dados, compreensão de recursos e análise de relacionamento;
- Conhecimento e interesse ativo em estudos e métodos, multi e transdisciplinares, nas áreas científicas, técnicas e sociais;
- Experiência substancial e atualizada em relação a *scripts* orientados a análises de dados, estruturas de dados, linguagens de programação e plataformas de desenvolvimento em Linux, na nuvem ou em ambiente distribuído;
- Fundamentação teórica e conhecimento do domínio para avaliação dos resultados das análises, tanto para benefícios técnicos quanto de negócio;

- Por fim, excelente comunicação verbal e escrita, além de habilidades organizacionais, com habilidade de comunicar os resultados obtidos a diferentes tipos de públicos.

Em uma abordagem mais abrangente e abstrata, Kelleher e Tierney (2018) afirmam que o papel do cientista de dados se tornou tão amplo que o debate acerca das competência que lhe cabem ainda está encerrado. Contudo, os autores atestam que é possível estabelecer uma lista com os itens com os quais a maioria dos envolvidos concorda que são relevantes para a área. A lista proposta por Kelleher e Tierney (2018), apresentada na Figura 15, é composta por oito competências que representam aspectos importantes de um projeto em Ciência de Dados. Os autores, no entanto, frisam que é raro um profissional dominar todas as oito competências, sendo que o mais usual é o cientista de dados ser especializado em um subconjunto deste arranjo. De qualquer forma, todos os envolvidos devem ter consciência e entendimento da contribuição de cada uma delas.

FIGURA 15 – O DESIDERATO CONJUNTO DE COMPETÊNCIAS DO CIENTISTA DE DADOS



FONTE: Adaptado de Kelleher e Tierney (2018).

Dentre os aspectos não citados anteriormente, destaca-se a competência referente à Ética e Regulação dos Dados. Os dados são o cerne de qualquer projeto de Ciência dos Dados. Entretanto, Kelleher e Tierney (2018) reafirmam que por mais que as organizações tenham acesso aos dados, estas não estão habilitadas a um uso irrestrito deste recurso. As legislações acerca da proteção de dados pessoais estão se expandindo, regulando e controlando o uso dos dados. Cabe ao cientista de dados, além de entender estes regulamentos, desenvolver uma compreensão ética das implicações de seu trabalho para utilizar os dados de maneira legal e apropriada.

Na sequência, são apresentadas pesquisas que exploram os papéis dos profissionais na Ciência de Dados, relacionando-os com as competências requeridas.

3.1.4 Os papéis do cientista de dados

Para iniciar a exploração dos papéis em Ciência de Dados, primeiramente, destaca-se a pesquisa realizada por Harris, Murphy e Vaisman (2013-), cujos detalhes são apresentados na seção 4.2.1. Neste momento, é pertinente identificar como os autores distinguem os papéis dos cientistas de dados, listando suas principais competências e combinando os resultados destes dos conjuntos de dados. Inicialmente, os respondentes deveriam indicar seu grau de concordância perante 11 questões iniciadas com “Eu vejo como um X”, onde X era substituído por um dos papéis levantados pelos autores. Com os dados coletados, foram identificados quatro grupos de autoidentificação para os cientistas de dados, conforme a Figura 16:

FIGURA 16 – AUTOIDENTIFICAÇÃO DO CIENTISTA DE DADOS

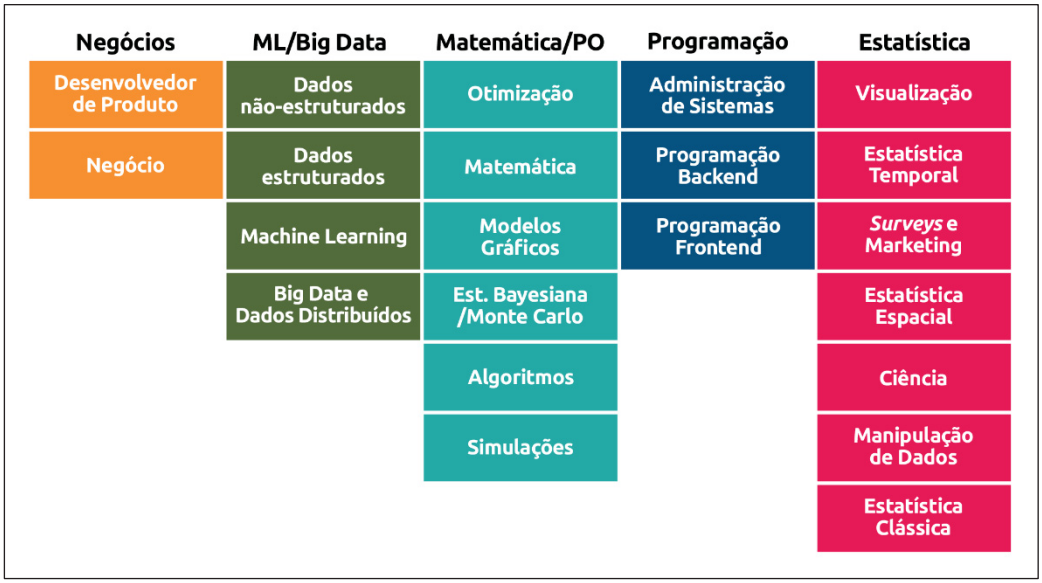
FONTE: Adaptado de Harris, Murphy e Vaisman (2013-).

Assim, com a definição dos quatro grupos pela autoidentificação do cientistas de dados, Harris, Murphy e Vaisman (2013-) definem as característica de cada um que aqui são sintetizadas:

- **Desenvolvedor de Dados:** corresponde aos profissionais cujo principal foco são questões técnicas do gerenciamento dos dados: como obter, armazenar e aprender com este recurso. Seu trabalho diário é escrevendo código, principalmente com sistemas em ambiente de produção;
- **Pesquisador de Dados:** nesta categoria estão os profissionais com formação acadêmica avançada, onde 75% já publicaram artigos em periódicos revisados por pares e mais de 50% possuem doutorado. As organizações reconhecem o valor da visão acadêmica na resolução de problemas complexos, mesmo que os domínios dos negócios sejam bastantes distintos dos campos científicos clássicos;
- **Criativo de Dados:** o grupo mais abrangente da Ciência de Dados e onde majoritariamente estão os profissionais mais jovens. São rotulados como “pau para toda obra” e se destacam pela aplicação de uma ampla gama de ferramentas e tecnologias a um problema, sempre buscando soluções inovadoras;
- **Pessoa de negócios (*Businessperson*) de Dados:** aquele que se classifica como líder e empreendedor, mais propensos a gerenciar pessoas e trabalhar como consultor ou mesmo ter sua própria empresa.

Em seguida, os cientistas de dados entrevistados deveriam ordenar, de acordo com sua afinidade e experiência, 22 competências agrupadas em cinco grupos apresentados na Figura 17:

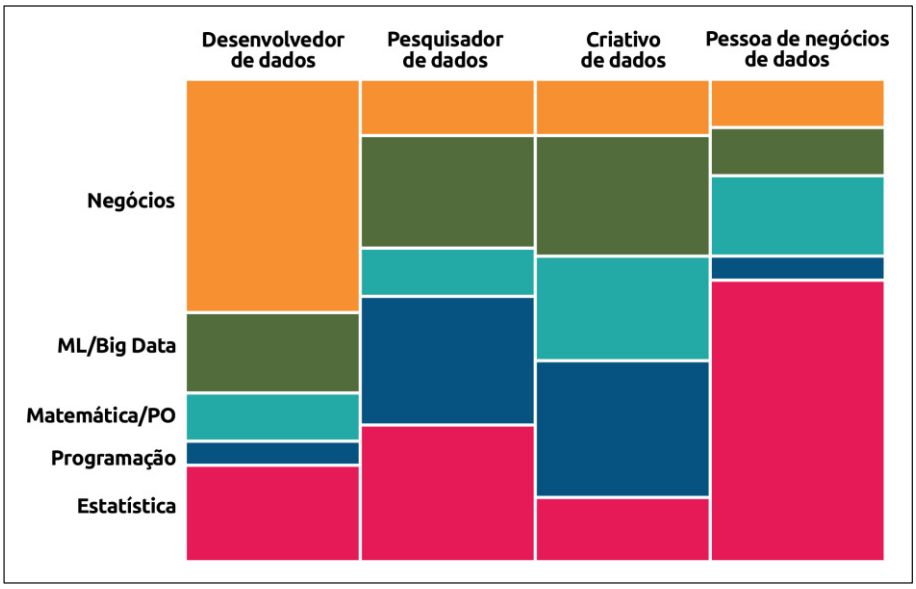
FIGURA 17 – GRUPOS DE COMPETÊNCIAS DOS CIENTISTAS DE DADOS



FONTE: Adaptado de Harris, Murphy e Vaisman (2013-).

As competências eram mostradas em ordem aleatória para os respondentes e, para evitar desentendimentos, apresentavam exemplos de cada item e quando compiladas e distribuídas pelos grupos de autoidentificação geraram o resultado exposto na Figura 18:

FIGURA 18 – COMBINAÇÃO ENTRE COMPETÊNCIAS E PAPÉIS



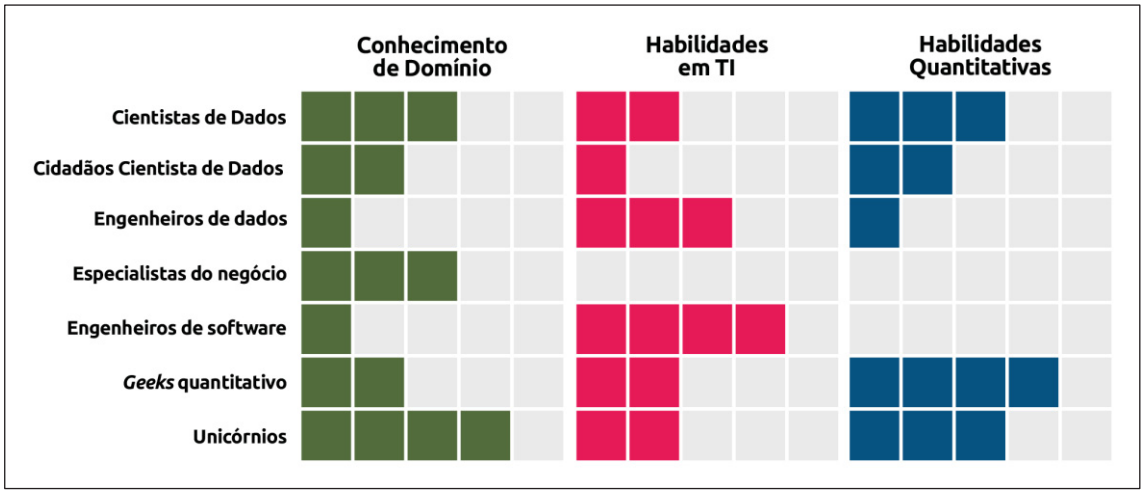
FONTE: Adaptado de Harris, Murphy e Vaisman (2013-).

A figura confirma resultados esperados como as competências para os negócios serem o grupo mais desenvolvido pelas pessoas de negócios de dados e

os pesquisadores de dados apresentarem profundo conhecimento em estatística e matemática. Os outros dois grupos, desenvolvedores e criativos, são aqueles que apresentam maior diversidade entres os grupos de competências, configurando-se como aqueles com maior capacidade para ML/Big Data e linguagens de programação (HARRIS; MURPHY; VAISMAN, 2013-, p. 13–14).

Para viabilizar a prática da Ciência de Dados, conforme a concepção proposta pela Gartner (apresentada na Figura 5), Linden et al. (2015) elencam os papéis que os integrantes de uma equipe devem desempenhar. Embora nem toda equipe tenha condições de permanentemente contar com um profissional de cada papel, é necessário entender que essas necessidades estão presentes nos estágios dos diferentes projetos realizados em uma organização. O mapeamento dos papéis perante as três disciplinas da Ciência de Dados é mostrado na Figura 19.

FIGURA 19 – MAPEAMENTO DAS HABILIDADES E PAPÉIS NA CIÊNCIA DE DADOS



FONTE: Adaptado de Linden et al. (2015).

Linden et al. (2015) assim descrevem os papéis identificados:

- **Cientistas de Dados:** integrantes fundamentais para uma equipe de dados, têm experiência em métodos de extração de conhecimento dos dados e possuem visão geral de todo o processo.
- **Cidadãos Cientista de Dados (ou Analista de Dados):** não possuem formação em Ciência de Dados, mas ainda podem executar uma variedade de tarefas "mais simples" de um projeto de dados;

- **Engenheiros de dados:** responsáveis pelo acesso e disponibilização dos dados aos integrantes do projeto, focados na infraestrutura e produtividade da equipe;
- **Especialistas do negócio:** indivíduos que possuem vasto conhecimento do domínio, sendo um líder do negócio ou um especialista da área;
- **Especialistas em sistema fonte:** possuem conhecimento dos dados no nível da aplicação do negócio;
- **Engenheiros de software:** requeridos esporadicamente, quando é necessária codificação personalizada;
- ***Geeks* quantitativo:** profissionais cuja principal característica é seu profundo conhecimento em métodos quantitativos. Para os projetos em dados estes conhecimentos ora são desejáveis, ora são mandatórios;
- **Unicórnios:** são cientistas de dados versados em toda a gama de habilidades. Eventualmente, são romantizados pela literatura e são extremamente raros.

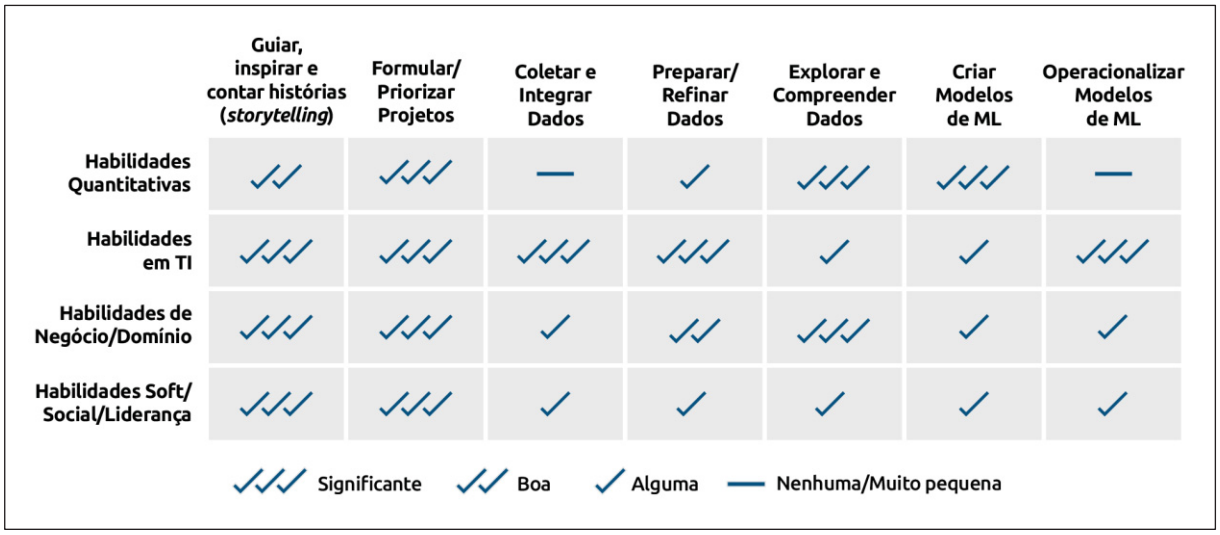
Em 2018, outra publicação pela Gartner, cujo principal autor também é Alexander Linden, apresenta uma abordagem alternativa à montagem de equipe de dados. Nesta proposta posterior, Linden et al. (2018) afirmam que, em projetos de Ciência de Dados, é necessário equilibrar habilidades e experiência em quatro áreas chave:

- **Análise quantitativa:** engloba competências matemáticas, treinamento e educação vinculados às disciplinas de Ciência de Dados, incluindo estatística, *machine learning*, pesquisa operacional e processamento de sinais;
- **Arquitetura e TI:** habilidades em tecnologias da área como infraestrutura tecnológica, preparação dos dados, engenharia de dados, operacionalização de *machine learning* e governança analítica de dados;

- **Domínio do negócio:** conhecimento do setor e das funções do negócio, estratégia e visão de negócio, requisitos regulatórios e legais relevantes.
- **Social:** competência interpessoal que estimulem a colaboração, liderança, comunicação e trabalho em equipe.

Com foco nestas quatro áreas, Linden et al. (2018, p. 4) estabelecem as tarefas gerais que são essenciais para o sucesso das iniciativas de dados: (1) Guiar, Inspirar e *Storytelling*; (2) Formular e priorizar projetos; (3) Coletar e integrar os dados; (4) Preparar e refinar os dados; (5) Explorar e compreender os dados; (6) Construir modelos de ML e; (7) Operacionalizar os modelos de ML. Assim, os autores estabelecem uma matriz estes as áreas e as tarefas, definindo o grau de relevância que cada competência possui para determinada tarefa (Figura 20).

FIGURA 20 – EQUILÍBRIO ENTRE HABILIDADES E EXPERIÊNCIA PARA EQUIPES DE DADOS



FONTE: Adaptado de Linden et al. (2018).

Assim, ao se estabelecer as tarefas, criando um consenso interno sobre elas, a composição da equipe é definida com a especificação das habilidades e a experiência necessárias para realizá-las. Nesta proposta, Linden et al. (2018, p. 8) instituem seis papéis fundamentais para a Ciência de Dados, sendo o de cientista de dados subdivido em três categorias (júnior, pleno e sênior), e indicam o nível de competência que lhes é necessária em relação a cada uma das áreas inicialmente definidas. Na Figura 21, a matriz entre as áreas de competência e os papéis contidos nas equipes de dados indica o nível necessário para cada função:

FIGURA 21 – COMPETÊNCIAS NECESSÁRIAS POR PAPEL DESEMPENHADO

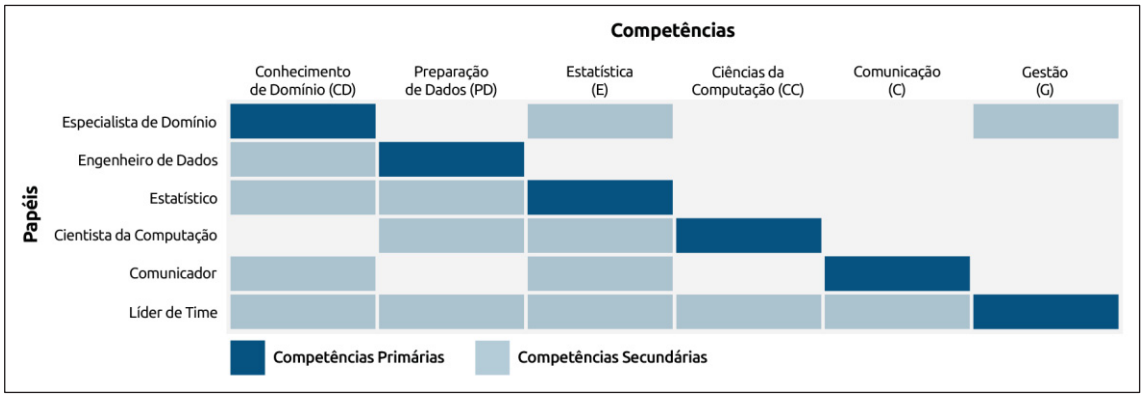
FONTE: Adaptado de Linden et al. (2018).

De qualquer maneira, Linden et al. (2018) recomendam que a estrutura proposta é uma referência e seu conteúdo deve ser adaptado para cada necessidade. É necessário criar uma fundamentação das tarefas críticas ao projeto, relacionando as capacidades da equipe aos quatro conjuntos de competências descritos na pesquisa. Também, é sugerido que haja uma otimização na combinação das tarefas necessárias aos papéis, adequando o nível de competência e experiência dos integrantes da equipe ao trabalho que deve ser realizado. Por fim, é aconselhado que outros profissionais alheios à Ciência de Dados, como especialistas em software ou hardware, por exemplo, integrem as equipes mesmo que na função de suporte. Assim, a diversidade de competência do conjunto é expandida.

A combinação entre competências e papéis requeridos à Ciência de Dados também é objeto da pesquisa de Baškarada e Koronios (2017). Seis competências são definidas como necessárias para projetos em dados: Conhecimento do Domínio, Preparação dos Dados, Estatística, Ciência da Computação, Comunicação e Gestão. Porém, os Baškarada e Koronios (2017, p. 67) “falham” na tentativa de encontrar um profissional “unicórnio”, ou seja, aquele que possui expertise, comparada aos especialistas, em distintas disciplinas dentre as pesquisadas. No lugar de um cientista de dados polivalente, os resultados apontam para a formação de equipes

multidisciplinares. Assim, os autores chegam a seis papéis fundamentais para a Ciência de Dados que quando cruzados com as competências apresentam o mapeamento contido na Figura 22:

FIGURA 22 – PAPÉIS, COMPETÊNCIAS PRIMÁRIAS E SECUNDÁRIAS



FONTE: Adaptado de Baškarada e Koronios (2017).

Além destes seis papéis, Baškarada e Koronios (2017, p. 69) propõem decompor as habilidades dos profissionais em primárias e secundárias, uma vez que nenhum dos integrantes de uma equipe de dados trabalha de maneira isolada. Assim, marcada em preto, a diagonal principal da matriz contida na Figura 22 apresenta a habilidade principal para cada papel. As demais células pintadas em cinza, destacam as habilidades secundárias que variam em quantidade de acordo com a quantidade de interação que um papel exercido apresenta com os demais. O líder do time, por exemplo, possui sua habilidade primária situada na disciplina de Gestão, mas deve possuir fluência em todas as demais, uma vez que é o profissional que interage com todos os integrantes da equipe.

No mesmo ano, Parks (2017, p. 9) afirma, em sua tese, que as competências do cientista de dados podem ser agrupadas em cinco áreas chave, que são assim sintetizadas:

- **Funções em Análise de dados:** abrangem métodos matemáticos e estatísticos, além de habilidades em modelagem necessárias para resolver problemas com dados;

- **Programação e Ferramentas:** habilidades técnicas necessárias para obtenção de dados, que englobam o domínio, ou mesmo a criação, de ferramentas de dados;
- **Técnicas e soluções em *Machine Learning*:** métodos utilizados para revisar dados e construir soluções em ML, com ou sem supervisão;
- **Conhecimento interdisciplinar:** representam o conhecimento abordado em disciplinas acadêmicas como contabilidade, economia, sistemas de informação, saúde;
- **Habilidades críticas:** habilidades interpessoais que viabilizem que os resultados dos dados sejam sumarizados e comunicados. Incluem técnicas de comunicação e visualização.

Por meio de três fontes de dados, revisão da literatura, pesquisa de levantamento e entrevistas, Parks (2017) identificou cincos áreas principais e 55 competências entre os praticantes da Ciência de Dados. A lista destas competências, vinculadas à área que as engloba, contém desde disciplinas acadêmicas clássicas a ferramentas e métodos específicos e é apresentada no Quadro 4

QUADRO 4 – COMPETÊNCIAS DO CIENTISTA DE DADOS

(continua)

Áreas	**Competências**
Funções em Análise de dados	• Estatística • Matemática Básica • Cálculo • Análise Multivariada • Álgebra linear e formas quadráticas • Programação inteira • Correlação de Person, MLlib, Funções Lambda ou Chi-Quadrado • Testes de Significância • Desvio padrão
Programação e Ferramentas	• Programação em R • Programação em Python • Excel • VBA • Java, C, C++ e HTML • SQL • SPARQL • TensorFlow • Radiant • Tableau/Qlikview • SAS • Data Wrangling • Hadoop/Tessera

(conclusão)

Técnicas e soluções em Machine Learning	• Árvores de Decisão • Métodos ordinários • Redes neurais • Vetores • Clustering • Análise de Componentes Independentes • Processamento de Linguagem Natural • Apache • Amazon Machine Learning • Azure ML, Caffe, H2O, Massive, MLIB mlpack, Pattern, Shogun, Torch, Tensorflow
Conhecimento interdisciplinar	• Contabilidade • Economia • Programação de Computadores/Sistemas de informação • Marketing • Tomada de Decisão
Habilidades críticas	• Gestão de Dados/Governança • Gestão de Banco de Dados (técnica) • Pensamento Estratégico • Habilidade em formular perguntas “inteligentes” • Organização (dos dados, de conceitos, de prioridades) • Visualização de Dados • Comunicação Escrita • Comunicação Verbal • Relações Interpessoais • Intuição para dados • Pensamento Crítico/Lógico • Curiosidade • Habilidades em Hacking • Habilidades científicas • Habilidades em Análise Quantitativa • Conselheiro de confiança • Especialista do Negócio/Especialista do Domínio • Foco na precisão

FONTE: Adaptado de Parks (2017).

Para Parks (2017), os grupos “Funções em Análise de dados” e “Programação e Ferramentas” são os pilares das competências da Ciência de Dados. No primeiro, competências em Estatística, Matemáticas e métodos relacionados são considerados fatores fundamentais para o sucesso dos projetos de dados. No segundo, é identificada uma discrepância entre o mercado e a literatura acadêmica, especialmente em relação às ferramentas citadas, como Tensorflow, Tableau, Qlikview e SAS. A conclusão do autor é que estas ferramentas ainda não foram incorporadas à cultura acadêmica, mesmo tendo sido tópicos recorrentes nas entrevistas com os profissionais.

Em relação ao *machine learning*, enquanto os acadêmicos destacam a importância dos métodos desse grupo para o cientista de dados, os entrevistados ressaltam que no mercado, eles ainda se encontram em estágio de desenvolvimento.

Em contrapartida, o conhecimento interdisciplinar é destacado em todos os aspectos da pesquisa, diferentemente do que ocorre para as habilidades críticas. A pesquisa de Parks (2017) revela que a capacidade de formular perguntas inteligentes para problemas de dados, habilidades interpessoais, pensamento crítico e curiosidade são tópicos recorrentes nas entrevistas com especialistas, mas não abordados pela literatura especializada. Outros aspectos da pesquisa de Parks (2017) são explorados na seção 4.1.2 desta tese.

O projeto europeu EDISON é outra iniciativa que procurou identificar as competências e os papéis necessários à prática da Ciência de Dados. A finalidade do projeto era estabelecer a profissão de cientista de dados e, para isso, buscava o alinhamento entre as necessidades do mercado com as trilhas de carreiras disponíveis (EDISON PROJECT, 2017). Neste sentido, o principal resultado do projeto foi a elaboração do EDISON Data Science Framework (EDSF) cujo objetivo era prover uma base para a construção de currículos eficazes em Ciência de Dados, potencializando todo o ecossistema de oferta e demanda da área (DEMCHENKO *et al.*, 2016).

O EDSF é composto por um conjunto de documentos, distribuídos gratuitamente, que se destinam ao desenvolvimento de educadores, empregadores e gestores, bem como os próprios cientistas de dados. Dentre estes documentos, o primeiro, e fundamental para o desenvolvimento dos demais, destina-se a formalizar as competências do cientista de dados.

O *Data Science Competences Framework* (CF-DS) apresenta uma visão abrangente das competências exigidas por múltiplas definições da Ciência de Dados, atendendo às demandas por análise de dados e engenharia de softwares, sempre com foco em organizações e métodos orientados a dados. Este contexto exige habilidades avançadas em gestão de dados heterogêneos e uso de métodos de pesquisa para revelar todo o valor dos dados (EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS)DEMCHENKO; BELLOUM; WIKTORSKI, 2017). Conforme o CF-DS, há cinco grupos de competências que fornecem uma base para demarcar de forma consistente programas de educação, treinamentos e certificação para profissionais relacionados à Ciência de Dados. As definições desses grupos, revisadas por especialistas da área reunidos em grupos ou individualmente, são assim apresentadas (EDISON Data Science Framework: Part 1.

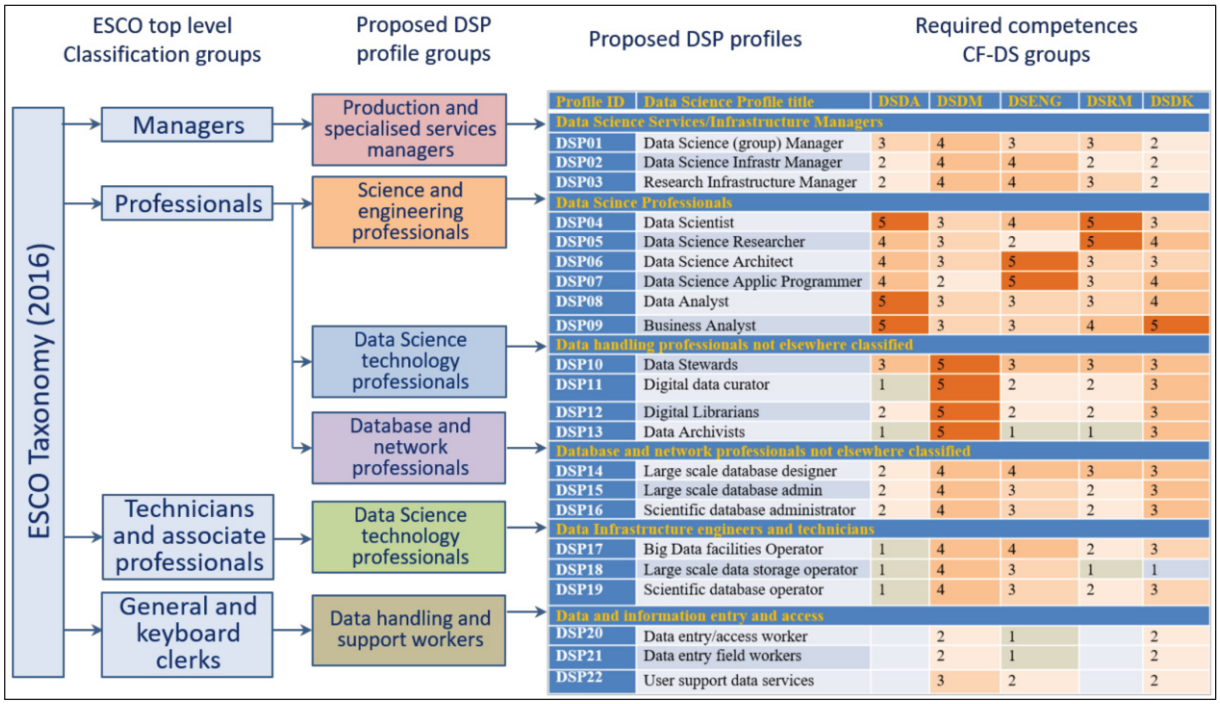
Data Science Competence Framework (CF-DS)DEMCHENKO; BELLOUM; WIKTORSKI, 2017, p. 16):

- **Análise em Ciência de Dados (DSDA):** utiliza análise de dados e técnicas estatísticas sobre os dados para descobrir novas relações e fornecer *insights* sobre problemas de pesquisa ou processos organizacionais, de modo a apoiar a tomada de decisão. Inclui, além de métodos estatísticos, soluções em *machine learning*, mineração de dados e algoritmos;
- **Engenharia em Ciência de Dados (DSENG):** adota princípios da engenharia e tecnologias computacionais modernas para pesquisar, projetar e implementar novas aplicações de análise de dados. Desenvolve experimentos, processos, instrumentos, sistemas, infraestruturas para apoiar o manuseio de dados durante todo o ciclo de vida dos dados;
- **Gestão de Dados e Governança (DSDM):** desenvolve e implementa a estratégia de gestão de dados para coleta, armazenamento, preservação e disponibilidade de dados para processamento posterior;
- **Métodos de Pesquisa e Gestão de Projetos (DSRM):** cria capacidades e entendimentos utilizando o método científico (hipótese, teste, validação) ou métodos de engenharia para descobrir novas abordagens para criar conhecimentos e alcançar objetivos de pesquisa ou organizacionais;
- **Conhecimento de domínio ou Expertise (DSDK):** emprega o conhecimento da área (científico ou comercial) para desenvolver relevantes aplicações de análise de dados.

No CF-DS, cada grupo de competência é subdivido em seis competências específicas que demandam diferentes perfis profissionais, com habilidades próprias à função exercida. Estas 36 competências são identificadas por meio de um código alfanumérico, tornando-se facilmente referenciadas em qualquer um dos documentos do EDSF.

Dentre os documentos do EDSF, existe um específico para discriminar os papéis dos profissionais que de alguma forma lidam com os dados. O *Data Science Professional Profiles* (DSPP) se baseia na taxonomia ESCO, utilizando quatro grupos de profissionais: 1) Chefia e Direção (*Managers*); 2) Especialistas das atividades intelectuais e científicas (*Professionals*); 3) Técnicos e profissões de nível intermédio (*Technicians and associate professionals*) e; 4) Empregados de escritório (*General and keyboard clerks*). Os perfis propostos para a Ciência de Dados, bem como suas relações com os grupos de competências são demonstrados na Figura 23:

FIGURA 23 – PERFIS PROFISSIONAIS DA CIÊNCIA DE DADOS



Profile ID	Data Science Profile title	DSDA	DSDM	DSENG	DSRM	DSDK
Data Science Services/Infrastructure Managers						
DSP01	Data Science (group) Manager	3	4	3	3	2
DSP02	Data Science Infrastr Manager	2	4	4	2	2
DSP03	Research Infrastructure Manager	2	4	4	3	2
Data Scince Professionals						
DSP04	Data Scientist	5	3	4	5	3
DSP05	Data Science Researcher	4	3	2	5	4
DSP06	Data Science Architect	4	3	5	3	3
DSP07	Data Science Applic Programmer	4	2	5	3	4
DSP08	Data Analyst	5	3	3	3	4
DSP09	Business Analyst	5	3	3	4	5
Data handling professionals not elsewhere classified						
DSP10	Data Stewards	3	5	3	3	3
DSP11	Digital data curator	1	5	2	2	3
DSP12	Digital Librarians	2	5	2	2	3
DSP13	Data Archivists	1	5	1	1	3
Database and network professionals not elsewhere classified						
DSP14	Large scale database designer	2	4	4	3	3
DSP15	Large scale database admin	2	4	3	2	3
DSP16	Scientific database administrator	2	4	3	2	3
Data Infrastructure engineers and technicians						
DSP17	Big Data facilities Operator	1	4	4	2	3
DSP18	Large scale data storage operator	1	4	3	1	1
DSP19	Scientific database operator	1	4	3	2	3
Data and information entry and access						
DSP20	Data entry/access worker		2	1		2
DSP21	Data entry field workers		2	1		2
DSP22	User support data services		3	2		2

FONTE: Adaptado de Demchenko (2017).

Expandindo os grupos do ESCO, no DSPP são propostos seis grupos para profissionais ligados aos dados que, ao final, resultam em 22 perfis com diferentes níveis de formação e responsabilidades. A combinação do framework de competências (CF-DS) à especificação dos perfis (DSPP), assim como os demais documentos do EDSF, são apresentados como uma ferramenta de *benchmarking* para a Ciência de Dados, para uso individual ou organizacional, para a montagem de equipes de dados ou para a criação de programas de educação para profissionais da área.

Para explorar os múltiplos papéis atribuídos aos cientistas de dados, Rawlings-Goss (2019, p. 7) explora o conceito do generalista em dados (*Data Generalist*). Para a autora, esses profissionais têm como principal característica sua polivalência, mas exercê-la apresentam uma ampla formação, são solucionadores de problemas rápidos e preparados para atuar em qualquer estágio do ciclo de vida de um projeto de dados. Há uma alta demanda sobre este perfil profissional, especialmente por organizações que conduzir diferentes tipos de projetos de dados e contam com uma equipe de dados com poucos integrantes. Rawlings-Goss (2019, p. 7) afirma que os generalistas são comumente tratados como de cientistas de dados, especialmente por estarem preparados para solucionar qualquer tipo de problema de dados.

Porém, as distintas atuações desse profissional geraram um conjunto de cargos que buscam especificar estas atividades, dos quais o autor ressalta:

- **Cientistas de Dados (solucionadores de problemas):** correspondem aos profissionais com perfil de cientista, uma vez que adotam método científico para diagnosticar um problema, determinar a metodologia e as ferramentas para resolvê-lo.
- **Analistas de Dados (tradutores):** aplicam várias técnicas de análises aos dados financeiros e operacionais, visando melhorar a produtividade e o desempenho das organizações. Estes profissionais têm o papel de traduzir os resultados estatísticos para uma linguagem acessível aos tomadores de decisão;
- **Arquiteto de Dados (construtores):** projetam a infraestrutura necessária para garantir que a organização possa coletar, integrar, armazenar e gerenciar dados advindos de diversas fontes, especialmente ferramentas *online*, ajudando no sucesso dos negócios digitais;
- **Engenheiros de Dados (testadores):** com conhecimento em engenharia de software e padrões de teste, aliado à experiência em codificação, este profissional implementa os projetos do arquiteto de dados, criando soluções robustas que mantêm os dados seguros e em perfeito funcionamento;

- **Administradores de Banco de Dados:** responsáveis pela tarefa de armazenar e realizar *backup* das informações organizacionais, em espaços físicos ou virtuais. Estipulam questões de segurança da informação, controlando o acesso aos dados e garantindo a restauração destes em caso de ocorrências não previstas;
- **Analistas de Negócio (comunicadores):** perfil menos técnico da equipe de dados, mas com profundo conhecimento nos processos administrativos que levam ao sucesso empresarial. Une o lado técnico à estratégia do negócio, propondo soluções que efetivem os objetivos de maneira bem-sucedida;
- **Gerente de Dados e Análise (*coaches*):** responsável por garantir que as pessoas certas sejam contratadas, bem como as metas e prioridades sejam corretamente definidas a todas as partes. Depende de habilidades sociais desenvolvidas, pois precisam lidar com diferentes perfis profissionais, mas também necessitam de conhecimento técnico para analisar e validar as descobertas dos dados.

Sobre as funções e cargos, Rawlings-Goss (2019, p. 7) reforça que existem outras especialidades que não foram citadas e que cada papel exige uma ênfase única em relação às competências. Os profissionais de cada subconjunto da Ciência de Dados fornecem sua própria perspectiva sobre a metodologia usada para pegar suas habilidades e transformá-las em soluções de dados, para pequenas e grandes organizações.

Diante do crescente entusiasmo e da multiplicação dos papéis dos cientistas de dados, Cao (2019, p. 35) ressalta o crescimento na oferta de cursos de formação da profissão, sejam do tipo livre, sejam de graduação de instituições tradicionais. Por outro lado, o autor salienta que há por parte do mercado uma constante queixa sobre disponibilidade limitada de profissionais qualificados, capazes de implementar um plano estratégico em relação aos dados. Neste cenário, onde o mercado educacional e o mercado profissional demonstram agir com pressa, há uma urgente necessidade de padronização e formalização da profissão, bem como da educação em Ciência de Dados.

Para contribuir com esse fim, Cao (2019, p. 36) ressalta as principais funções do cientista de dados: 1) entendimento de problemas complexos; 2) identificação e especificação de requisitos e limitações; 3) compreensão e quantificação das características dos dados; 4) conversão situações desafiadores para problemas analíticos; 5) planejar e projetar estratégias analíticas; 6) conduzir a exploração dos dados; 7) avaliar e otimizar os resultados analíticos; 8) extrair dos dados valor e *insights*; 9) comunicar e interpretar os resultados aos *stakeholders* e; 10) operacionalizar a exploração dos dados.

Para tais tarefas, são definidas as seguintes competências: 1) mentalidade de cientista da dados; 2) liderança científica; 3) doutorado nas áreas relacionadas; 4) capacidade em trabalhar com sistemas complexos e resolução de problemas; 5) sólidos fundamentos em estatística, análise de dados e aprendizagem; 6) experiência prática em computação; 7) atuação colaborativa, habilidade organizacional e comunicacional e; 8) conhecimento interdisciplinar e experiência em múltiplos domínios (CAO, 2019, p. 36).

De acordo com Cao (2019, p. 39), existem cursos de elevado nível para a formação do cientista de dados, em sua maioria *online*, de instituições acadêmicas ou não. Porém, a qualidade é altamente variável e poucos são os programas que oferecem uma abordagem sistemática para que o profissional atenda às necessidades estratégicas da economia dos dados. Assim, explorando a questão exposta pelo referido autor, a próxima seção apresenta abordagem e modelos na educação da Ciência de Dados.

3.2 EDUCAÇÃO PARA A CIÊNCIA DE DADOS

Como relatado, com a expansão do volume de dados criados e os avanços da tecnologia, há também um crescente déficit nas formação de profissionais preparados para o trabalho com dados (EUROPEAN COMMISSION, 2017; FINZER, 2013, p. 1; HAYES, 2017). Esta demanda é evidenciada pela criação de programas de educação especializados na área, em níveis de graduação, pós-graduação ou mesmo cursos livres, todos com o objetivo de capacitar os profissionais à prática em Ciência de Dados. Porém, por mais que estas iniciativas tenham efeito positivo, ainda há a necessidade de se estabelecer diretrizes educacionais e padrões de desempenho

capazes de definir quais profissionais são suficientemente qualificados para trabalhar como cientistas de dados (WALKER, 2015, p. 10).

Essa demanda por profissionais em dados, faz com que as próprias empresas criem iniciativas para descobrir e preparar talentos para atuarem em seus projetos de dados, especialmente por meio de parcerias com universidades (CISCO IT INSIGHTS, 2016). Além da formação de novos profissionais, estes programas buscam desenvolver uma cultura analítica que permeia todas as áreas e atividades da organização. A PwC (2017), por exemplo, é uma organização privada que, por meio da publicação de relatórios, orienta instituições educacionais sobre as demandas do mercado a fim de orientar a formulação de seus currículos.

Dentre as principais recomendações, está que, para extrair o máximo valor dos dados, é necessária uma abordagem multidisciplinar na solução de problemas, combinado Ciência de Dados a habilidades analíticas, com expertise do domínio, criatividade e liderança. Para transcender as habilidades técnicas, os currículos devem incorporar uma variedade de experiências que envolvam dar e receber *feedback*, resolução de problemas e conflitos, habilidades em comunicação e apresentação (DAVENPORT *et al.*, 2015, p. 18). Ademais, uma força de trabalho diversificada, com maior participação de minorias e mulheres nas equipes de tecnologia, é a chave para o sucesso beneficiando equipes, organizações, clientes e demais *stakeholders* (PWC, 2017).

Para um contínuo desenvolvimento, especialmente em um contexto onde os ciclos de vida das tecnologias são curtos, a educação em Ciência de Dados depende de uma combinação efetiva entre teoria, prática e contato com o local de trabalho (DEMCHENKO; COMMINIELLO; REALI, 2019, p. 124). Porém, uma vez que o exercício da profissão de cientista de dados não exige uma formação específica ou atende a alguma regulamentação, há no mercado inúmeras propostas de trilhas de aprendizagem que são essencialmente técnicas (CHANDRASEKARAN, 2013; METWALLI, 2020).

Aspectos como a ética na profissão de cientista de dados, que ganharia ênfase com a formalização da profissão (WALKER, 2015), ainda são pouco trabalhados pelas instituições educacionais. Curty e Serafim (2016), ao analisarem o conteúdo programático de 93 programas americanos em Ciência de Dados, constatam que apenas três deles ressaltam aspectos éticos e uso responsável dos dados. Um código

de conduta, para estes que têm acesso a dados sensíveis, envolvendo questões relativas à identidade e à privacidade, deveria ser amplamente difundido entre os praticantes da Ciência de Dados.

Cao (2019, p. 39), ao analisar cursos sobre Ciência de Dados, identifica uma série de lacuna ainda a serem preenchidas pelas IES, conforme as observações:

a) os cursos não ensinam os alunos a pensar sobre dados, ou seja, não estimulam uma mentalidade para a Ciência de Dados;
b) cursos verdadeiramente transdisciplinares são poucos, a maioria dos cursos abordam as disciplinas de forma desconexa, ministradas por diferentes departamentos;
c) a qualidade dos cursos é muito variada, impactando no avanço do conhecimento e, consequentemente, das competências;
d) as capacidades de pesquisa e inovação para solução de problemas baseada em dados estão faltando, mesmo sendo essenciais para conduzir projetos de dados na realidade;
e) há relevantes discrepâncias entre os complexos desafios do mundo real e o avanço do conhecimento proporcionado pela maioria dos cursos disponíveis.

Para o autor, se a Ciência de Dados emergiu como uma nova profissão da qual seus praticantes exercem um papel de destaque na sociedade, a educação desse profissional deveria habilitá-lo a, por meio dos dados, aumentar a produtividade, expandir a ciência, proporcionar o desenvolvimento e impulsionar a economia por meio de sua atuação. Para guiar sua proposta de educação do cientista de dados, Cao (2019, p. 39) propõe os seguintes questionamentos: 1) O que define a próxima geração de cientistas de dados qualificados? 2) Quais são as lacunas e problemas existentes nos cursos atuais de Ciência de Dados? 3) O que está disponível para criar os cientistas de dados da próxima geração? e 4) Como capacitar os alunos a lidar com desafios desconhecidos e conhecimentos indisponíveis a fim de criarem os conhecimentos necessários? A fim de contribuir com as reflexões de Cao (2019), a presente pesquisa buscou por modelos e propostas de educação para a Ciência de Dados, apresentando-os a seguir.

3.2.1 Propostas para a Educação de Cientistas de Dados

Cleveland (2001) foi o primeiro autor a elaborar uma proposta para a educação em Ciência de Dados, desenvolvendo um planejamento de como integrar a Ciência da Computação e a Matemática à Estatística. A sugestão curricular era composta por seis áreas técnicas, descritas e distribuídas seguindo as porcentagens:

- **(25%) Investigações multidisciplinares:** colaboração em análises de dados em áreas temáticas, domínios específicos.
- **(20%) Modelos e métodos de dados:** modelos estatísticos, métodos de construção de modelos, métodos de estimativa e distribuição baseados em inferência probabilística.
- **(15%) Computação com dados:** sistemas de hardware, sistemas de software e algoritmos computacionais.
- **(15%) Pedagogia:** planejamento curricular e abordagens educacionais para o ensino fundamental, médio, superior, pós-graduação, educação continuada e treinamento corporativo.
- **(5%) Avaliação de ferramentas:** análise de ferramentas utilizadas, necessidades percebidas de novas ferramentas e estudos de processos para desenvolvimento de novas ferramentas.
- **(20%) Teoria:** fundamentos da Ciência de Dados, abordagens gerais de modelos e métodos, computação com dados, educação e avaliação de ferramentas, avaliação matemática de modelos e métodos, computação com dados e didática.

As motivações para as mudanças vêm dos benefícios que as áreas técnicas da Estatística podem proporcionar aos analistas de dados, diante de mais recursos. A computação voltada a dados, que acabara de ser reconhecida pelos cientistas da computação, precisava se desenvolver e as abordagens multidisciplinares também se mostravam cada vez mais valorizadas. Outro aspecto considerado foi que formando da área estatística ingressavam na docência sem qualquer contato com a pedagogia durante sua formação. Por fim, a avaliação rigorosa das ferramentas e o desenvolvimento constante de novas soluções já se caracteriza como parte da rotina

do cientista de dados marcada pelo aprendizado contínuo (CLEVELAND, 2001, p. 22).

Já nos anos posteriores a 2010, Piety, Hickey e Bishop (2014) chamaram de Ciências de Dados Educacionais *(Educational Data Sciences)* o *framework* que desenvolveram para organizar as emergentes atividades educacionais que envolviam dados. Esta proposta, portanto, não é focada no cientista de dados, mas destinada a qualquer contexto no qual os dados são utilizados para informar os aprendizes e para a gestão da aprendizagem. Bastava que o contexto educacional adotasse tecnologias digitais voltadas à coleta, compartilhamento e representação de vastos volumes de informação. As áreas citadas pelos autores são: Análise Acadêmica/Institucional, Análise de Aprendizagem/Mineração de Dados Educacional, Análise/Personalização do Aluno e Melhoria Instrucional Sistêmica.

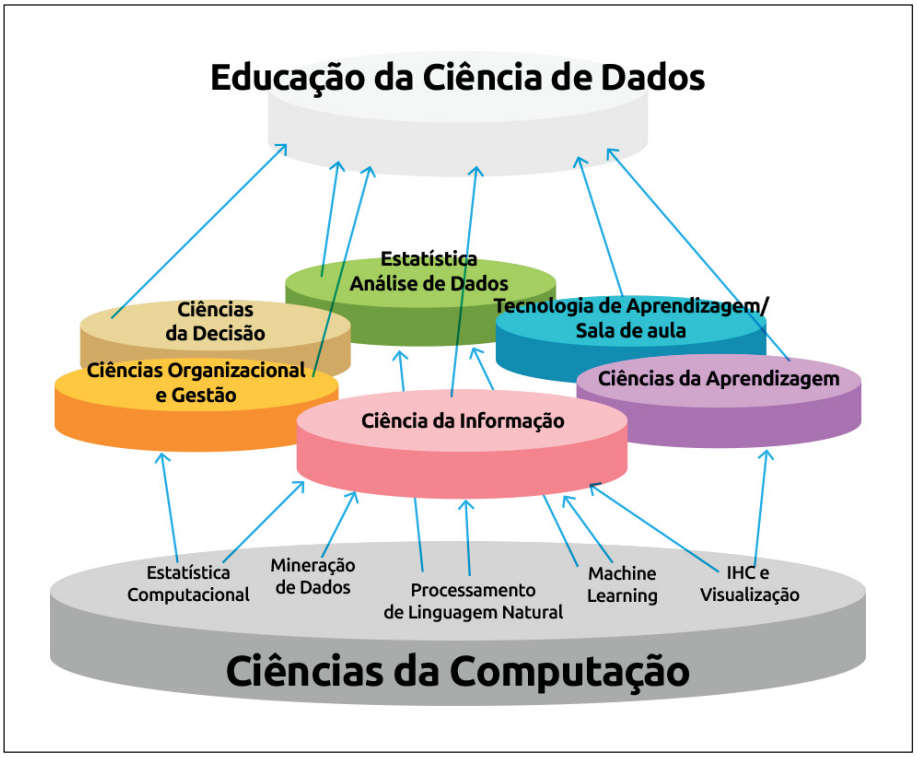
Assim, para evidenciar as peculiaridades da área Ciências de Dados Educacionais, os autores elencam cinco características:

- **Rápida evolução:** em poucos anos, a área apresentou significativo crescimento, passando de pequenas iniciativas a um amplo contexto, recebendo apoio de governos e fundações interessantes em incentivar um crescimento sociotécnico;
- **Problemas para demarcar fronteiras:** uma vez que a área cresce dentro de ecossistemas de comunidades acadêmicas já estabelecidas, as semelhanças e inter-relações dificultam a definição de distinções;
- **Disrupção nas práticas avaliativas:** diferentes tipos de informação que não estavam disponíveis proporcionam novas maneiras de se abordar a aprendizagem dos alunos e os próprios processos educacionais;
- **Visualização, interpretação e cultura:** surgem questões sobre a representação da informação e a popularização de *dashboards*, utilizados para ordenar e classificar elementos a fim de evidenciar o significado extraído de vastos e diversificados conjuntos de dados;
- **Ética, privacidade e arquitetura da informação:** questões de como a coleta e o uso de informações sobre alunos e professores podem ser

feitos de forma responsável, garantindo a privacidade daqueles indivíduos cujas informações são capturadas.

Ademais, Piety, Hickey e Bishop (2014, p. 199) para demonstrar as conexões entre as disciplinas componentes das Ciências de Dados Educacionais, propõem um modelo que é apresentado na Figura 24:

FIGURA 24 – CONEXÕES INTERDISCIPLINARES NA EDUCAÇÃO DA CIÊNCIA DE DADOS



FONTE: Adaptado de Piety, Hickey e Bishop (2014).

Sob esta perspectiva, para compreender as Ciências de Dados Educacionais é necessário abordar suas relações com outros campos. No modelo de Piety, Hickey e Bishop (2014, p. 199) a Ciência da Computação é base de todo o arranjo, uma vez que é a "fornecedora" de tecnologias e métodos para soluções em dados. Seus produtos potencializam as seis áreas (Ciência da Informação, Ciência da Aprendizagem, Ciência das Organizações e da Gestão, Tecnologia de Aprendizagem, Ciências da Decisão e Análise de Dados/Estatística) que, quando relacionadas, originam as Ciências de Dados Educacionais, um campo que promovem novas abordagens e soluções de educação em um contexto cada vez mais dinâmico (PIETY; HICKEY; BISHOP, 2014, p. 201).

Voltando à formação do cientista de dados, Anderson et al. (2014) relatam a experiência de 10 anos da implementação de um currículo de graduação com duração de quatro anos, englobando análise preditiva, *machine learning* e mineração de dados. Com base na Matemática, Estatística e Ciência da Computação, o currículo combinava teorias e conceitos a ferramentas e técnicas associadas à Ciência de Dados. Em tópicos, o conteúdo do currículo é assim descrito por

1) **Grandes conjuntos/Fluxo de dados:** criação/projeto, acesso, limpeza, análise, organização e visualização de dados;
2) **Banco de dados:** projeto, armazenamento, consulta e modelagem;
3) **Técnicas de Inteligência Artificial:** algoritmos genéticos, redes neurais, redes bayesianas, agentes inteligentes, *machine learning*, identificação de padrões, pesquisa heurística, representação do conhecimento e ontologias, processamento de linguagem natural;
4) **Software e Algoritmos:** projeto, programação, teste, algoritmos e análise;
5) **Recuperação de informação:** teoria da informação, mineração de dados, mineração de texto, mineração de imagem, indexação, análise de conteúdo, processamento linguístico, abstração, pesquisa e recuperação, filtragem de informação, formulação de consulta;
6) **Matemática:** lógica e métodos quantitativos, estruturas discretas, estatística, álgebra linear, modelagem e simulação;
7) **Comunicação oral e escrita:** aspectos da comunicação eficaz;
8) **Questões sociais, éticas e legais:** privacidade e segurança, propriedade, política, validação de informações, profissionalismo.

O currículo foi colocado em prática no curso de graduação da Faculdade de Charleston, nos Estados Unidos, em 2005, fornecendo à instituição um segundo curso interdisciplinar ligado à computação e atraindo alunos de alto desempenho devido ao seu rigor metodológico. Dentre os benefícios proporcionados pela experiência, Anderson et al. (2014) destacam a integração proporcionada pelo relacionamento desenvolvido entre os professores de disciplinas distintas, facilitado pelo tamanho reduzido do corpo docente da instituição. Muito do sucesso da

implementação do projeto se deve ao apoio fornecido pela administração em todos os níveis, proporcionando maior confiança e senso de parceria entre os envolvidos.

Além disso, o novo diploma fornecido atraiu novas matrículas para a IES, gerando alunos envolvidos em iniciação científica, focados em análises de dados. Anderson et al. (2014) ressaltam que o currículo passa por avaliações constantes que buscam prover melhorias ao programa educacional que vão desde gerar maior flexibilidade aos discentes a alinhamentos estratégicos em relação aos objetivos de aprendizagem. De qualquer forma, para os autores, a maior contribuição da iniciativa é mostrar que a Ciência de Dados pode ser o principal objeto de um curso de graduação.

No ano seguinte, Hardin et al. (2015) analisam o ensino da Ciência de Dados nos currículos dos cursos de Estatística. Os autores contextualizam a pesquisa informando que o número de estatísticos graduados mais que dobrou no período de 2008 a 2013, impulsionado pela demanda do mercado por analistas de dados. Exercendo este papel, é esperado que o profissional de estatística saiba utilizar bancos de dados (incluindo *data warehouses*), capturar dados da internet, desenvolver soluções complexas em múltiplas linguagens de programação e ter fluência em algoritmos como tem em modelos estatísticos. Todavia, esses tópicos não fazem parte da maioria dos cursos da área.

Mesmo assim, os autores relatam que os cursos de Estatística têm expandido os limites programáticos e aumentado o conteúdo relacionado a tecnologias voltadas aos dados e habilidades comunicativas. Desta forma, propiciam aos novos formandos competências necessárias para manipular grandes volumes de dados derivados do fenômeno *big data*. Para avaliar como os programas de Estatística estão inserindo os conceitos e métodos da Ciência de Dados em seus currículos, Hardin et al. (2015) descrevem sete iniciativas que buscam melhorias para atender estas novas demandas.

Por isso, os autores afirmam que a incorporação dos preceitos da Ciência de Dados aos currículos da Estatística proporciona aos docentes a oportunidade de abordarem métodos, mentalidade e práticas das soluções mais modernas em relação aos problemas com dados, incentivando os alunos atuais, além de atrair novos candidatos. Neste sentido, os tópicos considerados pela pesquisa são concentrados em cinco áreas principais: 1) Programação: estruturada, eficiente e computação de

alto desempenho; 2) Tecnologia de Dados: sistema de gerenciamento de banco de dados relacional, expressões regulares, XML, comandos Shell e *web scraping*; 3) Estrutura de dados: vetores, dados textuais e limpeza de dados; 4) Fluxo de Trabalho: reprodutibilidade, implementação na web e controle de versão e; 5) Estatística: simulações, métodos modernos e visualização de dados.

O sucesso dos casos estudados demonstra que esses tópicos podem e devem ser inseridos nos currículos de Estatística. Todavia, Hardin et al. (2015, p. 351) comentam que o conteúdo evidencia a heterogeneidade na educação fundamental dos estudantes, onde o contato com a computação é muito variado. Esse aspecto, demonstrado pelas práticas adotadas, pode se transformar em desafio para as instituições que planejam modernizar sua grade curricular.

Por outro lado, Baškarada e Koronios (2017, p. 72) apontam como problema à educação superior a tentativa de produzir profissionais do tipo "unicórnio". Cursos de graduação em Ciência de Dados se concentram em visualização de dados, manipulação/conversão de dados, estatística computacional, *machine learning*, análise espacial, mineração de dados, *big data*, mas também há aqueles que abordagem habilidades como comunicação, sociais e éticas como citado por Anderson et al. (2014). Mas dada a multidisciplinaridade da área, buscar um profissional proficiente em todos os tópicos é ilusório e pouco eficiente. A recomendação é que as equipes sejam multidisciplinares, formadas por profissionais com distintos papéis e especialidades. Para Baškarada e Koronios (2017, p. 72) , sem uma profunda expertise em pelo menos um dos papéis propostos pelo seu *framework* (ver Figura 22), um profissional, mesmo que graduado, não terá condições de contribuir efetivamente para uma equipe de Ciência de Dados.

Tendo em vista os papéis do cientista de dados definidos em seu documento sobre competências, o projeto EDISON elabora uma base curricular denominada *Data Science Model Curriculum* (MC-DS). Esta proposta conecta todos os componentes do EDSF em ferramenta cujo objetivo é auxiliar universidades e organizações de treinamento profissionalizante na criação de novos programas, para que estes englobem as competências e conhecimentos associados a cada perfil profissional definido pelo framework (EDISON Data Science Framework: Part 3. Data Science Model Curriculum (MC-DS)DEMCHENKO; BELLOUM; WIKTORSKI, 2017, p. 36).

O MC-DS é apresentado com um modelo que a se adotado em diversos perfis de instituições educacionais, podendo ser adaptado a necessidades específicas, considerando conceitos como: Alinhamento e Coerência, Escopo, Sequência, Continuidade e Integração (Ibidem, 2017b, p. 15). Baseado na abordagem de educação por competências, o projeto de currículo também pode ser adotado por instituições que utilizem outras práticas educacionais como o alinhamento construtivo, aprendizagem baseada em problemas e/ou projetos, e a Taxonomia de Bloom.

A partir de uma perspectiva da Ciência da Computação, Tosic e Beeston (2018) dissertam sobre as oportunidades e desafios na construção de novos cursos de graduação que são multidisciplinares e dependendo de apoio institucional e instrucional de diferentes departamentos. Um currículo em Ciência de Dados exige faculdades e recursos da Ciência da Computação, Matemática, Estatística, Negócios, dentre outras disciplinas "não técnicas", como Comunicação e Filosofia, referente a questões éticas. Assim, é fundamental que haja uma comunicação efetiva, respeitosa e cooperativa entre os profissionais envolvidos, sempre prezando pelos limites entre as disciplinas tradicionais, ainda que historicamente isto não ocorra (RAWLINGS-GOSS, 2019, p. 34; TOSIC; BEESTON, 2018, p. 2).

Apesar das dificuldades, estas relações vão ao encontro das demandas do mercado de trabalho. Programas e cursos interdisciplinares, como os de Ciência de Dados, são os melhores exemplos de como ajustar o ensino superior onde as gerações atuais e futuras de estudantes universitários recebem educação, desenvolvem competências e se encaixam nas oportunidades do século 21. Ademais, o envolvimento ativo nos aspectos educacionais e de pesquisa desses novos programas acadêmicos, embora desafiador, também tende a ser profissionalmente recompensador para os docentes, além de muito apreciado pelos estudantes (TOSIC; BEESTON, 2018).

3.2.2 Recomendações para a Educação de Cientistas de Dados

Diante da consolidação da Ciência de Dados, Cao (2019) elenca quesitos necessários para um desenvolvimento sistemático e de qualidade da profissão, bem como da educação, dos cientistas de dados das próximas gerações. Inicialmente, são recomendadas pesquisas de mercado frequentes, assim como a prática de

benchmarking, para mensurar a qualidade e a satisfação em relação aos serviços de dados, bem como identificação de lacunas, agilizando o desenvolvimento de melhorias para as empresas e formação dos profissionais.

Em relação à educação, Cao (2019) recomenda a colaboração, seja regional ou global, no desenvolvimento de diretrizes e currículo que diminuam as discrepâncias e aumentem a experiência positiva entre instituições, regiões e países. O autor ressalta ainda a necessidade de iniciativas educacionais que promovam o conhecimento interdisciplinar, proporcionando aos estudantes experiências em domínios distintos. Por fim, é ressaltada a demanda pode novas estratégicas, planos de ensino e programas educacionais que aprimorem as qualificações e competência dos profissionais existentes, mas que formem o cientista de dados da próxima geração.

Nos Estados Unidos, a Academia Nacional de Ciência, Engenharia e Medicina (*The National Academies of Sciences, Engineering, and Medicine*) diante da necessidade em produzir uma força de trabalho com conhecimento necessário para gerenciar, analisar e extrair conhecimento dos dados, seja na indústria privada, governamental ou acadêmica, criou um comitê para estabelecer uma visão geral da emergente disciplina Ciência de Dados, especialmente no nível de graduação. No relatório resultado da iniciativa, é ressaltando que nas próximas décadas a consciência e as competências fundamentais em Ciência de Dados serão benéficas a todos os alunos de graduação, não importante o curso no qual atuam (COMMITTEE ON ENVISIONING THE DATA SCIENCE DISCIPLINE, 2018).

Ainda é destacado que a valorização dos dados exige uma população alfabetizada em relação à área, além de um quadro substancial de graduados com competências específicas e conhecimento sólido na Ciência de Dados. Durante um período de aproximadamente um ano, o comitê americano avaliou estratégias para refinar a infraestrutura educacional e administrativa da atuação e da formação de cientistas de dados, com o objetivo de criar oportunidades de desenvolvimento profissional e atender às rápidas e constantes mudanças impostas ao segmento. As recomendações do relatório são aqui expostas:

1) **Desenvolvimento do corpo docente:** As instituições acadêmicas devem abraçar a Ciência de Dados como um campo vital que requer

preparação às especificidades da área. É fundamental o desenvolvimento de um corpo docente preparado para lecionar as disciplinas envolvidas;

2) **Percursos variados:** deve ser desenvolvida e fornecida uma variedade de percursos educacionais para preparar os alunos para uma série de funções de responsabilidade do cientista de dados no local de trabalho;
3) **Alfabetização de dados:** preparar seus graduados para esta nova era baseada em dados. As instituições acadêmicas devem encorajar o desenvolvimento de uma compreensão básica da Ciência de Dados em todos os alunos de graduação;
4) **Ética de dados:** um tópico que, dada a natureza da Ciência de Dados, os alunos devem aprender e praticar ao longo de sua formação. As instituições acadêmicas devem garantir que a ética esteja inserida no currículo desde o início e perdurando por todo o curso;
5) **Código de Ética:** a comunidade da Ciência de Dados deve adotar um código de ética, sendo que tal código deve ser firmado, e revisado com frequência, por membros de sociedades profissionais, incluído em programas e currículos de desenvolvimento profissional e educacional;
6) **Pontes institucionais:** as instituições com cursos de bacharelado e tecnológicos devem estabelecer fóruns para o diálogo sobre todos os aspectos da educação, treinamento e desenvolvimento da força de trabalho em Ciência de Dados;
7) **Diversidade do programa:** à medida que os programas de Ciência de Dados se desenvolvem, devem se concentrar em atrair alunos com origens e graus de preparação variados, preparando-se para o sucesso em carreiras variadas.
8) **Evolução constante:** ainda que seja o início da educação superior da Ciência de Dados, as instituições acadêmicas devem estar preparadas para desenvolver programas flexíveis que passíveis de melhorias incrementais ao longo do tempo;
9) **Educação Contínua:** durante o desenvolvimento de programas de Ciência de Dados, as instituições devem fornecer suporte para que o

corpo docente possa se tornar mais ciente dos diversos aspectos da área por meio de discussão, disciplinas compartilhadas entres diferentes professores, compartilhamento de materiais, cursos de curta duração e outras formas de treinamento.

10) **Avaliação construtiva:** deve-se garantir que os programas sejam continuamente avaliados além de trabalharem em conjunto para desenvolver novas abordagens de avaliação, buscando a excelência na educação. É necessário estabelecer relações com setores específicos do mercado para ajudar a avaliação do ensino perante os impactos dos profissionais recém-formados.
11) **Redes de relacionamento:** as associações profissionais devem se organizar para promoverem encontros regulares sobre a prática da Ciência de Dados entre seus membros. A revisão e discussão por pares são essenciais para compartilhar ideias, melhores práticas e dados.

Em relação ao conteúdo curricular, o relatório vai ao encontro das pesquisas aqui exploradas, ressaltando, porém, a necessidade de uma progressão dos tópicos e conjuntos de habilidades que guiarão os estudantes a desenvolverem sua proficiência em relação aos dados. Os conceitos necessários para um cientista de dados incluem fundamentos matemáticos, computacionais e estatísticos. Adicionalmente, deve-se desenvolver a capacidade em gerenciar, realizar a curadoria, descrição e visualização de dados, além da modelagem e avaliação de dados, implementação de fluxos de trabalho e reprodutibilidade. Por fim, também são consideradas essenciais habilidades em comunicação, conhecimento do domínio e soluções de questões éticas. As combinações possíveis dessas competências designam os papéis relativos aos cientistas de dados no mercado de trabalho (COMMITTEE ON ENVISIONING THE DATA SCIENCE DISCIPLINE, 2018).

Dentre as várias abordagens exploradas, destaca-se a necessidade do estudante de Ciência de Dados ter contato com problemas reais que envolvam negócios e o contexto organizacional (CURTY; SERAFIM, 2016, p. 324; RAWLINGS-GOSS, 2019, p. 24). Além disso, é vista como uma educação eficaz aquela que estimula no futuro profissional o senso de autodesenvolvimento contínuo, que

proporcione a adoção do modelo de aprendizagem ao longo da vida. Outro aspecto importante é a flexibilidade na oferta de currículos e cursos, que contribuem para a futura adoção de modelos de gestão de competências (DEMCHENKO; COMMINIELLO; REALI, 2019, p. 125).

3.3 SÍNTESE DO CAPÍTULO

Definir cientista de dados como o praticante da Ciência de Dados é o caminho mais lógico e simples. Todavia, bem como ocorre para a Ciência de Dados, não há consenso sobre o conceito do profissional cientista de dados. Devido à essa indefinição, este capítulo apresenta o conceito de cientista de dados, cuja evolução não ocorreu de forma síncrona à da sua área mãe (KIM; LEE, 2016, p. 162), identificando as competências desse profissional.

Dentre as características das definições do conceito de cientista de dados, estão a simplicidade de Donoho (2015), o foco disciplinar do Instituto Nacional (Americano) de Padrões e Tecnologia (NIST BIG DATA PUBLIC WORKING GROUP, 2015), o pragmatismo de Saltz e Grady (2017) e a visão corporativa de Linden et al. (2018, p. 7). Além de definições, são destacadas as particularidades e reforçada a necessidade de profissionalização do cientista de dados (BRANDT, 2016; CURTY; SERAFIM, 2016; HAYES, 2017; KIM; LEE, 2016; WALKER, 2015).

Em seguida, após apresentar o conceito de competência, com destaque para o modelo CHA, de Durand (2000), são exploradas as competências específicas do cientista de dados. Nesta seção, Loukides (2012) apresenta uma abordagem que foca em habilidades interpessoais como paciência, motivação, destacando a interdisciplinaridade associada à prática da Ciência de Dados. Por outro lado, outros pesquisadores trazem uma perspectiva mais técnica, focando nas habilidades específicas ou nas disciplinas necessárias para o cientista de dados (DHAR, 2013; HALL; PHAN; WHITSON, 2016; HARRIS; MURPHY; VAISMAN, 2013-).

Na sequência, são apresentadas pesquisas que buscam mapear habilidades, conhecimentos e papéis profissionais do cientista de dados. Inicialmente, utiliza-se o tesauro fornecido pelo ESCO (*European Skills, Competences, Qualifications and Occupations*) que traz além da descrição da ocupação do cientista de dados uma lista de competências essenciais a esse profissional. Depois, são apresentados os

requisitos que, segundo Cao (2017), fazem do cientista de dados um bom profissional, que englobam questões técnicas, educacionais e comportamentais, além de experiências profissionais passadas. Em relação aos papéis, são destacadas as publicações de Harris, Murphy e Vaisman (2013-), que estabelecem relações entre competências e papéis, e Linden et al. (2015, 2018), focados no mapeamento de profissionais para montagem de equipes de dados.

Em relação à educação, esta seção discute a demanda pela criação de cursos especializados em Ciência de Dados, sejam cursos livres, de graduação ou de pós-graduação, e a necessidade de padronização e diretrizes educacionais, capazes de certificar a qualificação de profissionais da área (WALKER, 2015). Ademais, são apresentadas propostas de distribuição de conteúdo curricular, particularidades da educação em Ciência de Dados e experiências já realizadas em cursos de nível superior na área (ANDERSON; MCGUFFEE; UMINSKY, 2014; CLEVELAND, 2001; PIETY; HICKEY; BISHOP, 2014). Por fim, são listadas recomendações da Academia Nacional de Ciência, Engenharia e Medicina para o desenvolvimento de cursos de graduação voltados à formação de cientistas de dados. Dentre as sugestões, destacam-se o desenvolvimento do corpo docente, alfabetização de dados, questões relacionadas à ética dos dados, promoção da diversidade e incentivo à educação contínua (COMMITTEE ON ENVISIONING THE DATA SCIENCE DISCIPLINE, 2018).

4 TRABALHOS RELACIONADOS

Neste capítulo, são apresentadas pesquisadas relacionadas à presente tese. Optou-se por dividi-la entre dois grupos distintos: teses e dissertações cujo objeto de estudo é a Ciência de dados enquanto profissão, *surveys* que procuram caracterizar o mercado de Ciência de Dados, incluindo pesquisas que apresentam métodos para o mapeamento de competências. Os procedimentos de coleta e seleção dos documentos são apresentados nas respectivas seções.

4.1 TESES SOBRE A PROFISSÃO EM CIÊNCIA DE DADOS

Conforme relatado na justificativa do primeiro capítulo (Seção 1.3), para buscar por dissertações e teses relacionadas ao tema aqui tratado, foram utilizadas três bases de dados. As duas primeiras, Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) e Catálogo de Teses e Dissertações da CAPES, visavam pesquisas publicadas no idioma português e não resultaram em nenhum registro utilizado. Para pesquisas internacionais, adotou-se a Networked Digital Library of Theses and Dissertations (NDLTD) onde foi possível resgatar três teses defendidas nos Estados Unidos, entre 2016 e 2018. As subseções seguintes sintetizam essas pesquisas, apresentando-as por ordem cronológica.

4.1.1 A emergência da profissão em Ciência de Dados

A primeira tese levantada é apresentada sob o título de *The emergence of the data science profession* e foi defendida por Philipp Soeren Brandt (2016), pela Universidade de Columbia. O autor estuda a formação de um novo profissional, o cientista de dados, e debate como este especialista, detentor de “conhecimentos arcanos”, obteve notoriedade pública. Além disso, um dos pontos de interesse da pesquisa é o debate bilateral envolvendo a Ciência de Dados. Se por um lado, tem-se as oportunidades relacionadas a novas formas de trabalho, do outro, está a preocupação relativa a questões de violação de privacidade.

Como comentado na seção 2.1, para uma introdução ao conceito de Ciência de Dados, Brandt (2016) utiliza a definição da Wikipedia[3], ressaltando a ocorrência de termos "obscuros" adotados na definição. Estes termos, ao mesmo tempo que tornam a definição confusa, demonstram a complexidade do tema. Brandt (2016) comenta que essa complexidade é confrontada com a notoriedade que a Ciência de Dados e, consequentemente, o Cientista de Dados vêm assumindo. Por isso, há a necessidade de estudos que busquem distinguir o papel desse profissional por meio de pesquisas empíricas. O autor considera que há muitas técnicas de análise que são capazes de explorar as minúcias do trabalho daqueles a que chama de *nerds* dos dados, provendo uma visão robusta e detalhada acerca dessa nova profissão.

Neste sentido, Brandt (2016) propõe duas técnicas para fornecer conjuntamente uma visão multifacetada sobre Ciência de Dados. A primeira consiste em uma abordagem qualitativa, classificada como observação participante (GIL, 2008), onde o autor participou por três anos de encontros organizados por e para cientistas de dados. Aqui, as observações do autor permitiram o reconhecimento das características de comunidade dos tais *nerds*, bem como a identificação de novas formas de trabalho providas pela transformação tecnológica da qual a Ciência de Dados faz parte. A segunda técnica adotada emprega técnicas quantitativas de análise de texto e análise de redes. Essa etapa investiga a Ciência de Dados sob o contexto do mercado, tendo como base de dados um conjunto de descrições de empregos, e sob o contexto acadêmico, analisando currículos de instituições que treinam os Cientistas de Dados.

4.1.2 Definindo a Ciência de Dados e o Cientista de Dados

Com o título *Defining Data Science and Data Scientist*, Dana M. Dedge Parks (2017) defendeu sua tese pela Universidade do Sul da Flórida. Inicialmente, a autora

---

[3] A definição utilizada pelo Brandt (2016) ("Ciência de Dados é um campo interdisciplinar sobre processos e sistemas para extrair conhecimento ou *insights* de dados em diversos formatos, estruturados ou não, que é uma continuação de alguns dos campos de análise de dados, como estatísticas, mineração de dados e análises preditivas, semelhante à Descoberta de Conhecimento em Bancos de Dados (KDD). "), tradução nossa, grifo do autor) não está mais vigente no endereço https://en.wikipedia.org/wiki/Data_science.

ressalta que a expansão na geração e armazenamento de dados advindos de atividades diárias, como as tais "pegadas digitais", requer uma mudança de mentalidade nas organizações para extrair todo potencial desses recursos. Porém, mesmo que essa transformação cultural exija uma compreensão fundamental do que é Ciência de Dados, esse conceito, juntamente com o termo Cientista de Dados, ainda não está formalmente definido. Como agravante, tem-se que diante de concepções mal-entendidas sobre os termos, instituições os incorporam a um nível incipiente, sem integração estratégica aos negócios da organização.

Sob a perspectiva dos profissionais, a tese apresenta a Ciência de Dados como um amálgama interdisciplinar de habilidades, técnicas e ferramentas que possibilitam às organizações a identificação de tendências e a construção de hipóteses que orientam a tomada de decisões. Sob o viés acadêmico, a Ciência de Dados é compreendida como uma metodologia pela qual o cientista de dados analisa questões não atendidas pela estatística, com rigor científico e capacidades sistêmicas para garantir que uma resposta a uma pergunta de dados seja acurada. Parks (2017) argumenta que tanto profissionais quanto acadêmicos trabalham para construir teorias, implementações e uma cultura focada em dados. Assim, a autora orienta sua tese para desenvolver a base para a definição do que é Ciência de Dados e do papel do cientista de dados.

Para essa finalidade, a autora adota três procedimentos metodológicos. O primeiro consistiu em uma revisão da literatura que analisou 22 artigos e um livro cujo conteúdo apresentava potencial de criar consenso ou apresentar visões opostas sobre os conceitos analisados. Em seguida, procedeu-se às entrevistas realizadas com profissionais de diversos cargos (analistas de dados, arquitetos de dados, cientistas de dados, diretores financeiros e acadêmicos) cujos tempos de experiência na área variavam de quatro a 20 anos. Nesta etapa, por meio de análise qualitativa do conteúdo, foi possível identificar repetição de frases, bem como frases únicas, além de termos adotados para descrever funções, competências, ferramentas e capacidades sistêmicas necessárias para o desempenho profissional da área. Por fim, a tese também adota uma pesquisa de levantamento com fonte de dados. Um questionário *survey*, composto por 11 questões de múltipla escolha e sete questões dissertativas, foi respondido por amostra de 78 respondentes, dividida entre universitários, executivos, especialistas em tecnologia, marketing, analistas e

cientistas de dados, professores, funcionários públicos, da saúde, da educação e construção civil. Para analisar os dados obtidos, foram adotados procedimentos quantitativos e qualitativos que permitiram, além de identificar características demográficas, descrever aspectos característicos do local de trabalho desses profissionais.

Entre os principais resultados obtidos pela pesquisa, destaca-se a existência de um consenso entre acadêmicos e profissionais sobre os atributos necessários para uma adoção bem-sucedida da Ciência de Dados. Embora haja discordância quanto à proporção, os resultados apontam que a combinação de estatística, matemática e programação é essencial para a extração de informação significativa dos dados. Dentro desses atributos, a autora identifica 55 competências agrupadas em cinco categorias: 1) Funções de análise de dados; 2) Programação e ferramentas; 3) Técnicas e soluções de aprendizado de máquina; 4) Conhecimento interdisciplinar e; 5) Habilidades críticas.

Como conclusão, Parks (2017) reforça que a Ciência de Dados se encontra em estágio evolucionário onde ocorrem mudanças diárias à medida que surgem novas tecnologias e competências até então inconcebíveis. Por fim, a autora valoriza o aspecto de expansão da área e a presença de pessoas com mentalidade científica dedicadas a descobrir conhecimento ocultos nos enormes armazenamentos de dados contidos em *data centers* espalhados pelo mundo.

4.1.3 Uma análise das oportunidades em Ciência de Dados

A terceira tese estudada foi defendida por Angel Krystina Washington Durr (2018) sob o título *A Text Analysis of Data-Science Career Opportunities and US iSchool Curriculum*, pela Universidade do Norte do Texas. A autora ressalta, na introdução da pesquisa, a demanda do mercado por profissionais com habilidades em dados que vem provocando o surgimento de uma série de novas profissões. Essa competência em lidar com dados, tratada na tese como literacia de dados, é vista como a capacidade do indivíduo em interagir com os dados a ponto de lhes atribuir significado, gerando informação.

Para Durr (2018), a Ciência de Dados é um entendimento científico dos dados e vem se tornando requerida em muitos papéis profissionais. A orientação a dados

das organizações contribui para o surgimento de novos cargos, além do cientista e do analista de dados. Arquiteto de dados, engenheiro de dados ou mesmo jornalista de dados são alguns exemplos. Os profissionais da informação também são vistos pela autora com importante papel na Ciência de Dados, especialmente nas etapas que antecedem a fase de análise. Com o crescente aumento no volume e nos formatos dos dados, parcerias entre o cientista da informação, especializado no tratamento e manutenção dos dados, e o cientista de dados, especializado nas análises, podem gerar economia de recursos.

Como o aumento da importância dos dados nas vidas das pessoas, nos negócios, nas organizações privadas e públicas, maior a necessidade de pessoas com competência em Ciência de Dados. Por outro lado, uma vez que as funções do cientista de dados não são claramente definidas e há surgimento contínuo de novos cargos especialistas em dados, a tarefa de recrutadores em selecionar o profissional adequado às necessidades da organização também se torna mais complicada. Esses fatores também tornam a educação em Ciência de Dados mais complexa, principalmente se orientada a uma função profissional específica.

Durr (2018), com objetivo de investigar se a formação profissional em Ciência de Dados atende as demandas do mercado, formula duas questões que guiam sua pesquisa: “Quais as características e os requisitos mencionados nas ofertas de trabalho em Ciência de Dados?” e “Como o currículo da iSchool contempla os requisitos de especificados nas ofertas de trabalho em Ciência de Dados?”. Ao responder essas perguntas, a autora objetiva desenvolver um *framework* específico para as habilidades e competências necessárias para os profissionais de Ciência de Dados, além de fornecer uma base para pesquisas futuras sobre treinamento e educação dos profissionais da área.

Para obter essas respostas, a autora inicia esclarecendo a definição de iSchool, que é um grupo de instituições de educação superior voltadas ao ensino da informação. Pelo site oficial, a organização iSchool (2020) foi fundada em 2005 por um coletivo de entidades educacionais, faculdades, universidades e departamentos, promotoras de programas formalmente focados em temas relacionados à informação, como tecnologia da informação, biblioteconomia, informática e ciência da informação. A organização é composta por 116 IES da Ásia, América do Norte e Europa, que juntas compartilham comum interesse nas relações entre informação, pessoas e

tecnologia (ISCHOOLS, 2020). Para Durr (2018), as iSchools são consideradas instituições pioneiras nas iniciativas de Ciência da Informação e ensinam os futuros líderes do campo da informação, e possuem notável reputação na comunidade acadêmica. Logo, é justificado utilizar a organização como critério de seleção para as instituições estudadas na pesquisa.

O *corpus* analisado por Durr (2018) era formado por 1603 documentos advindos de anúncios de vagas de empregos e 439 conteúdos programáticos de cursos de pós-graduação de instituições vinculadas à iSchool. As vagas foram retiradas do site indeed.com que, segundo o próprio site da empresa, é o maior site de emprego do mundo, presente em 38 países e recebe cerca de 250 milhões de visitantes únicos por mês (INDEED, 2020). Para seleção dos anúncios, buscou-se vagas que continham a expressão “data science” no título da vaga ou nas competências requeridas. Dentre as organizações vinculadas à iSchool, Durr (2018) entrou em contato com as 37 localizadas no território americano à época, das quais, apenas sete instituições (18,91%) aceitaram participar da pesquisa. A aparente falta de um repositório centralizado com todo o material completo dos cursos dificultou a coleta de dados e impediu que a pesquisadora atingisse sua meta inicial de 10 IES.

Com o uso da ferramenta SAS Enterprise Miner, Durr (2018) realizou uma análise de texto em 51.072 sentenças, derivadas dos 2049 documentos coletados. O levantamento das frequências dos termos permitiu comparar as competências mais discutidas nas ofertas de emprego ao conteúdo programático dos cursos. Para identificação das principais competências, adotou-se o procedimento de análise de autovalores (*eigenvalues*) que identificou 18 tópicos estatisticamente significativos que foram organizados em *rankings*, segundo a taxa de ocorrências dos anúncios e dos currículos, facilitando a comparação entre os dois contextos.

Em relação aos resultados, embora o teste de qui-quadrado tenha mostrado diferenças estatísticas nas distribuições das competências entre as duas amostras, ficou apontado que os currículos das iSchools cobrem com sucesso o conteúdo procurado pelo mercado americano de Ciência de Dados. Todavia, Durr (2018) aponta que pesquisas mais apuradas são necessárias para se aprofundar nessa investigação, incluindo uma análise mais específica sobre a relação entre a Ciência da Informação e a Ciência de Dados. Uma recomendação da autora está na aplicação de uma pesquisa *survey* junto aos profissionais e aos empregadores de Ciência de

Dados para determinar se existem necessidades cuja educação formal dos cientistas de dados não vem contemplando.

4.2 PESQUISAS DE LEVANTAMENTO

Durante a realização da pesquisa bibliográfica, foram identificadas pesquisas de levantamento de campo que buscavam caracterizar o perfil do profissional atuante na Ciência de Dados. Segundo Gil (2008, p. 55), esse tipo de pesquisa, também denominado *survey*, caracteriza-se pela interrogação direta de indivíduos cujo comportamento está sendo estudado. Assim, a partir de um número significativo de respostas e métodos quantitativos de análise, busca-se a ampliação no conhecimento sobre o público pesquisado. As próximas subseções sintetizam *surveys* acerca do cientista de dados que buscam identificar as características desse profissional, sob diferentes abordagens.

4.2.1 Analisando os analistas

A primeira pesquisa de levantamento analisada foi desenvolvida por Harris, Murphy e Vaisman (2013-) e publicada pela editora O'Reilly Media, especializada e reconhecida pelas suas publicações na área da tecnologia da informação. De forma anedótica, os autores, para ilustrar as recorrentes confusões relacionadas ao profissional de Ciência de Dados, apresentam quatro cientistas de dados com competências completamente distintas. Questionamentos sobre essa generalização orientam a pesquisa. Utilizar o termo "cientista de dados" para descrever todos esses é elucidativo, distingue pessoas com diferentes forças e permite que organizações tomem boas decisões? Esse agrupamento permite um plano de carreira viável e sugere opções de crescimento profissional? Ou, em vez disso, leva a confusão, mal-entendidos e perda de oportunidades?

Para Harris, Murphy e Vaisman (2013-), a excessiva valorização desse profissional cria expectativas sobre especialistas com capacidades miraculosas. Além disso, a falta de consciência sobre as variedades dos cientistas de dados leva as organizações a desperdiçarem seu tempo em busca de talentos extraordinários. A fim de responder às questões levantadas e investigar como as confusões acerca das

habilidades e papéis na Ciência de Dados suscitam perda de tempo e recursos, os autores aplicam uma *survey*, na qual cientistas de dados se descrevem a si próprios, bem como suas habilidades. Mais do que ferramentas e técnicas, a pesquisa buscou entender e definir subgrupos de cientistas de dados, segundo sua própria perspectiva, não com base nos anos de experiência, nível educacional ou cargos.

O questionário, enviado via *web*, levava menos de 10 minutos para ser preenchido e focava em cinco áreas: habilidades, experiências, educação, autoidentificação e presença na *web*, item que apresentou baixa participação e foi retirado da análise. Para coletar os dados das habilidades, o questionário apresentava uma lista com 22 itens, considerados úteis para a realização do trabalho de cientista de dados, dispostos de maneira aleatória, para que os respondentes ordenassem de acordo com seu nível de proficiência. Em relação à autoidentificação, os respondentes indicavam, por meio de uma escala Likert, seu nível de identificação perante 11 categorias profissionais vinculadas à Ciência de Dados. Durante o ano de 2012, a pesquisa coletou 250 preenchimentos completos, advindos de vários países.

Os resultados apontaram quatro grupos, fatores latentes, de cientistas de dados: a) Desenvolvedor de dados (*data developer*), formado por desenvolvedores e engenheiros; b) Pesquisador de dados (*data reseacher*), englobando pesquisadores, cientistas e estatísticos; c) Criativo de dados (*data creative*), formado por profissionais artistas, *hackers* e "pau para toda obra" e; 4) Pessoas de negócios em dados (*data businessperson*), incluindo líderes, pessoas de negócio e empreendedores (HARRIS; MURPHY; VAISMAN, 2013-, p. 10). Após essa etapa, as habilidades foram combinadas às 22 habilidades, separadas em cinco grupos: Negócios, *Machine Learning*/*Big Data*, Matemática/Pesquisa Operacional, Programação e Estatística.

Após as análises específicas para cada grupo, Harris, Murphy e Vaisman (2013-) identificam a amplitude em habilidades como característica definidora dos cientistas de dados. A maioria é capaz de, ainda que sozinho, construir novos produtos de dados, incluindo todas as etapas necessárias, ainda que em nível de protótipo. Os autores também constataram que os cientistas de dados mais bem-sucedidos são aqueles com experiência substancial e profunda em pelo menos uma área das categorias da Ciência de Dados, indo ao encontro do conceito do profissional com "habilidades em forma de T" (*T-Shaped*).

Esse tipo de profissional apresenta amplitude de habilidades na parte superior e profundidade em uma área, criando uma barra vertical e formando o “T”. Algumas características do profissional T são a facilidade para trabalhar em equipes interdisciplinares e a maior eficiência dos que profissionais sem profundidade. A Ciência de Dados é um campo inerentemente colaborativo e criativo, onde um mesmo profissional precisa lidar com administradores de bancos de dados, estatísticos, matemáticos, empresários, dentre outros profissionais multidisciplinares, sempre buscando inovar no trabalho com os dados.

Harris, Murphy e Vaisman (2013-) concluem que os times de Ciência de Dados lidam com dados brutos, mas também se comunicam com os mais altos níveis da tomada de decisão. Seus integrantes precisam de uma diversidade de habilidades para que essas interações se deem da melhor maneira. Todavia, os autores reforçam que sua categorização é apenas uma sugestão que pode ser útil à comunidade em geral, sendo apenas um passo para o melhor entendimento dessa profissão, ainda em desenvolvimento. Por fim, salientam que pesquisas adicionais podem ajudar a esclarecer essas categorias, além de demonstrarem como o campo está mudando ao longo do tempo, como novos caminhos educacionais e de carreira.

4.2.2 O estado da Ciência de Dados

Em 2021, o site Kaggle publicou a quinta edição da pesquisa que realiza junto a seus membros. O Kaggle é uma comunidade focada nos praticantes de Ciências de Dados e, além de contar com mais de três milhões de usuários registrados, disponibiliza em seu portal mais de 50 mil bases de dados e 400 mil *scripts* de análises, na forma de conteúdo público (KAGGLE, 2020). Os dados dessa *survey* anual são abertos e compartilhados, resultando na “maior e mais compreensível” base de dados disponível sobre a Ciência de Dados e Aprendizagem de Máquinas (KAGGLE, 2019, p. 2).

A edição de 2021 coletou 25.973 respostas (KAGGLE, 2021). Com abrangência global, a comunidade Kaggle é composta por praticantes de todos os níveis sobre a qual a pesquisa buscou dados sobre a formação, o emprego e as ferramentas utilizadas na prática da Ciência de Dados. Todavia, o sumário executivo publicado se restringiu a analisar 14% dessa amostra, que correspondiam aos

profissionais formalmente identificados como cientistas de dados. As demais respostas, embora possam ser acessadas pela base de dados disponibilizada, não compõem os resultados apresentados pela organização.

Dentre os principais achados em relação às características dos profissionais, a pesquisa destaca a Ciência de Dados como uma "profissão masculina", onde 82,2% dos praticantes são homens, confirmando as edições anteriores. Além disso, a amostra indica que os cientistas de dados são jovens, mais da metade tem menos de 30 anos, e altamente qualificados, sendo que 92,8% dos respondentes possuem ensino superior, 47,7% fizeram mestrado e, 15,0%, doutorado. Por outro lado, a maioria dos entrevistados possui menos de cinco anos de experiência em programação.

Quanto à remuneração, com valores médios entre 100 e 125 mil dólares anuais, os resultados apontam os EUA com os maiores salários, muito acima dos demais países. Segundo a pesquisa, a média salarial dos cientistas de dados no Brasil fica entre 20 e 25 mil dólares anuais e é a quinta maior dos países listados. Além disso, os valores de remuneração dos profissionais brasileiros são superiores à edição de 2020, quando a faixa salarial encontrada era de 10 a 15 mil dólares anuais.

Além de disponibilizar ao público os dados coletados pela pesquisa, o Kaggle promove um concurso para promover as melhores análises advindas da comunidade. Com um total de 50 mil dólares em prêmios, sendo o prêmio principal de 10 mil dólares e conta com um sistema de pontuação formal para definir os vencedores. Em 2020, o trabalho que liderou o *ranking* foi desenvolvido por Parul Pandey e, sob título *Geek Girls Rising: Myth or Reality!* ("A ascensão das garotas *geeks*: mito ou realidade", tradução nossa), apresenta uma análise da representação feminina no universo da Ciência de Dados e do *machine learning* (PANDEY, 2019).

O sumário executivo enfatiza ainda que, por mais que muitos cientistas de dados possuam altos graus de educação formal, é característica desse profissional a busca constante pelo desenvolvimento de novas competências. Nesse sentido, uma vez que muitas empresas ainda estão iniciando na prática de *machine learning*, também há pelo âmbito organizacional, uma evidente necessidade de aprendizagem contínua. A pesquisa destaca, então, blogs, fóruns, incluindo o próprio Kaggle, além de sites como Coursera e YouTube como métodos recorrentes de educação continuada.

A *survey* realizada pelo Kaggle, além de fornecer uma robusta caracterização do cientista de dados em uma escala global, ao compartilhar abertamente seus dados, torna-se uma fonte de dados a ser utilizada na presente pesquisa. Com os dados disponibilizados, é possível isolar as respostas de profissionais atuantes no Brasil para traçar um perfil preliminar e, assim, comparar com os dados aqui coletados primariamente. Os procedimentos desta etapa são detalhados na seção de metodologia.

4.2.3 Salários de cientistas de dados

Outra *survey* sobre o perfil profissional do cientista de dados é a pesquisa realizada pela empresa de recrutamento executivo de "talentos quantitativos", Burtch Works (2019, p. 9). A empresa define como um profissional quantitativo aqueles cuja especialidade está entre análises preditivas, Ciência de Dados, engenharia de dados, análise quantitativa de negócios, análises web, análise de crédito ou risco, pesquisas de marketing, dentro outros. Publicado em junho de 2019, o relatório apresenta, além de questões demográficas, dados salariais de cientistas de dados e profissionais de análises preditivas (PAP) que em 2018 haviam sido publicados separadamente.

Dos 40.000 profissionais com os quais a Burtch Works mantém contato, 2.261 participaram da pesquisa, sendo que 421 são identificados como cientistas de dados. Porém, se o tamanho da amostra não é um diferencial da pesquisa, seus autores ressaltam os pontos que a tornam única. O primeiro deles é o foco exclusivo nos cientistas de dados e PAP, excluindo outras áreas como *business intelligence*, pesquisas operacionais, tecnologia da informação, por exemplo. Outros diferenciais são decorrentes da experiência da companhia no segmento, que possibilitam aos seus analistas estabelecer com confiança compensações decorrentes de diferentes regiões, nível do cargo, segmento de mercado, nível de escolaridade e gênero.

Ademais, o relatório fornecido pela Burtch Works realiza uma distinção entre os dois profissionais estudados. Segundo o estudo, o profissional de análise preditiva é aquele que consegue aplicar habilidades quantitativas sofisticadas a dados derivados de transações, interações e outros comportamentos para gerar informação e guiar ações (BURTCH WORKS, 2019, p. 4). Por outro lado, os cientistas de dados são definidos como "um subconjunto de PAPs que possuem as habilidades de ciência

da computação necessárias para adquirir e limpar ou transformar dados não estruturados ou de streaming contínuo, independentemente de seu formato, tamanho ou origem" (BURTCH WORKS, 2019, p. 4, tradução nossa). São apresentados como exemplos de dados não estruturados *streams* de vídeo, dados de áudio, *scraps* da web, incluindo de redes sociais digitais, dados de sensores, arquivos de *log* ou longos blocos de linguagem escrita.

Para apresentar os salários, a pesquisa separou os respondentes em duas categorias: a primeira, chamada de colaborador individual, corresponde ao profissional que não gerencia outros colaboradores; aos cientistas de dados que executam esse papel, a pesquisa chama de gerente. Ambas as categorias são divididas em três níveis que variam de acordo com as responsabilidades e tempo de experiência de cada papel. Enquanto no primeiro nível dos colaboradores, o salário anual médio é de 95 mil dólares, mesmo valor de 2018, no nível mais alto, chega a 167 mil dólares, 1% maior que no ano anterior. Para a categoria gerencial, o nível mais baixo apresenta salário anual médio de 146 mil dólares e chega a 250 mil, no terceiro nível. A variação para a pesquisa anterior foi de 0% e 1%, respectivamente.

Se comparado à *survey* do Kaggle (2019), os cientistas de dados da pesquisa apresentaram um nível de escolaridade ainda mais elevado, sendo que entre os respondentes, 47% possuíam mestrado enquanto outros 47%, doutorado. Um nível maior de escolaridade está relacionado a uma maior remuneração, com exceção para os níveis gerenciais 2 e 3. Nesses casos, a experiência de trabalho e as competências em gestão representam elementos mais valorizados que a educação formal (BURTCH WORKS, 2019, p. 20).

Em relação ao gênero, 83% dos cientistas de dados são homens, enquanto 17%, mulheres. A participação feminina apresentou um acréscimo de 2% se comparada ao ano anterior. Esse percentual é bem menor que o encontrado para PAPs, onde 26% dos profissionais são mulheres. Para ambos os profissionais, essa diferença é menor no primeiro nível da categoria colaborador individual, onde 32% das respostas advêm de mulheres. Esse resultado pode indicar uma inserção menos díspar de novos profissionais no mercado. Por outro lado, infelizmente, o tamanho da amostra feminina não permitiu verificar a ocorrência de diferenças salariais entre os gêneros.

Diferentemente do levantamento realizado pelo Kaggle (2019), a pesquisa de Burtch Works não aborda as especificidades do trabalho dos cientistas de dados, como métodos e ferramentas. Por outro lado, se aprofunda em aspectos demográficos, como categorização do profissional e regiões geográficas, estabelecendo distinções e identificando tendências de curto e longo prazos para esse segmento do mercado. Mesmo assim, é possível comparar os estudos que apresentarem resultados congruentes para o nível de escolaridade e da proporção entre os gêneros. Entretanto, no quesito salário é encontrada uma discrepância considerável entre as pesquisas. Essa diferença pode ser explicada pelo procedimento de categorização realizado pela Burtch Works e pelo fato dessa pesquisa ter se restringindo ao mercado americano.

4.2.4 Desmistificando a Ciência de Dados

Em meados de 2015, Bob E. Hayes, doutor e consultor em Ciência de Dados, realizou, em parceria com a AnalyticsWeek, empresa independente de tecnologia e pesquisa no mercado de análises, realizou um levantamento com profissionais atuantes na área de dados (HAYES, 2015). A coleta de dados da pesquisa ainda se encontra ativa[4], porém a última análise encontrada está publicada no *website* do autor e data do ano de 2017, sendo o primeiro sumário executivo da pesquisa apresentado ainda em 2015 (HAYES, 2017).

Nesta pesquisa, podendo escolher mais de uma opção, os respondentes indicavam a classificação que melhor descrevia seu trabalho: Pesquisador (56,8%), Gerente de Negócios (40,2%), Criativo (37,1%) e Desenvolvedor (36,1%). Em seguida, indicavam seu nível de proficiência, variando de “Não conheço” a “Especialista”, em 25 competências inerentes à Ciência de Dados. Destas, destacaram-se comunicação, gestão de dados estruturados, mineração de dados e ferramentas de visualização. Em contrapartida, competências como *big data*, dados distribuídos e gestão de computação em nuvem foram os itens nos quais os

[4] Até o dia 06 de setembro de 2020, o formulário da pesquisa permanece ativo no endereço https://loyaltywidget.com/limesurvey/index.php?sid=42831.

profissionais apresentaram o menor nível de destreza. No geral, cerca de 23% dos respondentes se declaram especialistas em uma competência, 10%, especialistas em duas competências e apenas 1% indicou profundo conhecimento em cinco competências, indicando a dificuldade de se encontrar o profissional “unicórnio”.

A amostra original era composta por 420 profissionais majoritariamente advindos da América do Norte (64%), trabalhando para empresas B2B (80%), com menos de 1.000 colaboradores (56%) e das áreas de TI, consultoria, saúde e finanças (68%). Em relação ao nível de escolaridade, 45% apresentaram mestrado completo enquanto 19% possuíam doutorado, confirmando o alto nível de formação educacional do segmento. Por fim, quanto ao gênero, os profissionais masculinos representaram 77% dos respondentes, valor que, ainda que em menor proporção, reafirmou o desequilíbrio entre homens e mulheres atuando na área. Uma das hipóteses da atenuação dessa diferença é que a pesquisa englobou profissionais da área de negócios, excedendo os cargos técnicos.

Para Hayes (2015), a interdisciplinaridade da Ciência de Dados a torna uma área guarda-chuva que requer inúmeras competências para o profissional que busca extrair conhecimento a partir de dados. Em sua abordagem mais recente, Hayes (2017) investiga o campo sobre três aspectos: Pessoas, Processos e Ferramentas. Assim, além das competências profissionais, o autor questiona sobre o fluxo de trabalho e investiga as ferramentas e plataformas utilizadas pelos profissionais de dados em seu cotidiano.

4.2.5 O relatório sobre o Cientista de Dados

Outra pesquisa de levantamento cuja primeira edição data do ano de 2015 é o relatório publicado pela Figure Eight. Conhecida como CrowdFlower até o exemplar de 2018, a empresa é desenvolvedora de uma plataforma de inteligência artificial voltada a equipes de Ciência de Dados. Assim como o ambiente da Ciência de Dados mudou durante os anos, com o crescimento dos dados e da capacidade de processamento, com a popularização de projetos de *machine learning* e a ampliação no número de vagas, a pesquisa se adaptou durante os anos. Na versão de 2018, a pesquisa contrapôs a visão dos cientistas de dados à opinião de profissionais éticos,

como médicos, teóricos e legisladores sobre a adoção da inteligência artificial (FIGURE EIGHT, 2018, p. 3). Este ponto destaca a pesquisa das demais estudadas.

Em 2015, a amostra da pesquisa era formada por 153 profissionais que detinham o título de “cientista de dados” em seus cargos no LinkedIn. Era trabalhadores de empresas de diversos portes e setores, majoritariamente localizadas nos EUA. Para 2018, a Figure Eight conseguiu 240 respostas advindas de questionário enviado via e-mail ou de eventos da área. O relatório surgiu com o objetivo de descobrir o que não funcionava no campo de Ciência de Dados e dar às organizações informações para formarem times de dados mais estratégicos, produtivos e “felizes” (CROWDFLOWER, 2015, p. 5).

A felicidade dos cientistas de dados é um item realçado desde a primeira edição do relatório. Segundo a pesquisa, os cientistas de dados “amam” seu trabalho, sendo que o percentual de profissionais “felizes” ou “muito felizes” passou de 67% em 2015 para 89%, em 2018 (FIGURE EIGHT, 2018, p. 4–5). Mesmo satisfeitos, os cientistas de dados são frequentemente deparados à possibilidade de troca de emprego. No relatório mais recente, quase 30% dos respondentes são contatados várias vezes por semana para novas oportunidades de trabalho e, para 85% da amostra, essa frequência é de pelo menos uma vez por mês (FIGURE EIGHT, 2018). Dados que representam a demanda por esse tipo de profissional.

Entre outros resultados do primeiro relatório, foram revelados que os principais obstáculos enfrentados pelo cientista de dados é o tempo gasto na limpeza dos dados desorganizados (57,5%) e má qualidade dos dados (52,3%). Limitações tecnológicas foram indicadas por apenas 30,1% dos respondentes, não figurando entre os itens mais selecionados. Por outro lado, as análises preditivas (53,6%) e reconhecimento de padrões (52,3%) foram indicadas como as tarefas mais interessantes da rotina da profissão. Dentro dos múltiplos papéis assumidos pelos cientistas de dados, a função de pesquisador (54,4%) e cientista da computação (52,3%) foram as mais recorrentes, seguidos pelo papel de analista de BI, indicado por 36% dos respondentes.

Como comentado, a versão de 2018 do relatório aborda questões éticas que estão sendo discutidas pela sociedade como nunca. Além da privacidade dos dados, que sempre foi uma preocupação primordial, a inteligência artificial (IA) vem gradativamente sendo adota para tomar decisões importantes, como diagnósticos

médicos ou sentenças judiciais. Assim, um amplo debate público, que inclua perspectivas alheias à área tecnológica, faz-se necessário.

Logo na primeira questão que compara as respostas dos dois públicos, fica nítida a divergência entre eles. Perguntados sobre o impacto da IA ao mundo, 75% dos cientistas de dados acreditam que ser faria será "bom", enquanto 16% consideram que será "ruim" e, consequentemente, 9% pensam que não haverá mudanças significativas. Para os profissionais éticos, esses percentuais são 39%, 45% e 15%, respectivamente. Ou seja, enquanto a maioria dos profissionais de dados acreditam que os efeitos da IA serão benéficos, a maior fatia dos profissionais éticos pensa que, majoritariamente, os efeitos serão negativos.

Outra pergunta ilustrativa que demonstrou a divergência entre os dois grupos, referia-se à adoção de carros autônomos. Enquanto 75% dos cientistas de dados afirmaram que prefeririam passear em um carro autônomo, este foi o percentual de profissionais éticos que preferem dirigir por conta própria. Porém, perguntados se as decisões advindas de algoritmos são menos tendenciosas que decisões todas por pessoas, ambos os grupos concordaram, ainda que os algoritmos sejam desenvolvidos por programadores humanos. Outros itens apresentados nesta seção do relatório verificam em quais cenários seriam indicadas aplicações baseadas em IA para tomada de decisões sem interação humana e quais as implicações éticas de lançamento de produtos que não sejam acessíveis de maneira igualitária a todos.

Nos quatro relatórios publicados entre 2016 e 2018, há um consenso sobre o tempo gasto na limpeza e organização dos dados. Assim, para extrair todo o potencial estratégico dos cientistas de dados, como provedores de informação e suporte à tomada de decisão, os relatórios recomendam que esses se concentrem nos aspectos analíticos do seu trabalho. Caso contrário, o trabalho com dados de má qualidade resultará em desperdício de recursos e talentos.

4.2.6 A construção da profissão em Ciência de Dados

Por fim, julga-se pertinente citar um levantamento, ainda que nenhum relatório tenha sido publicado[5]. A pesquisa é derivada do projeto europeu EDISON cuja finalidade era estabelecer a profissão do cientista de dados (EDISON PROJECT, 2017). Introduzido na seção 3.1.4, o EDSF é organizado em seis documentos que visam atenuar a complexidade das habilidades e competências necessárias para definir a Ciência de Dados como profissão: 1) Data Science Competence Framework (CF-DS); 2) Data Science Body of Knowledge (DS-BoK); 3) Data Science Model Curriculum (MC-DS) e; 4) Data Science Professional Profiles (DSPP).

A pesquisa de levantamento realizada pelo projeto EDISON é dedicada a colaborar com a formação de futuros cientistas de dados, uma vez que investiga detalhadamente diversos aspectos da profissão. Por isso, destina-se aos diversos papéis envolvidos: cientistas, estudantes, pesquisadores, gerentes de RH e educadores. Divido em 13 seções, o questionário apresenta 31 questões, em sua maioria de múltipla escolha, abordando aspectos demográficos, educacionais, das organizações empregadoras, ainda que o foco esteja nas habilidades e competências.

Em 07 de setembro de 2020, contatou-se a administração do projeto EDISON para verificar o status atual e os planos futuros da pesquisa, mas não se obteve nenhum retorno.

4.3 SÍNTESE DO CAPÍTULO

O capítulo 4 sintetiza os trabalhos que serviram como referência para esta tese, organizados em dois grupos: 1) teses e dissertações que investigam a profissão em Ciência de Dados e; 2) pesquisas de levantamento que apresentam o perfil dos profissionais e do mercado da Ciência de Dados, englobam aquelas que fazem

[5] Até o dia 06 de setembro de 2020, o formulário da pesquisa permanece ativo no endereço https://www.surveymonkey.com/r/QRPQ9VC.

qualquer tipo de mapeamento de competência do cientista de dados. As pesquisas utilizadas são:

- **A emergência da profissão em Ciência de Dados** (BRANDT, 2016)**:** tese que pesquisa a formação do cientista de dados e como este profissional obteve relevância na sociedade. Empregou abordagem qualitativa, optando por uma observação participante, e quantitativa por meio de análise de texto e análise de redes.
- **Definindo a Ciência de Dados e o Cientista de Dados** (PARKS, 2017)**:** tese que para definir o que é Ciência de Dados e o cientista de dados utiliza revisão da literatura, entrevistas com profissionais, análise de conteúdo e pesquisa de levantamento. Dentre as conclusões da tese, é valorizada o estágio evolutivo da Ciência de Dados que sofre mudanças diárias à medida que engloba novas tecnologias e competências.
- **Uma análise das oportunidades em Ciência de Dados** (WASHINGTON DURR, 2018)**:** tese que analisa 1603 anúncios de vagas de emprego e 439 conteúdos programáticos de cursos de pós-graduação. Por meio de mineração de texto, apresenta as principais competências dos profissionais da Ciência de Dados.
- **Analisando os analistas** (HARRIS; MURPHY; VAISMAN, 2013-)**:** primeira pesquisa de levantamento encontrada no estágio inicial desta tese, buscava distinguir os diferentes papéis dos cientistas de dados e suas respectivas habilidades. Com 250 preenchimentos válidos, a pesquisa definiu quatro grupos de profissionais (desenvolvedor de dados, pesquisador de dados, criativo de dados e pessoa de negócio de dados) e cinco conjuntos de habilidades (Negócios, *Machine Learning*/*Big Data*, Matemática/Pesquisa Operacional, Programação e Estatística). Dentre as conclusões, os autores reforçaram que os cientistas de dados mais bem sucedidos são aqueles com habilidades em forma de T, ou seja, possuem fluência em todos os conjuntos de habilidades, possuindo experiência e conhecimento profundo em um deles.

- **O estado da Ciência de Dados** (KAGGLE, 2020, 2021)**:** maior pesquisa de levantamento direcionada a cientistas de dados é realizada anualmente, fornecendo muitos parâmetros sobre o perfil desses profissionais. Dados demográficos, educacionais e profissionais foram subsídios de comparação para as análises desta tese.
- **Salário de cientistas de dados** (BURTCH WORKS, 2019, 2021)**:** pesquisa de levantamento, realizada por uma empresa de recrutamento e dedicada a profissionais de dados, que traz informações para comparações com o mercado dos Estados Unidos, uma vez que se concentra no território norte americano.
- **Desmistificando a Ciência de Dados** (HAYES, 2017)**:** pesquisa de levantamento que agrupava os profissionais de dados em quatro grupos (pesquisador, gerente de negócios, criativo e desenvolvedor) e, em seguida, solicitava o grau de proficiência dos respondentes em 25 competências, realizando os devidos cruzamentos. Com uma amostra de 420 profissionais, o autor conclui que a Ciência de Dados, por meio de sua relação com a interdisciplinaridade, torna-se uma área guarda-chuva, onde o profissional precisa de diferentes competências para extrair conhecimento dos dados.
- **O relatório sobre o cientista de dados** (CROWDFLOWER, 2015, 2017; FIGURE EIGHT, 2018)**:** realizada entre 2015 e 2018, com amostras entre 153 e 240 respondentes, esta pesquisa de levantamento aborda desde questões de satisfação do cientista de dados com seu trabalho ao tempo gasto em cada tarefa cotidiana da profissão.
- **A construção da profissão em Ciência de Dados** (EDISON PROJECT, 2017)**:** pesquisa derivada do projeto EDISON (EDISON PROJECT, 2017) que ainda está disponível (https://www.surveymonkey.com/r/QRPQ9VC), porém não teve seu relatório publicado.

As pesquisas relatadas neste capítulo foram identificadas na etapa inicial desta tese. Uma busca suplementar foi realizada em maio de 2022, com foco na busca de novas teses, mas nenhum registro adicional foi encontrado. De qualquer forma, os

trabalhos relacionados aqui apresentados foram considerados satisfatórios e contribuíram para o desenho metodológico da presente pesquisa, detalhado no próximo capítulo.

5 PROCEDIMENTOS METODOLÓGICOS

Este capítulo apresenta o esquema metodológico adotado na pesquisa. Inicialmente, é realizada a classificação em relação à natureza, quanto à abordagem do problema e dos objetivos e, por fim, é realizada especificação conforme seus procedimentos técnicos. Em seguida, são detalhados os materiais e métodos de análise empregados no desenvolvimento da tese.

5.1 CLASSIFICAÇÃO DA PESQUISA

Em uma visão geral, esta pesquisa analisa o cientista de dados e, consequentemente, a própria Ciência de Dados. Como resultado, espera-se a aplicação prática das descobertas obtidas, seja para fins educacionais, para a formação de equipes de dados por parte das organizações ou para a orientação de futuros profissionais da área. Logo, quanto à sua natureza, a pesquisa pode ser classificada como aplicada (SILVA; MENEZES, 2005-, p. 20).

Quanto aos objetivos, a pesquisa é predominantemente classificada como descritiva, uma vez que descreve o cenário da Ciência de Dados no Brasil com ênfase nas competências do cientista de dados. Para Gil (2008, p. 28), o principal objetivo das pesquisas deste tipo é descrever “as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis”. Conforme o autor, a pesquisa descritiva, ao extrapolar a simples identificação de relações entre as variáveis e explicar a natureza destes relacionados, aproxima-se da pesquisa explicativa. Neste sentido, Sampieri, Collado e Lucio (2013, p. 100) também defendem que um estudo descritivo pode derivar para pesquisas correlacionais e explicativas, a depender dos critérios metodológicos estabelecidos e resultados obtidos.

A pesquisa apresenta abordagem quantitativa, uma vez que é baseada na utilização de procedimentos estatísticos. Gil (2008, p. 17) define métodos estatísticos como aqueles fundamentados na aplicação de teorias probabilísticas que permitem a determinação numérica das probabilidades admitidas de acerto e erro (GIL, 2008, p. 17). Por outro lado, a pesquisa também possui abordagem qualitativa, uma vez que adota coleta de dados sem medição numérica e buscará, por meio de dados textuais

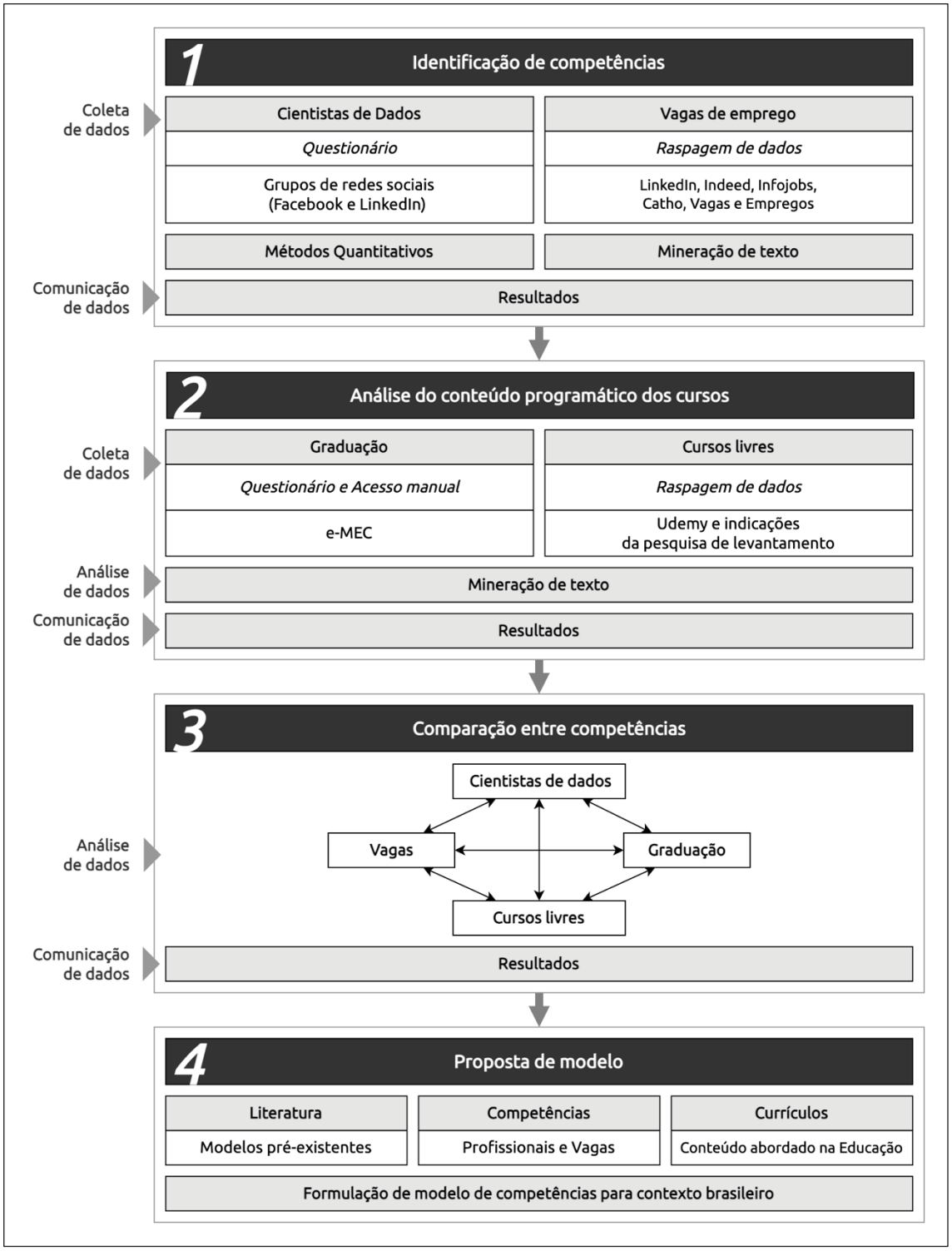
e não estruturados, compreender o contexto geral da área estudada a partir de observações particulares (SAMPIERI; COLLADO; LUCIO, 2013, p. 33).

Neste sentido, Creswell (2010, p. 238) afirma que uma abordagem de métodos mistos, aquela que combina métodos qualitativos e quantitativos, tende a proporcionar mais *insights*, possibilitando uma maior compreensão dos problemas de pesquisa. As pesquisas mistas buscam aproveitar os pontos fortes dos métodos quantitativos e dos métodos qualitativos, com a finalidade de "realizar inferências como produto de toda a informação coletada (metainferências) e conseguir um maior entendimento do fenômeno em estudo" (SAMPIERI; COLLADO; LUCIO, 2013, p. 546). Assim, optou-se pela abordagem mista, considerada mais adequada para lidar com as múltiplas variáveis relacionadas ao assunto pesquisado.

Para os procedimentos técnicos, esta pesquisa pode ser enquadrada como pesquisa de levantamento, visto que utiliza questionário como uma das formas de coleta de dados para descrever numericamente características, tendências e atitudes de uma população por meio de uma amostra (CRESWELL, 2010, p. 33). Ademais, também é uma pesquisa documental, pois realiza tratamento analítico de materiais que não foram analisados conjuntamente (GIL, 2008, p. 51), como anúncios de vagas de emprego, *websites* e matrizes curriculares.

Antes do detalhamento da metodologia, é apresentada na Figura 25 uma visão geral das etapas adotadas ao longo da pesquisa.

FIGURA 25 – VISÃO GERAL DAS ETAPAS METODOLÓGICAS DA PESQUISA



FONTE: O autor (2022).

Assim, verifica-se que os passos metodológicos são separados em quatro etapas principais. A primeira corresponde ao conjunto de procedimentos para identificação das competências dos cientistas de dados por meio dos profissionais

que já atuam na área e de anúncios de vagas de emprego. Em seguida, prossegue-se à análise do conteúdo de cursos de formação para cientista de dados, seja graduação (bacharelado e tecnológico), seja do tipo curso livre. Em um terceiro momento, tem-se a exploração das relações entre os resultados das duas primeiras etapas que embasam, juntamente com a literatura da área, a proposta de um modelo de competências para a Ciência de Dados que corresponde à quarta fase.

5.2 UNIDADES DE ANÁLISE

A escolha pela abordagem de métodos mistos para responder à questão de pesquisa implicou na definição de mais de uma unidade de análise, isto é, elementos, participantes, eventos ou comunidades referentes ao objeto de pesquisa em questão (SAMPIERI; COLLADO; LUCIO, 2013, p. 191). Para Alves-Mazzotti e Gewandsznajder (1998, p. 169–170), uma unidade de análise se refere à organização dos dados para posterior análise. Assim, diferentes procedimentos metodológicos podem demandar unidades de análise distintas para uma pesquisa, abordando diferentes aspectos de um mesmo problema.

Na presente pesquisa, ao iniciar a investigação das competências relacionadas à Ciência de Dados, a primeira unidade de análise são os próprios cientistas de dados. Por isso, o principal procedimento de coleta de dados para este fim é uma pesquisa de levantamento com os profissionais da área. Adicionalmente, para compreender o que as organizações buscam destes especialistas, opta-se pela coleta e posterior análise de anúncios de vagas de emprego para cientistas de dados, formando uma nova unidade de análise.

Outrossim, correspondente a uma terceira unidade de análise, é analisado o conteúdo programático de cursos voltados à formação em Ciência de Dados, seja de graduação ou cursos livres. Os critérios para a composição destas unidades de análise são detalhados a seguir.

5.2.1 População e Amostra

A população, ou universo, é o conjunto de elementos a serem estudados que partilham uma ou mais características em comum (MALHOTRA, 2019, p. 289;

MARCONI; LAKATOS, 2003, p. 223). Em um processo quantitativo, a amostra é um subgrupo definido e representativo da população de interesse, sobre o qual se faz uma coleta de dados para o desenvolvimento de inferências. Assim, espera-se que "os resultados encontrados na amostra consigam ser generalizados ou extrapolados para a população" (SAMPIERI; COLLADO; LUCIO, 2013, p. 192).

Nesta pesquisa, a população em questão corresponde aos cientistas de dados atuantes no Brasil e, uma vez que esta profissão não é regulamentada, não é possível estimar um número exato em relação ao tamanho da população. Logo, com uma população infinita, a amostra a ser utilizada se caracteriza como não probabilística. Este tipo de amostragem, também conhecida como amostra por julgamento, supõe um procedimento de seleção menos formal e prejudica o cálculo do nível de confiança das estimativas, uma vez que não permite o cálculo preciso do erro padrão (SAMPIERI; COLLADO; LUCIO, 2013, p. 208). Assim, se por um lado este tipo de amostragem permite um acesso mais ágil aos dados, por outro, ela traz desvantagens em relação à generalização dos resultados.

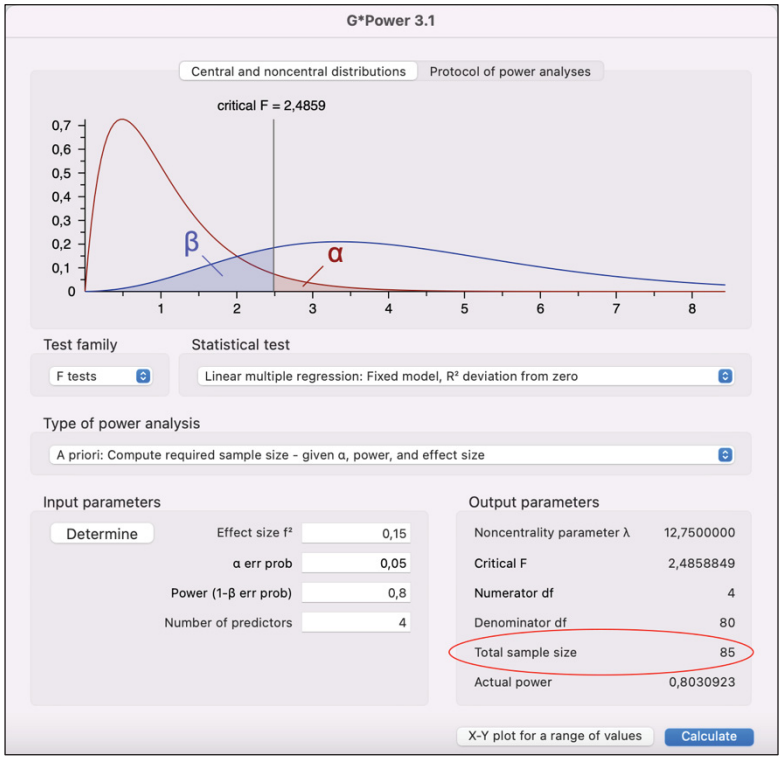
Para se obter ao menos uma projeção do tamanho da população estudada, foram adotados dois procedimentos. O primeiro foi procurar profissionais no LinkedIn por meio da expressão "cientista de dados". Em 05 de maio de 2021, um número aproximado de 7.200 profissionais foi retornado pela busca. Como segundo procedimento, verificou-se a quantidade de membros do maior grupo do Facebook sobre Ciência de Dados no Brasil. O "Data Science Brasil"[6] contava com 23.635 membros em 02 de julho de 2021. Porém, diante da discrepância dos valores e da fragilidade dos procedimentos (nem todo cientista de dados possui perfil no LinkedIn e nem todo membro do grupo Data Science Brasil é cientista de dados), optou-se por manter a população como infinita.

Assim, para se definir o tamanho necessário da amostra foi utilizado o G*Power, software focado em calcular o poder de diferentes testes estatísticos que é comumente utilizado em pesquisas das ciências sociais, comportamentais e biomédicas (ERDFELDER *et al.*, 2009). Para este cálculo foram seguidas as recomendações de Memon et al. (2020): 1) selecionar o tipo de teste como

[6] Disponível em https://www.facebook.com/groups/DataScienceMachineLearningBR.

“Regressão múltipla linear: modelo fixo, R^2 desvio de zero”, uma vez que que foi empregada análise fatorial confirmatória; 2) definir o tipo de análise como “a-priori” que calcula o tamanho necessário da amostra diante do fornecimento de α (nível de significância), do poder estatístico e do tamanho do efeito; 3) especificar o tamanho do efeito como 0,15 (efeito médio), α em 0,05 e poder estatístico em 0,80, que é a configuração recomendada para pesquisas em ciências sociais (FIELD; MILES; FIELD, 2012-) e; 4) definir o número de preditores, ou variáveis latentes, que nesta pesquisa são quatro (Tecnologia, Análise de Dados, Entendimento de Negócios e Competências Socioculturais). Para que o modelo estatístico atinja esse resultado, a amostra mínima deve conter 85 observações, conforme demonstrado na Figura 9:

FIGURA 26 – CÁLCULO DA AMOSTRA MÍNIMA NECESSÁRIA



FONTE: O autor, com base em Memon et al. (2020).

Segundo o critério adotado pelo G*Power, 85 respostas válidas seriam suficientes para a construção e fundamentação do modelo estatístico. Porém, Bruni (2013, p. 175) salienta que para a realização de inferências ou generalizações, o estudo estatístico deve contar com uma amostra probabilística com as mesmas proporções e características do todo estudado. O que não é o caso desta pesquisa. Além disso, sobre a definição do tamanho mínimo de amostra para uma análise

fatorial, Matsunaga (2010) aponta que 200 observações são suficientes e Hair et al. (2014) sugerem uma relação mínima de cinco observações para cada variável da estrutura fatorial. Como o número total de variáveis que compõem o modelo é 43, o valor mínimo de observações esperado para o questionário foi 215 respostas válidas.

5.2.2 Pesquisa Documental

Uma pesquisa documental é aquela que utiliza dados sobre uma população, porém sua obtenção não é realizada de maneira direta. Por meio de documentos, como livros, jornais, fotos, filmes, dentre outros registros, o pesquisador tem acesso a informações relacionadas ao problema de pesquisa que contribuem para a investigação do fenômeno sem as desvantagens da coleta feita diretamente realizada junto às pessoas, como falta de colaboração, desperdício de tempo e situações constrangedoras (GIL, 2008, p. 147).

Nesta pesquisa, duas fontes de documentos são utilizadas. Inicialmente, serão coletados os anúncios de vagas de emprego para cientistas de dados. Kim e Lee (2016, p. 162) argumentam que este tipo de anúncio tem se tornado um objeto de estudo recorrente em pesquisas sobre profissões, principalmente por conterem a terminologia relacionada à ocupação em questão. Conforme os autores, enquanto entrevistas e questionários mostram uma expectativa pessoal em relação às competências e tarefas profissionais, os anúncios de vagas de emprego apresentam uma tendência mais real por parte do mercado.

Para coletar estes anúncios de trabalho foram selecionados cinco *websites* destinados a este fim: Infojobs, Catho, Indeed, Vagas, Empregos e LinkedIn. Este último, embora sua principal característica seja ser uma rede social de negócios, possui uma seção de divulgação de anúncios de vagas de emprego. Como nenhuma das APIs (*Application Programming Interface*) dos *websites* pesquisados fornecia acesso aos dados de interesse da pesquisa (conteúdo detalhado na seção 5.4.2), foi necessária a utilização de raspagem de dados. Esta técnica, também conhecida como *web scraping*, corresponde ao uso de programas que percorrem páginas HTML procurando seções desejadas na estrutura de marcação e compilando os dados para a formação de um conjunto passível de ser analisado (LANTZ, 2015, p. 383).

Ainda que a raspagem de dados possa infringir os termos de utilização dos *websites* acessados, a técnica é considerada legal (FINDDATALAB.COM, 2020; WHITTAKER, 2022). As discussões acerca da prática se fundamentam, inicialmente, em potenciais prejuízos às plataformas acessadas, seja financeiro, seja de desempenho, mas principalmente por utilizar os dados para fins distintos daqueles inicialmente autorizados. Todavia, reforça-se que a presente pesquisa está de acordo com a Lei Geral de Proteção de Dados Pessoais (LGPD), uma vez que tem finalidade exclusivamente acadêmica, além de que todos os dados coletados passaram pelo processo de anonimização, isto é, foram utilizadas técnicas que impede a associação, direta ou indireta, a um indivíduo (Lei Nº 13.709/2018 - Lei Geral de Proteção de Dados Pessoais (LGPD)BRASIL, 2018).

Em seguida, o segundo conjunto de documentos foi formado pelo conteúdo programático dos cursos de formação de cientistas de dados. Para os cursos de graduação foi adotado o portal e-MEC como base de dados para a busca por registros que contenham "ciência de dados" ou "data science" em seus títulos. A coleta desse conteúdo foi realizada manualmente, uma vez que a diversidade dos *websites* consultados inviabilizou o procedimento de raspagem. Além disso, a quantidade encontrada possibilitou o processo manual. Por fim, os cursos livres em português foram buscados na plataforma Udemy, *marketplace* líder mundial em ensino à distância, juntamente com indicações obtidas pela pesquisa de levantamento (verificar detalhamento na seção 5.4.4).

Sendo assim, com a seleção dos documentos descritos, tem-se a complementação da primeira unidade de análise, os anúncios de vagas de emprego para cientistas de dados, da segunda unidade de análise, conteúdo de cursos superiores para cientistas de dados, e da terceira, conteúdo programático de cursos livres voltados à Ciência de Dados. Dessa maneira, segundo os critérios de Marconi e Lakatos (2003, p. 175–176), a pesquisa documental aqui realizada é considerada de fontes escritas, fontes primárias, uma vez que as informações são coletas e compiladas pelo autor, e de fato contemporâneo, visto que a pesquisa é realizada enquanto o fenômeno ocorre.

5.3 DEFINIÇÃO DO INSTRUMENTO DE COLETA DE DADOS

Conforme mencionado, para a coleta de dados submetidos a métodos quantitativos, optou-se pela realização de uma pesquisa de levantamento juntamente aos cientistas de dados atuantes no Brasil. Este tipo de pesquisa, também conhecida como *survey*, caracteriza-se por interrogar de forma direta um grupo significativo de pessoas, integrantes de um fenômeno a ser estudado (GIL, 2008, p. 55). Desta forma, com a devida aplicação de métodos de análise, obtém-se uma descrição quantitativa de tendências, atitudes ou opiniões da população em questão.

Dentre as vantagens da adoção da pesquisa de levantamento estão a economia e rapidez, uma vez que permite angariar uma grande quantidade de respostas em pouco tempo, a um custo baixo, especialmente pela utilização de questionários autoaplicáveis. O questionário, este importante elemento das pesquisas de levantamento, é definida por Gil (2008) como uma

> técnica de investigação composta por um conjunto de questões que são submetidas a pessoas com o propósito de obter informações sobre conhecimentos, crenças, sentimentos, valores, interesses, expectativas, aspirações temores, comportamento presente ou passado etc. (GIL, 2008, p. 121)

Além de econômico, Gil (2008, p. 121) cita outras vantagens da aplicação de questionários: a) acesso a indivíduos distantes geograficamente do local da pesquisa; b) garantia de anonimato dos respondentes; c) flexibilidade de horário, possibilitando que o respondente preencha o questionário quando lhe for mais conveniente e; d) os respondentes não são influenciados pelos entrevistadores. Por outro lado, a impossibilidade de ajuda perante dúvidas e diminuição na taxa de conclusão, especialmente em questionários longos, são características consideradas limitantes para dos questionários.

Para a presente pesquisa, foi adotado um levantamento com corte transversal, ou seja, os dados são coletados em um período específico (CRESWELL, 2010, p. 179), sendo as questões baseadas na literatura utilizada e em outras pesquisas de levantamento relacionadas. Dentre estas referências de projetos de levantamento, destaca-se a pesquisa *Data Science Survey*, conduzida por Hayes (2020) cujos resultados preliminares foram detalhados na seção 4.2.4. Hayes também autorizou,

via mensagem pela rede social Twitter, a utilização de seu questionário por esta tese. As outras pesquisas que serviram de referência foram os levantamentos de Harris, Murphy e Vaisman (2013-) e Kaggle (2020).

Para adequar as perguntas do questionário ao objetivo de identificar as competências dos cientistas de dados, foi feito um levantamento na literatura das competências associadas a este profissional. Como critério para manutenção das competências, foram selecionadas aquelas citadas por pelo menos três autores utilizados. Após a seleção, os itens restantes passaram por um processo de padronização de terminologia e agrupamento semântico que deu origem às cinco primeiras seções do questionário.

Essas cinco primeiras seções, são compostas por variáveis contínuas onde o respondente indica seu grau de proficiência em relação a cada item apresentado. As cinco categorias de proficiência são baseadas no questionário de Hayes (2020). Porém, enquanto o autor utiliza variáveis qualitativas e ordinais, na presente pesquisa, foi estabelecida uma escala com valores entre 0 e 10, configurando-se em variáveis quantitativas. Desta forma, uma série de métodos quantitativos de análise podem ser aplicados (FIELD; MILES; FIELD, 2012-, p. 9). Os níveis de proficiência, com o intervalo correspondente entre colchetes, são assim descritos:

- [>**0 a 2] Consciente:** quando o profissional tem consciência da existência do tópico ou possui um entendimento básico sobre técnicas e conceitos relacionados.
- **[>2 a 4] Novato(a):** possui experiência obtida em sala de aula e/ou cenários experimentais ou mesmo como estagiário. Porém, este profissional carece de ajuda na execução de tarefas relacionadas à competência em questão.
- **[>4 a 6] Intermediário(a):** o profissional com este nível de proficiência conclui com êxito as tarefas referentes ao tópico, geralmente de forma independente e apenas eventualmente, solicita ajuda de um especialista.
- **[>6 a 8] Avançado(a):** neste nível, o profissional. além de executar as ações associadas sem assistência, é reconhecido na organização como uma referência quando surgem perguntas difíceis sobre tópico referido.

- **[>8 a 10] Especialista:** o profissional deste nível pode fornecer orientação, solucionar problemas e responder a perguntas relacionadas a essa área de especialização e ao campo em que a competência é utilizada.

A escolha por uma escala numérica com valores entre 0 e 10, com a possibilidade de utilizar uma casa decimal, é justificada pela possibilidade de indicar níveis dentro das próprias categorias. Assim, de acordo com a autopercepção dos respondentes, há diferença entre um intermediário de uma competência com valor 4,1 e outro intermediário, com valor 5,8. Além disso, o respondente foi orientado a assinalar 0 somente quando não tivesse nenhum conhecimento sobre o item em questão. Todas as seções do questionário podem ser conferidas no Quadro 5:

QUADRO 5 – SEÇÕES DA PESQUISA DE LEVANTAMENTO

Nº	**Seção do Questionário**	**Descrição**	**Principais Referências**
1	Tecnologia	Com base nas seções “Tecnologia” e “Programação” do questionário de Hayes (2017), este grupo de competência busca mensurar a proficiência dos respondentes em relação ao uso da tecnologia em problemas de dados.	Harris, Murphy e Vaisman (2013), Hayes (2017), Linden (2018), Cao (2019)
2	Análise	Novamente duas seções do instrumento de coleta de dados de Hayes (2017), “Modelagem e Matemática” e “Estatística” são referência para a composição deste agrupamento que investiga as competências relativas a métodos de análise de dados.	Harris, Murphy e Vaisman (2013), Baškarada e Koronios (2017), Hayes (2017), Parks (2017), Cao (2019)
3	Entendimento de negócios	Grupo que visa mensurar os conhecimentos específicos ao segmento de atuação dos cientistas de dados que contribuem para o desenvolvimento de projetos de dados.	Tierney (2012), Harris, Murphy e Vaisman (2013), Hayes (2017)
4 e 5	Sociocultural	Questões relativas a elementos sociais importantes à ciência de dados, como conhecimento interdisciplinar e comunicação. Neste grupo, estão também questões relacionadas às características culturais das organizações e não dos respondentes diretamente, como valorização da ética e colaboração entre os integrantes das equipes de dados.	Patil (2011), Linden (2015, 2018), Cao (2017, 2019), Parks (2017), Kelleher e Tierney (2018), Rawlings-Goss (2019)
6	Informações Adicionais	Neste conjunto de questões, estão os itens associados à caracterização do respondente em relação a questões pessoais, educacionais, profissionais, além de permitir comentários sobre a pesquisa.	Hayes (2017), IBGE (2020), Kaggle (2020)

FONTE: O autor (2022).

A coluna “Principais Referências” apresenta os autores que mais citam as competências referentes ao grupo em questão. Todas as referências que formam os grupos de competências estão dispostas no **Apêndice 1**.

Uma vez que a seção de informações adicionais contém questões mais sensíveis como remuneração ou mesmo a identificação opcional por meio de endereço de e-mail, optou-se por posicioná-la ao final do questionário. Desta maneira, mesmo diante de possíveis desistências, esperava-se que ao menos as questões relativas às competências apresentassem uma maior taxa de preenchimento. De qualquer forma, os preenchimentos incompletos foram descartados, configurando-se como respostas inválidas para a pesquisa.

5.3.1 Aplicação de pré-teste

Para identificar possíveis falhas, inconsistências, ambiguidades, dentre outras dificuldades no preenchimento do questionário, o instrumento de pesquisa foi submetido a um pré-teste antes de sua utilização definitiva. Martins e Theóphilo (2009, p. 94) afirmam que a aplicação do questionário a uma amostra pequena, com cerca de 10 colaboradores, leva ao aprimoramento do instrumento de coleta de dados, aumentando sua confiabilidade e validade. Para os autores, a etapa de pré-teste é fundamental para que o instrumento atenda aos objetivos da pesquisa e garanta que se pretende descrever ou medir está sendo realmente descrito, ou medido.

Gil (2002, p. 120) reforça que a amostra deve ser composta por indivíduos que possuam as mesmas características do universo pesquisado e sugere um número entre 10 e 20 participantes. Além disso, o autor reforça os elementos mais relevantes a serem considerados no pré-teste: a) clareza e precisão dos termos; b) quantidade de perguntas; c) forma das perguntas; d) ordem das perguntas e; e) introdução. Para coletar esses aspectos, é necessário solicitar aos respondentes que relatem suas dificuldades durante o preenchimento do questionário.

Para a realização do pré-teste, foi definida uma amostra composta por 12 voluntários, sendo sete profissionais de dados (cientistas de dados, analistas de dados, estatístico e engenheiro de dados) e cinco professores universitários, das

áreas de Estatística, Gestão da Informação e Ciência de Dados. O período de resposta foi iniciado em 07 de junho de 2021 e encerrado no dia 15 do mesmo mês.

Dentre as alterações mais significativas decorrentes da aplicação do pré-teste, destacam-se:

- ajustes e correções ortográficos;
- reformulação de redação em questões que geraram dúvidas;
- adição de opções de resposta para cenários não previstos;
- remoção da aleatoriedade de questões – estava prejudicando o agrupamento de questões relacionadas, além de atrapalhar na retomada de um questionário iniciando anteriormente;
- exclusão de questão referente à área de ciência da computação que causou estranhamento em três respondentes.

Outros dois aspectos levantados pelos respondentes também foram alvo de reflexão. O primeiro foi o tamanho do questionário, considerado longo por dois participantes. Porém, uma vez que o pré-teste indicou um tempo médio de 22 minutos para completar o questionário, optou-se por manter o questionário em sua integralidade. Além disso, uma vez que dois respondentes levaram mais de uma hora e meia para completar o preenchimento, o que indica pausas no processo, o tempo médio final deve ser inferior ao encontrado no pré-teste.

Mesmo que não tenha sido alvo de reclamação, o segundo ponto de questionamento foi a utilização de *sliders* (controles deslizantes) para mensurar o nível de proficiência dos respondentes quanto ao item apresentado. A opção por este tipo recurso na interface do questionário se deve aos benefícios em relação ao uso de escalas estáticas com número definido de opções. Dentre esses benefícios, estão a maior variabilidade dos dados coletados que, devido à sua natureza contínua, aumenta a confiabilidade da escala e a diminuição do "efeito teto", onde os respondentes selecionam as principais opções (CHYUNG; SWANSON, 2019; COOK, 2013).

Além disso, procurou-se compensar os problemas do uso de *sliders* com elementos da interface do usuário. Dentre os possíveis prejuízos dos sliders estão: a) o aumento no tempo de preenchimento pela ação de arrastar e soltar (*drag-and-drop*); b) maior taxa de incompletude (usuários têm uma tendência maior a não interagir com

o elemento quando se compara com botões do tipo rádio) e; c) dificuldade de interação em dispositivos móveis (CHYUNG; SWANSON, 2019). Assim, para contornar esses problemas, o questionário contou com validação para evitar respostas em branco, *feedback* em tempo real de acordo com a movimentação do *slider* e escala de cores para facilitar e estimular o preenchimento.

5.3.2 Questões definitivas

A primeira seção é referente ao grupo de competências relacionadas ao uso de tecnologias em problemas de dados. Com 12 variáveis, juntamente com a seção de análise de dados, esse é o conjunto com maior número de itens refletindo sua ênfase na literatura da pesquisa. O Quadro 6 apresenta as variáveis da seção de tecnologia, conforme aparecem no questionário, indicando o nome a ser utilizado na etapa de análise:

QUADRO 6 – VARIÁVEIS DA SEÇÃO DE TECNOLOGIA

Nº	**O que medem**	**Variável**
1	Algoritmos (por exemplo, complexidade computacional, teoria da ciência da computação).	tecAlgoritmos
2	Administração de banco de dados.	tecBancoDeDados
3	Gestão de dados semiestruturados (por exemplo, SQL, JSON, XML).	tecDadosSemiEstruturados
4	Gestão de dados não estruturados (por exemplo, NoSQL, imagens, áudios, textos).	tecDadosNaoEstruturados
5	Desenvolvimento de softwares (isto é, planejar, executar, testar, implantar e manter sistemas de dados).	tecDesenvSoftware
6	Engenharia de dados (projetar, desenvolver e gerenciar a infraestrutura para projetos em dados).	tecEngDeDados
7	Gestão de dados (por exemplo, preparação, limpeza, integração de fontes de dados diferentes, raspagem da web).	tecGestaoDeDados
8	*Hacking* (habilidades técnicas e lógicas para construir soluções rápidas sem necessariamente possuir os fundamentos da Ciência da Computação).	tecHabHacking
9	Inteligência artificial (por exemplo, machine learning, deep learning, árvores de decisão, redes neurais, SVM).	tecAIML
10	Programação voltada a projetos de dados, seja *back-end*, *front-end* ou *full-stack*.	tecProgramacao
11	Administração de sistemas de informação.	tecSistemas
12	Sistemas de dados distribuídos de alto desempenho (por exemplo, Hadoop, MapReduce, Spark).	tecDadosDistribuidos

FONTE: O autor (2022).
NOTA: Todas as variáveis são consideradas contínuas e indicam o grau de proficiência do respondente referente ao item, variando de 0 (nenhum conhecimento) a 10 (especialista).

Na sequência, no Quadro 7 são demonstradas as 11 variáveis que compõem a segunda seção do questionário. Neste grupo, estão as competências relativas aos métodos de análise, majoritariamente quantitativos.

QUADRO 7 – VARIÁVEIS DA SEÇÃO DE ANÁLISE

Nº	**O que medem**	**Variável**
13	Análise de dados, de forma geral.	anAnaliseDeDados
14	Análises preditivas (tais como otimização de campanhas de marketing, detecção de fraude, redução de risco).	anAnalisesPreditivas
15	Estatísticas e modelagem estatística (por exemplo, modelo linear geral, ANOVA, MANOVA, modelo espaço-temporal, geoestatística).	anEstatistica
16	Matemática (tópicos como álgebra linear, análise real, cálculo, dentre outros).	anMatematica
17	Ferramentas de mineração de dados (como R, Python, SPSS, SAS, Weka, dentre outras).	anMineracaoDeDados
18	Modelagem em grafos (por exemplo, relacionamentos em redes sociais, definição de rotas, dentre outros).	anModelagemGrafos
19	Otimização (por exemplo, linear, inteira, convexa, global).	anOtimizacao
20	Processamento de linguagem natural (NLP) e mineração de texto.	anNLP
21	Formulação de questões que possam ser convertidas em problemas relacionados a dados passíveis de serem solucionados.	anFormQuestoes
22	Método científico (por exemplo, desenho experimental, projeto de pesquisa).	anMedotoCientifico
23	Visualização de dados (por exemplo, gráficos, mapas, visualização baseada na web, dentre outros).	anVisualizacao
24	Técnicas de regressão (linear, múltipla, logística).	anRegressao

FONTE: O autor (2022).
NOTA: Todas as variáveis são consideradas contínuas e indicam o grau de proficiência do respondente referente ao item, variando de 0 (nenhum conhecimento) a 10 (especialista).

Enquanto alguns autores (verificar **Apêndice 1**) se referem à competência de análise de dados de forma genérica, outros citam técnicas desta competência, gerando competências mais específicas. Assim, nota-se que a variável *anAnaliseDeDados* mede de forma geral a competência do respondente em analisar dados e engloba outras específicas como *anEstatistica* ou *anAnalisesPreditivas*. Optou-se por mantê-las, gerando inclusive a possibilidade de se verificar a correlação entre as variáveis do grupo. Neste grupo, também foram inseridas as questões relacionadas à proficiência em métodos científicos, como *anFormQuestoes* e *anMedotoCientifico*.

As variáveis da terceira seção do questionário correspondem ao grupo de competências relativas ao entendimento de negócios e são mostradas no Quadro 8:

QUADRO 8 – VARIÁVEIS DA SEÇÃO DE ENTENDIMENTO DE NEGÓCIOS

Nº	O que medem	Variável
25	Conhecimento em domínios específicos relativos à sua área de atuação (por exemplo, assistência médica, finanças, educação, imobiliário, automotivo) relevante para o seu trabalho.	enConhecimento
26	Projeto e desenvolvimento de novos produtos.	enDesenvolvimentoDeP
27	Gestão de Projetos em Ciência de Dados.	enGestaoDeProjetos
28	Desenvolvimento de novos negócios (capacidade de empreender).	enNegocios
29	Planejamento financeiro para projetos em Ciência de Dados.	enFinanceiro*
30	Governança de dados (definição de estruturas organizacionais, políticas e processos relacionados a dados).	enGovernanca*
31	Garantia de conformidade (*compliance*) com leis, normativas e demais regulamentações vigentes.	enCompliance*
32	Atendimento integral à LGPD.	enLGPD*

FONTE: O autor (2022).
NOTA: Todas as variáveis são consideradas contínuas e indicam o grau de proficiência do respondente referente ao item, variando de 0 (nenhum conhecimento) a 10 (especialista).

Embora o conhecimento de domínio seja uma das competências mais associadas ao cientista de dados (verificar **Apêndice 1**), poucas especificações deste elemento foram encontradas na literatura utilizada. Pelo menos, não a ponto de atender o critério de ser citada por três autores para compor o questionário da pesquisa. Por isso, nesta seção, foram adicionadas outras competências referentes a entendimento de negócios que são recorrentemente relacionadas à Ciência de Dados, identificadas na literatura. Dentre essas, estão desenvolvimento de produtos de dados, gestão de projetos e capacidade empreendedora.

Além dessas variáveis, três outras competências (assinaladas no Quadro 8 com asteriscos) foram sugeridas por especialistas no pré-teste. A primeira foi a capacidade de planejar financeiramente um projeto para Ciência de Dados (EUROPEAN E-COMPETENCE FRAMEWORK 3.0, 2014), seguida pela garantia de atender leis e regulamentações vigentes (DEMCHENKO *et al.*, 2019; GRANVILLE, 2014). Por fim, também foi sugerida uma inserção específica ao domínio de leis de

proteção de dados pessoais como a brasileira LGPD - Lei Geral de Proteção de Dados (Lei Nº 13.853/2018BRASIL, 2018) e europeia GDPR - *General Data Protection Regulation* (MANTELERO; VACIAGO, 2017; STODDER, 2018). Dessa forma, com oito variáveis, a seção referente ao entendimento de negócios em Ciência de Dados é a menor seção do questionário.

No Quadro 9, estão as variáveis de comportamentos socioculturais, componentes do penúltimo grupo de competências do questionário.

QUADRO 9 – VARIÁVEIS DA SEÇÃO DE COMPETÊNCIAS SOCIOCULTURAIS

Nº	**O que medem**	**Variável**
33	Conhecimento interdisciplinar que permite lidar com todos os aspectos de um problema referente a dados, da coleta às conclusões, interagindo com profissionais de diversas áreas.	socInterdisciplinari
34	Comunicação (por exemplo, compartilhamento de resultados a público não-técnico, comunicação oral e escrita, apresentações, redação de artigos).	socComunicacao
35	Liderança técnica.	socLidTec
36	Liderança estratégica.	socLidEst
37	Competência para solucionar qualquer tipo de problema de dados.	socSolucaoProblemas

FONTE: O autor (2022).
NOTA: Todas as variáveis são consideradas contínuas e indicam o grau de proficiência do respondente referente ao item, variando de 0 (nenhum conhecimento) a 10 (especialista).

Embora não seja propriamente uma competência, a interdisciplinaridade é uma das características da Ciência de Dados mais recorrentes na literatura (CAO, 2017; FINZER, 2013; LEY; BORDAS, 2018). Outras competências socioculturais relacionadas ao sucesso em Ciência de Dados são a comunicação (BÖRNER *et al.*, 2018; DAVENPORT *et al.*, 2015; LINDEN *et al.*, 2018; RAWLINGS-GOSS, 2019), a liderança (CAO, 2019; LINDEN *et al.*, 2018; RAWLINGS-GOSS, 2019) e a busca por solucionar problemas de dados (BAUMEISTER; BARBOSA; GOMES, 2020; IBM, 2020). As cinco primeiras variáveis socioculturais dentre as onze presentes no questionário podem ser aperfeiçoadas pelo indivíduo, logo é coerente manter a escala de proficiência utilizada até este ponto do questionário.

Todavia, para mensurar outros atributos socioculturais, foi orientado ao respondente que considerasse seu ambiente organizacional e indicasse seu grau de

concordância. As questões desse grupo, apresentadas no Quadro 10, correspondem à quinta e última seção do formulário relacionada a competências.

QUADRO 10 – VARIÁVEIS DA SEÇÃO DE COMPETÊNCIAS SOCIOCULTURAIS PROMOVIDAS PELA ORGANIZAÇÃO

Nº	O que medem	Variável
38	A colaboração é uma característica visível entre os membros das equipes de dados.	socColaboracao
39	Cientistas de dados são pessoas criativas que buscam abordagens inovadoras para extrair significado dos dados.	socCriatividade
40	A curiosidade e inquietação perante as possibilidades são atributos necessários para se trabalhar com Ciência de Dados.	socCuriosidade
41	A ética é princípio fundamental em todos os projetos que envolvem dados.	socEtica
42	Há constante preocupação com o tratamento de dados sensíveis.	socDadosSensiveis*
43	O risco de análises tendenciosas (bias) é foco contínuo de atenção.	socBias*

FONTE: O autor (2022).
NOTA: Todas as variáveis são consideradas contínuas e indicam o grau de concordância do respondente referente ao item, variando de 0 (nenhum conhecimento) a 10 (especialista).

As seis competências socioculturais investigadas sobre o contexto da organização começam pela colaboração, curiosidade e criatividade (HARRIS; MURPHY; VAISMAN, 2013-; KELLEHER; TIERNEY, 2018; LINDEN *et al.*, 2015, 2018; PARKS, 2017). Curiosidade e criatividade são questões que reforçam a mudança de escala, diante do pressuposto que o respondente não se avaliaria mal nestes itens. Todavia, considerando o ambiente organizacional, as respostas seriam menos tendenciosas. Os três últimos itens investigados no grupo sociocultural estão relacionados: ética, tratamento a dados sensíveis e atenção a vieses (FORGÓ *et al.*, 2021; KAMPAKIS, 2020; KELLEHER; TIERNEY, 2018). Destaca-se, ainda, que os dois últimos itens foram incorporados ao questionário durante a fase de pré-teste, por recomendação de especialistas.

Finalmente, as variáveis referentes a informações adicionais de caracterização são apresentadas no Quadro 11:

QUADRO 11 – VARIÁVEIS DA SEÇÃO DE INFORMAÇÕES ADICIONAIS

Nº	O que medem	Variável e Natureza	Subseção
44	O entendimento sintético do respondente em relação à Ciência de Dados.	**caDescCDT1, caDescCDT2, caDescCDT3** – Qualitativa, texto.	Informação adicional
45	A idade do respondente.	**caIdade** – Quantitativa, discreta e aberta.	Informações pessoais
46	O gênero do respondente com a opção "Prefiro não dizer" ou campo para sugestão de autoidentificação.	**caGenero** – Qualitativa, nominal e fechada.	
47	Cidade atual do respondente.	**caCidade** – Qualitativa, nominal e aberta.	
48	Estado atual do respondente.	**caUF** – Qualitativa, nominal e fechada.	
49	O mais alto nível de educação formal obtido pelo respondente.	**caNivelEducacional** – Qualitativa, ordinal e fechada.	Informações educacionais
50	O(s) curso(s) de graduação do respondente.	**caCursoGraduacao** – Qualitativa, texto.	
51	Possível(is) curso(s) de pós-graduação do respondente.	**caPosGraduacao** – Qualitativa, texto.	
52	Formação adicional relativa à Ciência de Dados.	**caFormacao** – Qualitativa, texto.	
53	Função atual do respondente.	**caFuncao** – Qualitativa, nominal e fechada.	Informações profissionais
54	Metodologia para projetos em Ciência de Dados.	**caMetodologia** – Qualitativa, nominal e fechada.	
55	Anos de experiência no trabalho com dados.	**caExperiencia** – Quantitativa, discreta e aberta.	
56	Remuneração financeira mensal do respondente.	**caRemuneracao** – Qualitativa, nominal e fechada.	
57	Número de colaboradores da organização.	**caColaboradores** – Qualitativa, ordinal e fechada.	Informações organizacionais
58	Número de pessoas que atuam diretamente como cientista de dados ou cargos análogos.	**caEquipeDados** – Quantitativa, discreta e aberta.	
59	Segmento no qual a organização atua.	**caSegmento** – Qualitativa, texto.	
60	Endereço eletrônico do respondente.	**caEmail** – Qualitativa, texto.	Identificação
61	Espaço livre para comentários adicionais sobre a pesquisa.	**caComentarios** – Qualitativa, texto.	Comentários

FONTE: O autor (2022).

Inicialmente, seguindo a questão proposta por Hayes (2020), é solicitado ao respondente que indique sinteticamente qual seu entendimento sobre Ciência de Dados. Hayes (2020) pede em seu questionário que o respondente utilize apenas uma palavra, mas aqui, permitiu-se o uso de até três termos para responder à questão. As questões de informações pessoais, como idade e gênero, além da cidade e estado dos respondentes, permitem investigar diferenças entre grupos perante remuneração, formação, função, dentre outros critérios, conforme pesquisas anteriores (BURTCH WORKS, 2019; PWC, 2017).

Da mesma forma, as questões relativas às informações educacionais permitirão identificar padrões na formação dos profissionais que atuam com Ciência de Dados. Na sequência, são mensuradas questões relativas à atuação profissional, como função, remuneração e tempo de experiência, cujos relacionamentos serão também verificados. Ainda na caracterização, estão as questões referentes à organização na qual o respondente atua. Verificando o porte pelo número de colaboradores, o tamanho das equipes de dados e o setor de atuação.

Para a remuneração do respondente, por recomendação dos respondentes do pré-teste a questão foi alterada de uma pergunta aberta (variável contínua) para uma pergunta fechada, com opções de seleção, gerando uma variável categórica. Os intervalos definidos foram baseados no salário-mínimo nacional, com as opções para o respondente não querer responder ou, ainda, informar que não possui remuneração no momento.

De maneira similar, também se optou por uma variável categórica para mensurar o número de colaboradores. Como referência, foi seguido o critério do Cadastro Central de Empresas fornecido pelo IBGE - Instituto Brasileiro de Geografia e Estatística (2020) que separa as empresas em quatro faixas: (1) com até nove pessoas; (2) de 10 a 49 pessoas; (3) de 50 a 249 pessoas e; (4) com mais de 249 pessoas. Este caminho foi escolhido considerando que o número de exato de colaboradores nem sempre é uma informação acessível o que dificultaria o preenchimento desta questão. Pela mesma dificuldade de acesso à informação, decidiu-se não questionar sobre o faturamento das organizações, utilizando-se apenas o número de colaboradores para indicar o porte das empresas.

Com a descrição do instrumento de coleta de dados, a próxima seção detalha os procedimentos de coleta e tratamento dos dados.

5.4 COLETA E TRATAMENTO DE DADOS

Nesta seção, são descritos os passos de coleta e tratamento dos dados da pesquisa. Inicialmente, são abordados os dados da pesquisa de levantamento e relatados os processos de escolha da plataforma do questionário, de divulgação, controle e tratamento do conteúdo coletado. Em seguida, são detalhados os procedimentos realizados para coleta dos anúncios e dos conteúdos curriculares dos cursos analisados.

5.4.1 Pesquisa de levantamento

Duas etapas basilares da pesquisa de levantamento já foram detalhadas anteriormente: a formulação do questionário e a aplicação do pré-teste. A seguir são detalhados os procedimentos de desenvolvimento da plataforma de coleta de dados, o processo de divulgação e acompanhamento, bem como do tratamento do conteúdo obtido. As etapas da pesquisa de levantamento, anteriores à fase de análise dos dados, estão dispostas no Quadro 12.

QUADRO 12 – ETAPAS DE COLETA E TRATAMENTO DOS DADOS DO QUESTIONÁRIO

Etapa	**Descrição**	**Materiais**
Formulação do questionário	Formular e organizar as questões com o objetivo de mapear as competências dos profissionais da Ciência de Dados atuantes no Brasil, com base em pesquisas de levantamento preliminares e na literatura da área.	Revisão teórica.
Aplicação de pré-teste	Como avaliação preliminar do questionário, a ferramenta foi enviada a 12 indivíduos com características similares à população pesquisada: sete profissionais de dados e cinco professores universitários. O objetivo foi identificar falhas na redação, complexidade das perguntas, necessidade de retirar ou adicionar questões, tempo de preenchimento (GIL, 2008).	LimeSurvey. E-mail.
Ajustes e correções	Alterações e correções na ferramenta de coleta de acordo com o resultado do pré-teste.	LimeSurvey.
Divulgação	Divulgação da pesquisa por meio de posts em grupos e perfis de redes sociais.	Post em redes sociais (Facebook, LinkedIn, WhatsApp e Telegram).
Acompanhamento de respostas	Controle no número de respostas obtidas e investigação de eventos adversos (preenchimento incorreto, abandono de pesquisa etc.). Formas alternativas de divulgação, destacando-se o uso de mensagem direta e individual pelo LinkedIn.	LimeSurvey. Mensagem direta pelo LinkedIn.
Tratamento dos dados	Preparação dos dados para os procedimentos de análise como exclusão de preenchimento incompleto, formas “anormais” de preenchimento, bem como padronização e categorização de respostas.	Excel, Python.

FONTE: O autor (2022).

Após a definição das questões, optou-se pela versão comunitária da plataforma de questionário LimeSurvey, uma solução *web*, de código aberto e totalmente gratuita (LIMESURVEY, 2020). Foi utilizada a versão 3.27.1 do software, instalado em um servidor Linux, utilizando PHP versão 7.3.28 e banco de dados MySQL 5.6.36-82.0. Para que a plataforma atendesse plenamente às demandas da pesquisa, especialmente em relação à utilização de sliders, foi desenvolvido um novo tema, baseado em princípios de design, visando a melhor experiência dos usuários. Embora a UFPR utilize o LimeSurvey como ferramenta padrão para suas pesquisas de levantamento (https://questionarios.ufpr.br/), a versão adotada é 2.00+, e não permite a instalação de novos temas. Dessa forma, foi providenciada, além de uma hospedagem particular, a compra do domínio https://cienciadedados.info, no qual a pesquisa ainda se encontra acessível.

Com o questionário concluído, a divulgação da pesquisa de levantamento foi inicialmente realizada pelo envio de mensagem em grupos de redes sociais cujo título continham a expressão “ciência de dados” ou “data science”. Além disso, foram selecionados apenas grupos cujo idioma era português, para ficar coerente à população estudada, e continham mais de dois mil membros cadastrados. Em busca realizada no dia 19 de outubro de 2020, foram encontrados oito grupos na rede social Facebook que atendiam aos critérios estabelecidos e apenas um registro foi encontrado no LinkedIn. A busca com o termo “ciência de dados” não retornou nenhum registro com mais de 2 mil integrantes, porém, ao se buscar por “data science”, foi identificado o grupo “BI e Ciências de Dados / Data Science” com 2.286 membros. Nota-se que o título adotou “Ciências” no plural, sendo assim desconsiderado pela primeira busca nesta rede social.

A divulgação começou em 02 de julho de 2021, quando foram verificados os números atualizados de membros dos grupos. Esse processo se repetiu em 05 de maio de 2022, sendo que os valores obtidos são encontrados no Quadro 13. Nota-se que todos os grupos aumentaram a quantidade de membros nas datas pesquisadas, onde a média de aumento foi de 21,94% para o intervalo 2020/2021 e, 13,53%, para 2021/2022. O maior grupo, cujo título “Data Science Brasil” é compartilhado por outros dois grupos, conta com mais de 32 mil membros na verificação mais recente, tendo crescido 38,94% em relação a outubro de 2020.

QUADRO 13 – GRUPOS DE CIÊNCIA DE DADOS NO FACEBOOK E NO LINKEDIN

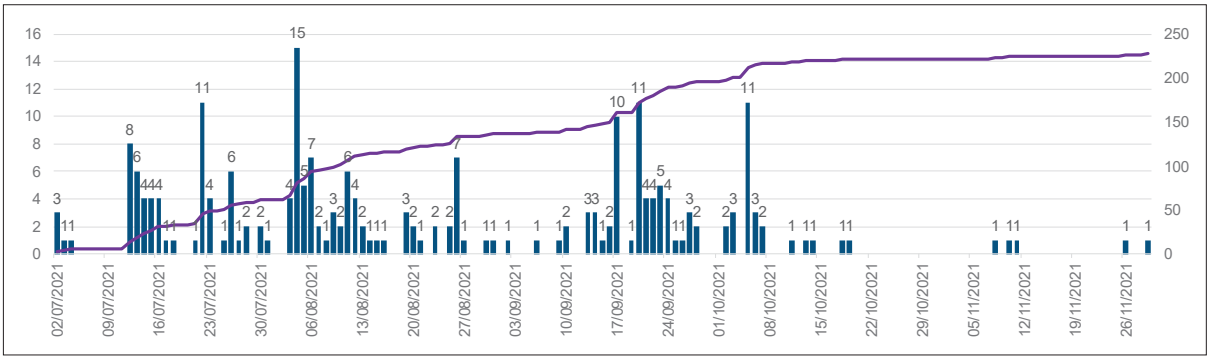
Nome	Privacidade	Out/2020	Jul/2021	↑2020/2021	Maio/2022	↑2021/2022	Rede Social	URL
Data Science Brasil	Privado	23.635	30.968	31,03%	32.838	6,04%	Facebook	https://www.facebook.com/groups/DataScienceMachineLearningBR
R e Python aplicados a Ciência de Dados	Público	13.039	16.172	24,03%	17.039	5,36%	Facebook	https://www.facebook.com/groups/1870674756507304
Data Science Brasil	Público	10.213	12.368	21,10%	14.720	19,02%	Facebook	https://www.facebook.com/groups/1976050415995026
Ciência de Dados Brasil	Público	8.089	8.769	8,41%	11.512	31,28%	Facebook	https://www.facebook.com/groups/cienciadedados/
Ciência de Dados Brasil	Privado	7.400	7.926	7,11%	8.085	2,01%	Facebook	https://www.facebook.com/groups/CienciaDadosBR/
Ciência de dados na veia	Privado	5.700	6.545	14,82%	7.417	13,32%	Facebook	https://www.facebook.com/groups/cienciadedadosnaveia/
Data Science Brasil	Público	2.494	3.102	24,38%	3.529	13,77%	Facebook	https://www.facebook.com/groups/2994360697275625/
Big Data, Data Science and IOT.	Privado	2.071	2.403	16,03%	2.708	12,69%	Facebook	https://www.facebook.com/groups/122514451732318/
BI e Ciências de Dados / Data Science	Privado	2.286	3.442	50,57%	4.070	18,25%	LinkedIn	https://www.linkedin.com/groups/3971821/

FONTE: O autor (2022).

Mesmo com dezenas de milhares de membros, entre 02 e 15 de julho de 2021, a divulgação em grupos sobre Ciência de Dados se mostrou pouco efetiva para a presente pesquisa, gerando poucas respostas válidas. No próprio “Data Science Brasil”, com seus mais de 30 mil membros à época, a publicação[7] produziu 21 reações do tipo “curtida” e quatro comentários de pessoas de fora da pesquisa. Por isso, foi necessário adotar novas estratégias para obter novos respondentes. Inicialmente, as novas formas de divulgação começaram pela publicação em grupos de WhatsApp e Telegram, dentre os quais se destaca o “Pt-BR Data Science & Python” com mais de 5.300 em maio de 2022.

Todavia, o que realmente surtiu efeito para a participação de profissionais na pesquisa foi o envio de mensagens diretas por meio do LinkedIn. Esse processo foi realizado de forma individualizada, onde, em um primeiro momento, era feita uma solicitação de conexão juntamente com um texto com menos de 300 caracteres (limite para este tipo de ação na rede social). Em caso de aceite, uma mensagem com mais detalhes era enviada. O recurso que apresentou resultados promissores esbarrou nos controles de segurança do LinkedIn que identificaram este comportamento como anormal, onde, por vezes, o limite semanal de solicitações foi ultrapassado e houve restrição no envio de novos convites. Mesmo assim, essa forma de divulgação se mostrou fundamental, visto que a origem de 66,08% das 227 respostas válidas da amostra foi a rede social LinkedIn. A Figura 27 demonstra a evolução da coleta de dados da pesquisa de levantamento.

FIGURA 27 – EVOLUÇÃO DA COLETA DE DADOS DA PESQUISA DE LEVANTAMENTO



FONTE: O autor (2022).

[7] Acessível, mediante participação no grupo, pelo endereço eletrônico: https://fb.com/groups/DataScienceMachineLearningBR/permalink/2960984447490322/

Pelo gráfico, percebe-se que a divulgação em grupos do Facebook, iniciada em 02 de julho de 2021, trouxe cinco participações. O ritmo foi intensificado a partir do dia 12 do mesmo mês, quando se começa o envio de mensagens diretas por meio do LinkedIn. A coleta da amostra foi encerrada em 29 de novembro de 2021, com um total de 457 participações iniciadas. Porém, cerca de metade desse número concluiu o preenchimento, resultado em 232 respostas completas. O tempo médio para completar a pesquisa foi de 22 minutos e 58 segundos. Porém, se desconsideradas as participações que duraram mais de uma hora, ou seja, onde há indicativo de pausa durante o preenchimento, este tempo cai para 13 minutos e 36 segundos.

Das 232 respostas completas, duas foram descartadas porque as variáveis quantitativas, aquelas que compõem o modelo de competências, apresentaram desvio padrão inferior a 0,5, pois este comportamento indica que o respondente colocou o mesmo valor para todas as questões. Outros três registros também foram excluídos por serem de profissionais que não residem no Brasil: dois residem em Portugal e o terceiro não quis informar sua localidade. Assim, a amostra final para análises quantitativas é composta por 227 observações, acima do mínimo estipulado de 205.

Com essa quantidade de respostas e nível de significância de 0,05, a amostra obtida proporciona um tamanho de efeito 0,1, classificado como médio, e um poder estatístico de 0,97, para quatro construtos (ERDFELDER *et al.*, 2009; FIELD; MILES; FIELD, 2012-). Considerando que o modelo proposto tem 43 variáveis observáveis, o valor da relação entre observações e variáveis é de aproximadamente 5,28, também acima do mínimo recomendado (HAIR *et al.*, 2014). Entretanto, ressalta-se que a amostra é considerada não probabilística e não permite generalização dos resultados, ainda que colabore para a investigação de uma população que não é totalmente conhecida (BRUNI, 2013; SAMPIERI; COLLADO; LUCIO, 2013).

Com a definição da amostra, os dados coletados passaram por alguns procedimentos de tratamento, especialmente nas questões abertas. Por isso, as questões referentes aos termos que definem a Ciência de Dados, cidade, curso de graduação, pós-graduação, formação e segmento de atuação precisaram passar por um processo manual de padronização. Nesse processo, foi necessário identificar e padronizar as instituições de educação citadas pelos respondentes. Por exemplo, a Universidade Federal do Paraná foi citada por extenso, mas também pela sigla UFPR. Outros casos que precisaram de tratamento manual foram os respondentes com mais

de uma graduação. Nestas situações, eram separados os cursos e as instituições de ensino.

Posteriormente, novas variáveis foram criadas para possibilitar novos processos de análise. Para cara grupo do questionário (Tecnologia, Análise de Dados, Entendimento de Negócios e Competências Socioculturais) foi cria uma variável específica contendo a média das respostas fornecidas para o grupo. Por exemplo, a variável “tecnologia” contém a média das respostas referentes ao grupo de questões relacionadas à tecnologia. Além disso, segundo esta média cada respondente foi classificado em relação ao seu nível de proficiência no grupo específico. Assim, a variável “tecnologiaCat” indica o nível de proficiência do respondente (Consciente, Novato(a), Intermediário(a), Avançado(a) ou Especialista) para o grupo tecnologia.

Por fim, também seguindo a escala definida no questionário, para cada variável numérica do modelo foi criada uma correspondente categórica. Dessa forma, se o respondente respondeu que seu nível de proficiência em processamento de linguagem natural, cuja variável é “anNLP”, é 7,5, este registro contará com o valor “Avançado” para a variável “anNLPCat”. Ou seja, variável “anNLPCat” é a correspondente categórica da variável “anNLP”. Dessa maneira, novos tipos de visualização foram proporcionados.

5.4.2 Anúncios

Nesta seção, os passos metodológicos adotados para coletar os anúncios de vagas de emprego para cientistas de dados são sucintamente descritos, desde a estratégica de busca aos procedimentos de análise utilizados. Em geral, a pesquisa com os anúncios seguiu a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*), conjunto de tarefas voltadas a projetos em Mineração de Dados (BEDREGAL-ALPACA; ARUQUIPA-VELAZCO; CORNEJO-APARICIO, 2020, p. 595; CHAPMAN *et al.*, 2000, p. 10) e considerada uma das práticas mais populares entre os profissionais da área (PIATETSKY-SHAPIRO, 2014). Naturalmente, em uma pesquisa científica, as etapas da CRISP-DM precisaram ser adaptadas, uma vez que o resultado não é necessariamente uma implementação.

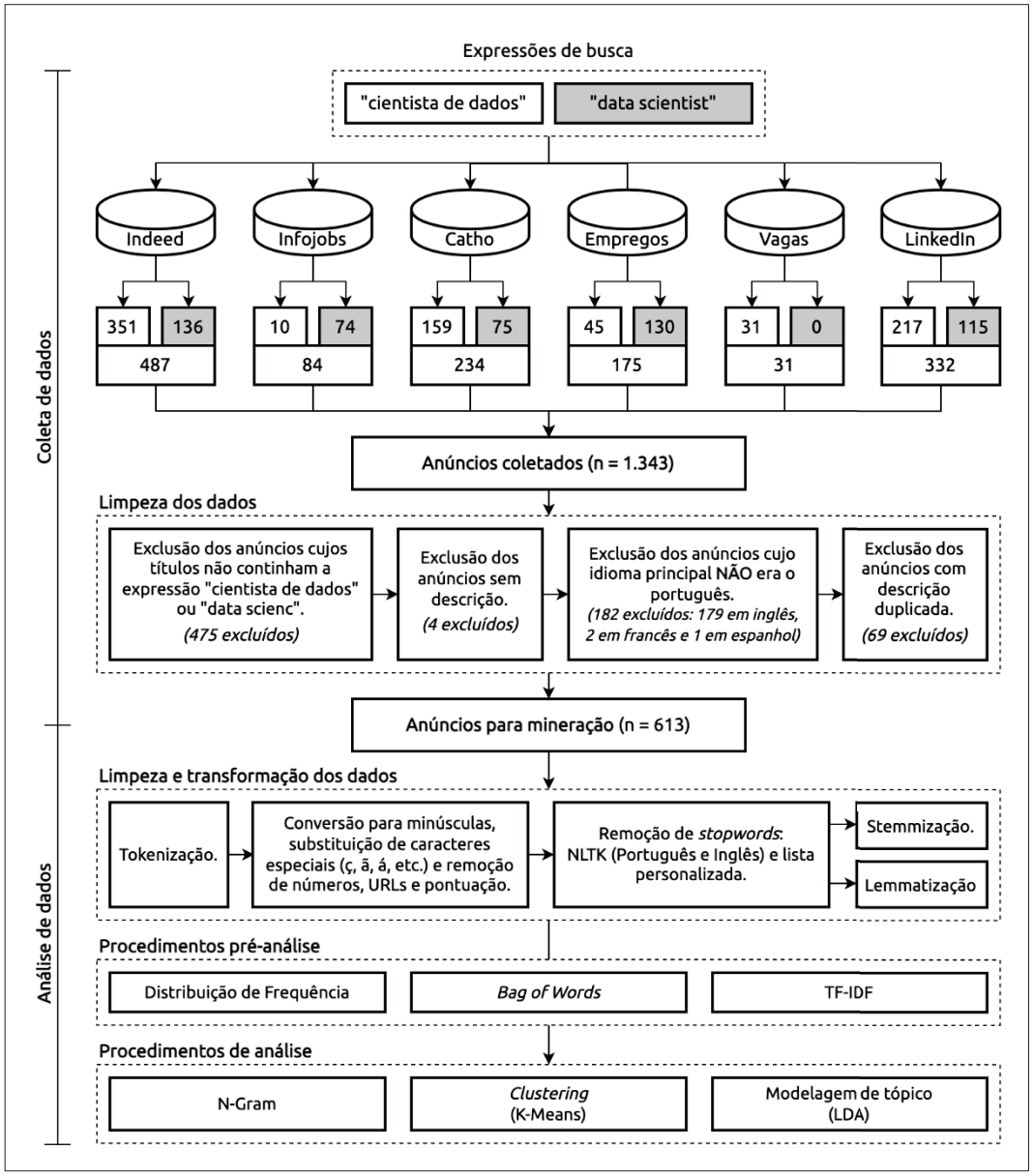
As fontes de dados da pesquisa foram *websites* especializados na divulgação de vagas voltadas à área de tecnologia, selecionados a partir da experiência do autor e do resultado de pesquisa, com a expressão “vagas de emprego”, realizada no

mecanismo de busca Google. Foram selecionados três *websites* estrangeiros com versões voltadas ao público brasileiro (LinkedIn, Indeed e Infojobs) e três empresas brasileiras destinadas ao mesmo fim (Catho, Empregos e Vagas). Nesta etapa, correspondente ao entendimento de negócio proposto pela CRISP-DM, procurou-se por APIs (*Application Programming Interface*) nos *websites* definidos, a fim de se verificar a disponibilidade de dados de interesse da pesquisa pelas plataformas consultadas. Uma vez que nenhuma API encontrada fornecia os dados requeridos, optou-se por empregar técnicas de raspagem de dados para acesso e coleta do material. A raspagem de dados, também conhecida como *web scraping*, corresponde ao uso de programas que percorrem páginas HTML procurando informações desejadas na estrutura de marcação e compilando os dados para a formação de um conjunto passível de ser analisado (LANTZ, 2015).

Assim, como parte do entendimento dos dados, foram identificados quais informações dos anúncios de emprego poderiam ser extraídas pelo procedimento de raspagem: endereço URL do anúncio (*jobUrl*), título (*jobTitle*), resumo (*snippet*), descrição (*jobDescription*), nome da empresa (*companyName*), localidade da vaga (companyLocation), avaliação da empresa (*companyRating*), URL da empresa (*companyUrl*), remuneração (*salary*) e data de cadastro (*date*). Ainda que nem todos os seis *websites* e nem todas as vagas forneciam todas essas informações, sempre que possível, esses foram os dados coletados.

Os programas web scraper foram desenvolvidos em linguagem de programação Python, sendo adaptados para cada website acessado, visto que a estrutura HTML de cada um é única. Além disso, para todos os websites, a raspagem foi dividida em duas etapas: 1) coleta de todos os anúncios com dados presentes na listagem (título, URL e resumo) e; 2) acesso individual a cada anúncio para coletar os demais dados disponíveis. Para buscar os anúncios, em cada website foram utilizadas as expressões “cientista de dados” e “data scientist” (dado que muitas empresas brasileiras utilizam termos em inglês em seus anúncios), filtrando por vagas no território brasileiro. Nesta etapa, entre 21 e 29 de setembro de 2021, foi possível coletar um total de 1.343 anúncios, cujos valores individuais de cada base são mostrados na Figura 28.

FIGURA 28 – ETAPAS DE COLETA DE DADOS E PROCEDIMENTOS DE ANÁLISE DOS ANÚNCIOS DE VAGA DE EMPREGO



FONTE: O autor (2022).

Uma vez que os mecanismos de busca dos *websites* não são atrelados aos títulos dos anúncios, percebeu-se que muitas vagas encontradas não se destinavam a cientistas de dados. Por isso, como primeiro procedimento de tratamento, foram excluídos 475 anúncios que não continham nem “cientista de dados” e nem “data scien” em seus títulos. Em seguida, foram excluídos também quatro anúncios que não possuíam descrição, campo utilizado na mineração de texto, e 182 anúncios cujo idioma principal não era o português. Por fim, também foram removidos os anúncios com descrições duplicadas, uma vez que a mesma empresa poderia ter várias vagas para um cargo. Anúncios com descrições duplicadas afetariam a frequência com que

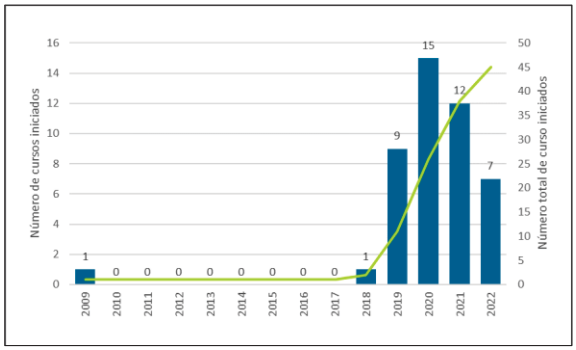
os termos aparecem na mineração. Desse modo, ao final desses procedimentos, relacionados à “Preparação dos dados” da CRISP-DM, o *corpus* a ser analisado apresentava 613 documentos.

5.4.3 Cursos superiores

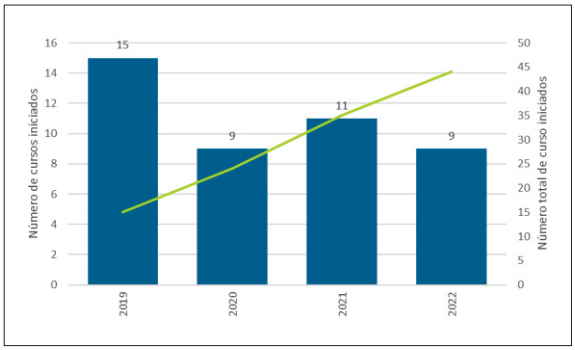
Para a coleta dos conteúdos relacionados à educação superior em nível de graduação, foi tomada como base o Cadastro Nacional de Cursos e Instituições de Educação Superior (e-MEC) (http://emec.mec.gov.br). No e-MEC, foram buscados cursos de graduação com os termos “ciência de dados” ou “data science” em seus títulos. Após o resultado obtido, os registros foram exportados para o formato de planilha eletrônica (Excel) que continha um total de 28 dados referentes aos cursos, dentre eles a instituição de educação superior (IES), código da IES, sigla da IES, código e nome do curso, grau, modalidade, número de vagas autorizadas, data do início de funcionamento, situação atual do curso, dentre outros.

Em busca atualizada em 1 de junho de 2022, foram retornados um total de 89 cursos de graduação, sendo que 70 (78,65%) registros correspondiam ao termo “ciência de dados” e 19 (21,35%), a “data science”. Destes cursos, 30 (33,71%) eram de grau de bacharelado e 59 (66,29%) se referiam a formações tecnológicas, enquanto 49 (55,06%) eram de modalidade presencial e 40 (44,94%) correspondiam ao ensino a distância. No entanto, ao filtrar pelos cursos que já tinham iniciado seu funcionamento e continuavam em atividade, o número total de cursos caiu de 89 para 45 registros.

FIGURA 29 – INÍCIO DE FUNCIONAMENTO E CRIAÇÃO DOS CURSOS SUPERIORES



(a) Início de funcionamento dos cursos em atividade

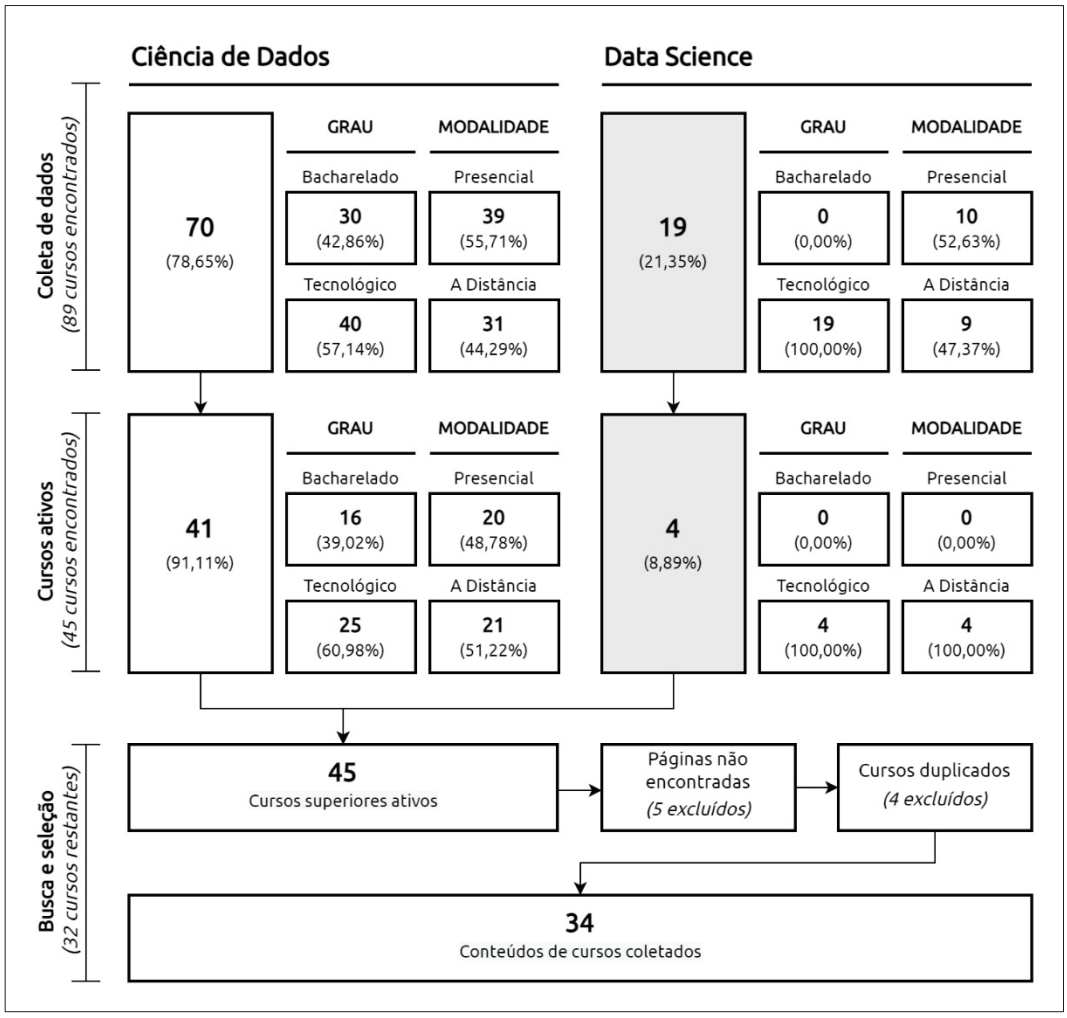


(b) Anos de criação dos cursos não iniciados

FONTE: O autor (2022).

De qualquer maneira, conforme demonstra Figura 29, nota-se uma intensificação na criação de novos cursos a partir de 2019, seja para cursos ativos, seja para cursos registrados no e-MEC que ainda não iniciaram suas atividades. Dos cursos ativos, a proporção de cerca de 1/3 entre bacharelado (16 – 35,56%) e tecnológico (29 – 64,44%) se manteve. Por outro lado, a modalidade de cursos à distância é maioria dos cursos ativos: 25 (55,56%) contra 20 (44,44%). A Figura 30 demonstra o processo de coleta e filtragem dos cursos de graduação obtidos a partir do e-MEC:

FIGURA 30 – ETAPAS DE COLETA DE DADOS REFERENTE AOS CURSOS DE NÍVEL SUPERIOR



FONTE: O autor (2022).

Para coletar os dados dos cursos, foi necessário um procedimento manual diretamente na página web de cada curso, em cada instituição de ensino. Uma vez que a forma de apresentação dos cursos é única a cada website visitado, a possibilidade de raspagem de dados foi descartada. Além disso, a quantidade de 45

registros foi considerada viável de se coletar individualmente. Assim, as páginas dos cursos foram procuradas por meio do mecanismo de busca do Google, utilizando o nome do curso combinado com a sigla ou o nome da instituição. Adicionalmente, quando o curso não era identificado, mas o site da IES sim, era realizada uma busca combinando os termos "ciência de dados" ou "data science", filtrando pelo domínio da instituição. Por exemplo, a busca seguinte procura páginas que contenham a expressão exata "ciência de dados" no domínio da UFPR:

$$\text{“}ciência\ de\ dados\text{”}\ site{:}\ ufpr.br$$

Nesse processo, porém, houve a exclusão de 11 registros contidos na planilha exportada: cinco cursos exportados não foram localizados nos sites das instituições e seis representaram conteúdo duplicado, onde a instituição continha a mesma página para cursos presenciais e à distância, por exemplo. Nestes casos, embora para o MEC fossem cursos distintos, apenas um deles foi considerado para esta pesquisa. Todos os cursos cujo conteúdo foi identificado e coletado estão listados no Apêndice 3.

Os procedimentos de análise empregados são os mesmos utilizados para analisar os anúncios de emprego. Isto é, os conteúdos coletados passam pelas etapas de limpeza e transformação dos dados, procedimento de pré-análise (distribuição de frequência, *bag of words* e TD-IDF) e procedimentos de análise (n-gram, *clustering* e modelagem de tópicos), conforme já apresentado na Figura 28.

5.4.4 Cursos livres

Cursos livres correspondem a uma categoria de educação não-formal e, consequentemente, não regulamentada pelo Ministério da Educação (MEC). Todavia, mesmo não estando sob uma regulamentação específica, este tipo de ensino é considerado válido, legal e é ofertado em modalidade presencial ou à distância. O objetivo desses cursos é a formação inicial e continuada ou qualificação e, comumente, resultam em certificados que trazem benefícios ao profissional, ainda que a instituição emissora não seja reconhecida pelo MEC.

Para Ciência de Dados, Cao (2017) afirma que um dos gêneros de cursos livres que mais crescem no mercado são cursos online abertos em larga escala, ou *Massive Open Online Course* (MOOC). Dentre as plataformas destacadas pelo autor, encontra-

se a Udemy, *marketplace* líder mundial em ensino e aprendizado que conecta alunos e instrutores ao redor do mundo. Segundo o site oficial da plataforma, em dezembro de 2021, a Udemy possuía uma comunidade com mais de 52 milhões de alunos, mais de 68 mil instrutores, 196 mil cursos e 712 milhões de matrículas, em 75 idiomas diferentes (UDEMY, 2022).

Por esses números, e por apresentar cursos no idioma português, a Udemy foi a primeira fonte de dados para coletar os conteúdos de cursos em Ciência de Dados. Novamente, a busca, realizada em 22 de abril de 2022, baseou-se nos termos "ciência de dados" e "data science", filtrando os resultados pelo idioma português. Foram retornados 50 cursos para "ciência de dados" e 106, para "data science", totalizando 156 registros. Porém, após a remoção de duplicados, foi possível exportar o conteúdo de 142 cursos. Para essa etapa foi utilizada uma combinação de *scripts* Python com o complemento do navegador Chrome chamado Data Miner (2021), necessário para contornar os recursos de segurança contra raspagem de dados impostos pelo portal da Udemy.

Além disso, foram consideradas as formações em Ciência de Dados indicadas pelos respondentes na pesquisa de levantamento. As formações indicadas por mais de um respondente foram: Data Science Academy (13 indicações), Coursera (9 indicações), IGTI (6 indicações), Udacity (5 indicações), PUC-MG (3 indicações) e DataCamp (2 indicações). Enquanto IGTI e PUC-MG foram desconsideradas por serem educação formal de especialização, Udacity e DataCamp não oferecem cursos em português.

No website Coursera, considerado um concorrente da Udemy, a expressão "ciência de dados" retornou 25 cursos enquanto "data science" identificou 967 resultados, em 01 de junho de 2022. Em ambas as situações, utilizou-se o filtro para idioma português que, neste caso, apresentou três variações: Português, Português (Portugal) e Português (Brasil). Todavia ao analisar esses registros, foram identificados apenas sete cursos produzidos em português, os demais continham eram produzidos em outros idiomas e eram legendados. Todos os cursos identificados foram produzidos pela mesma instituição: Fundação Instituto de Administração (FIA), empresa que atua nas áreas de consultoria, educação executiva e pesquisa.

Uma vez que os respondentes não indicaram quais cursos ou formações haviam feito, a seleção foi realizada arbitrariamente, privilegiando os cursos com "ciência de dados" ou "*data science*" nos títulos e que fossem o mais amplo possível.

Por isso, para o portal IGTI, foi considerado o *bootcamp*, uma espécie de treinamento de curto período e alta densidade de conteúdo, intitulado “Bootcamp Cientista de Dados”. Na Data Science Academy, foram selecionados três cursos que melhor atendiam aos critérios definidos. O Quadro 14 traz todos os cursos derivados das indicações da pesquisa de levantamento:

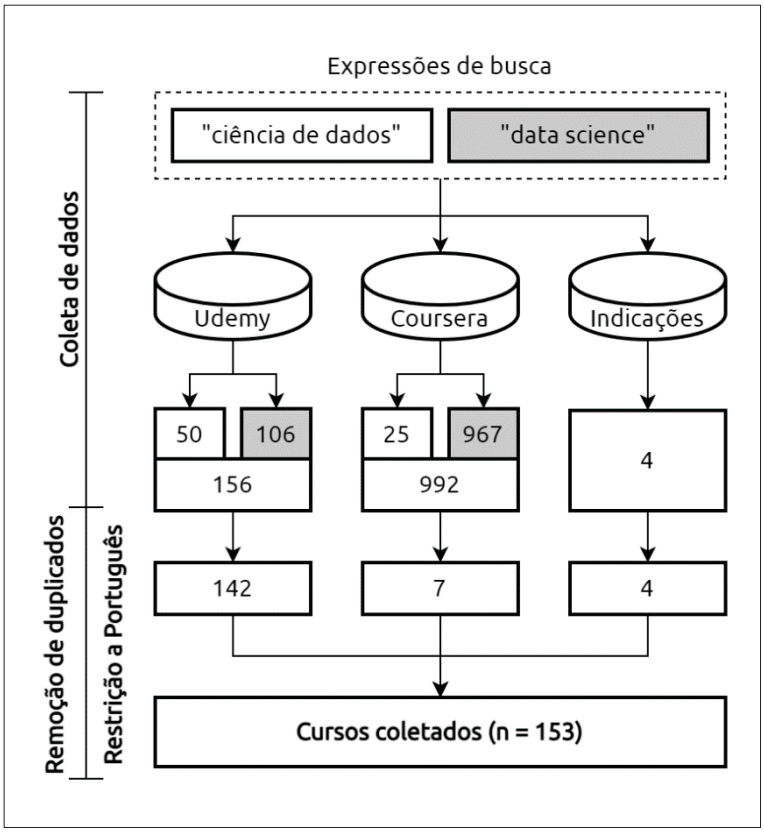
QUADRO 14 – CURSOS LIVRES DERIVADOS DE INDICAÇÕES DA PESQUISA DE LEVANTAMENTO

Instituição	**Curso**
Data Science Academy	Preparação Para Carreira de Cientista de Dados
	Bootcamp de Certificação Cientista de Dados
	Introdução à Ciência de Dados 3.0
IGTI	Bootcamp Cientista de Dados
Coursera / FIA	Ferramentas para Ciência de Dados: Introdução ao R
	Ciência de Dados para Finanças
	Marketing e Data Science
	Finanças Orientada a Dados
	Introdução à Ciência e Engenharia de Dados
	Análise de Dados para Workforce Management
	Marketing Science e Estratégia de Marketing

FONTE: O autor (2022).

Como foram apenas onze registros, os conteúdos dos cursos foram coletados manualmente. Por fim, a Figura 31 apresenta a visão geral da coleta dos 153 conteúdos de cursos livres analisados nesta pesquisa.

FIGURA 31 – ETAPAS DE COLETA DE DADOS REFERENTE AOS CURSOS LIVRES



FONTE: O autor (2022).

Novamente, os procedimentos de análise obedecem ao padrão estipulado para análise de anúncios e cursos superiores, tanto limpeza e transformação, quanto os procedimentos de análise. A próxima seção detalha todo o protocolo de análise da presente pesquisa.

5.5 PROTOCOLO DE ANÁLISE

Nesta seção, são descritos os procedimentos de análise, tanto para a abordagem quantitativa quanto para a abordagem qualitativa. Inicialmente, apresentam-se os métodos quantitativos destinados às variáveis numéricas e, em seguida, são detalhados os métodos qualitativos de análise que serão empregados nesta tese.

5.5.1 Métodos Quantitativos

Como principal ferramenta de análise quantitativa, foi adotada a Análise Fatorial Confirmatória (AFC), técnica muito utilizada em pesquisas aplicadas, em especialmente em Ciência Sociais, que pode confirmar o quão bem variáveis observáveis representa um número menor de construtos (HAIR *et al.*, 2014; SOUZA; ALEXANDRE; GUIRARDELLO, 2017). A adoção da AFC para validar estruturas de mensuração em pesquisas sobre competências é encontrada em pesquisas preliminares (KOENIGSFELD *et al.*, 2012; LIU *et al.*, 2016; WANG, 2013), demonstrando a utilidade desta técnica para este tipo de análise.

Brown (2015) explica que AFC é um tipo que modelagem de equação estrutural (MEE) voltada especificamente a modelos de mensuração, onde são verificadas as relações entre medidas observadas, também chamadas de indicadores, e as variáveis latentes, denominadas fatores. Para Hair et al. (2014), a principal vantagem da AFC é o pesquisador poder testar analiticamente uma teoria conceitualmente fundamentada. Desta forma, é possível explicar como itens mensurados descrevem medidas sociológicas, psicológicas ou de negócios. Para realização do procedimento, foram seguidos os passos descritos na Quadro 15:

QUADRO 15 – PROTOCOLO DE ANÁLISE QUANTITATIVA

(continua)

Estágio	**Objetivos**	**Procedimentos**	**Referências**
Pré-condições	Testar se o conjunto de dados é apropriado para a análise fatorial.	Estatísticas descritivas: informações básicas sobre os dados e sua distribuição.	Field, Miles and Field (2012); Hair et al. (2014); Korkmaz, Goksuluk e Zararsiz (2014); Souza, Alexandre e Guirardello (2017); Thode (2002)
		Teste de Shapiro-Francia: Verificar a normalidade univariada das variáveis.	
		Teste de Mardia: Verificar a normalidade multivariada das variáveis, baseado nas medidas de assimetria e curtose.	
		Alfa de Cronbach: medida diagnóstica que avalia a consistência interna do instrumento de pesquisa. Valor acima de 0,7 são aceitáveis.	
		Teste de Kaiser-Meyer-Olkin (KMO): medida de adequação da amostra para o procedimento de Análise Fatorial Exploratória (AFE). O valor da estatística KMO varia entre 0 e 1, sendo que valores acima de 0,9 são esperados. Embora esse procedimento não seja requerido para AFC, optou-se pela sua realização.	
		Teste de esfericidade de Bartlett (BS): teste estatístico da significância geral de todas as correlações em uma matriz de correlação.	

(conclusão)

Estimação	Definição do método de estimação das cargas fatoriais.	Mínimos quadrados ponderados diagonalmente (DWLS): método de estimação que utiliza matriz de correlação policórica, considerada superior à tradicional máxima verossimilhança (*maximum likelihood*, ML), especialmente quando há restrição de normalidade e/ou no tamanho da amostra.	Li (2016); Xia e Yang (2019)
Validade do modelo de mensuração	Verificação da significância e dos índices de ajuste do modelo.	χ^2: um indicador tradicional para avaliar a qualidade geral do modelo onde um bom ajuste deve apresentar um resultado insignificante, isto é, o p-valor deve ser maior que 0,05. Porém, por ser uma medida muito rigorosa, a divisão do χ^2 pelo número de graus de liberdade (gl) também é útil para indicar o ajuste do modelo. Neste caso, idealmente, essa relação deve ser inferior ou igual a 2 ou, no máximo, a 3.	Schreiber et al. (2006); Hair et al. (2014); Xia e Yang (2019);
		Erro quadrático médio de aproximação (RMSEA): um indicador de discrepância entre os dados e o modelo por grau de liberdade do modelo. Valores mais baixos são esperados e um índice entre 0,05 e 0,08 é aceitável.	
		Índice Goodness-of-fit (GFI): índice absoluto de ajustamento que indica a razão de variância que é contabilizada pela covariância de população estimada. O valor GFI varia entre 0 e 1, onde valores mais altos indicam melhor ajuste e valores acima de 0,95 indicam um ajuste muito bom.	
		Índice Tucker Lewis (TLI): índice incremental de adequação obtido pela comparação do valor de χ^2 do modelo com o valor de χ^2 do modelo nulo. É uma retificação do índice de ajuste normalizado (NFI) que é sensível ao tamanho da amostra. TLI também apresenta valor entre 0 e 1 onde valores mais altos indicam melhor ajuste.	
		Índice Bentler's Comparative Fit (CFI): outro índice incremental de ajuste que é uma versão aprimorada do NFI. CFI varia entre 0 e 1 e valores acima de 0,9 são esperados.	
		Raiz Padronizada do Resíduo Médio (SRMR): índice útil para comparar a qualidade dos modelos ajustados. São esperados valores abaixo de 0,08.	
Validade dos construtos	Avaliar se a estrutura dos construtos é válida.	Confiabilidade dos construtos/Confiabilidade Composta (*Composite Reliability*) (CR): indicador de confiabilidade e consistência interna das variáveis observadas para representação de um construto latente. Valores entre 0,70 e 0,90 são considerados satisfatórios.	Hair et al. (2014, 2019); Henseler, Ringle e Sarstedt (2015);
		Variância média extraída (AVE): medida sumária de convergência entre um conjunto de variáveis medidas que representam um construto latente. Espera-se encontrar valores superiores a 0,50.	
		Heterotrait-monotrait ratio (HTMT): medida utilizada como validade discriminante, i. e., avalia se um construto é realmente distinto dos outros construtos. Espera-se valores abaixo de 0,90.	

FONTE: O autor (2022).

O primeiro conjunto de procedimento do protocolo de análise se refere à verificação de adequação dos dados ao procedimento de análise fatorial. Todavia, além das estatísticas descritivas dos dados referentes ao modelo estatístico, isto é, dos grupos com as variáveis referentes às competências em Ciência de Dados, são apresentadas as principais características de todas as variáveis do questionário. Dessa forma, foi possível obter uma visão geral dos dados do questionário para, então, prosseguir com os demais passos (FIELD; MILES; FIELD, 2012-, p. 179). Assim, por meio de análise univariada, pode-se determinar o que é típico no grupo, indicar a variabilidade entre os indivíduos e verificar a distribuição dos elementos em relação às variáveis (GIL, 2008, p. 161). Com medidas de tendência central, como média aritmética e mediana, e medidas de dispersão, amplitude e desvio padrão, tem-se uma visão geral do comportamento das variáveis do conjunto coletado.

Em seguida, foi verificada a normalidade dos dados, tanto de forma univariada quanto multivariada. Uma vez que os dados qualitativos foram coletados inicialmente pelos grupos de cientistas de dados em redes sociais, em todo o território brasileiro, esperava-se um número de respostas que garantiria uma distribuição normal e, consequentemente, que os dados sejam classificados como paramétricos. Field, Miles e Field (2012-, p. 167) frisam que uma vez que inúmeros testes estatísticos são baseados no conceito de distribuição normal, utilizá-los para dados não-paramétricos pode gerar resultados impróprios. Porém, a normalidade dos dados não foi evidenciada, conforme detalhado na seção 6.1.2, e, portanto, não foi necessário verificar outras características necessárias à aplicação de testes paramétricos (FIELD; MILES; FIELD, 2012-, p. 168). Os procedimentos seguintes foram adotados mediante a observação de não-normalidade dos dados.

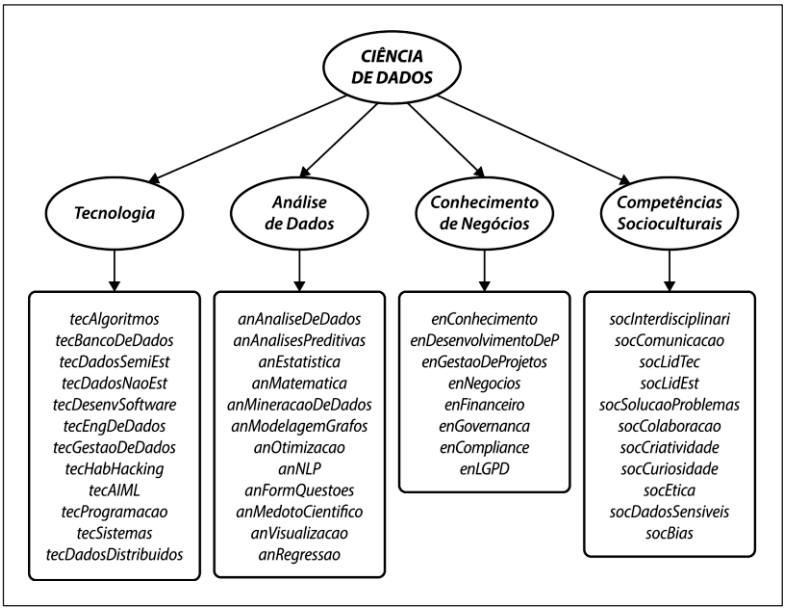
Ainda no estágio de verificação de pré-condições, verificou-se a confiabilidade (*reliability*) do instrumento de mensuração. Ou seja, no caso de questionário, é averiguar se uma variável ou um conjunto de variáveis é consistente com aquilo que se pretende mensurar (FIELD; MILES; FIELD, 2012-, p. 798; HAIR, 2009, p. 101). Dentre os índices que mensuram a confiabilidade, o alfa de Cronbach é um dos mais utilizados, sendo “universalmente aconselhável” (MARÔCO; GARCIA-MARQUES, 2006, p. 66). Sampieri, Collado e Lucio (2013, p. 226) também destacam a medida de consistência interna, alfa de Cronbach, como um dos coeficientes mais adotadas, destacando que seu coeficiente pode variar de 0 (zero), nenhuma confiabilidade, a 1

que indica o máximo de confiabilidade. Field, Miles e Field (2012, p. 17) indicam que valores entre 0,7 e 0,8 já são aceitáveis.

Além do alfa de Cronbach, foram empregados os testes de Kaiser-Meyer-Olkin (KMO) e de esfericidade de Bartlett (BS). O KMO é um teste indicado para verificar a adequação de uma amostra para o procedimento de análise fatorial exploratória, onde cujos valores variam entre 0 e 1, sendo que valores acima de 0,8 são considerados ótimos (FIELD; MILES; FIELD, 2012-, p. 920). Já o teste estatístico de esfericidade de Bartlett verifica a correlação entre as variáveis e deve apresentar significância estatística para atestar a adequabilidade dos dados a uma estrutura fatorial (HAIR *et al.*, 2014).

Assim, com esses passos, foram averiguadas as condições para confirmar a relação definida para o questionário. Isto é, foi possível verificar se o agrupamento das competências, baseado na literatura consultada tem validade estatística. Os grupos "Tecnologia", "Análise de Dados", "Conhecimento de Negócios" e "Competências Socioculturais" (Figura 32) são, neste cenário, constructos latentes, pois não podem ser medidos diretamente, sendo mensurados pode meio de uma ou mais variáveis (HAIR, 2009, p. 540). Dessa forma, a Análise Fatorial Confirmatória (*Confirmatory Factor Analysis* – AFC) foi utilizada para testar as hipóteses de relações determinadas entre variáveis latentes relacionadas à Ciência de Dados (FIELD; MILES; FIELD, 2012-, p. 915).

FIGURA 32 – MODELO CONCEITUAL DAS COMPETÊNCIAS DA CIÊNCIA DE DADOS



FONTE: O autor (2022).

Ressalta-se, ainda, que o modelo desenvolvimento é considerado de ordem superior ou, mais especificamente, de segunda ordem. Dessa forma, seguindo a definição de Hair et al. (2014), o modelo apresenta duas camadas de construtos latentes, sendo que o fator latente de segunda ordem, denominado Ciência de Dados, é a causa dos quatro fatores latentes de primeira ordem (Tecnologia, Análise de Dados, Entendimento de Negócios e Competências Socioculturais) que são a causa das variáveis mensuradas, por sua vez.

Os passos seguintes do protocolo de análise quantitativa, descritos no Quadro 15, referem-se à estimação das cargas fatoriais do modelo, à validação do modelo de mensuração e à validação dos construtos obtidos. Para a execução dos passos definidos neste protocolo, foi utilizado principalmente o software JASP (https://jasp-stats.org), que é uma iniciativa gratuita e de código aberto, que oferece uma interface gráfica para realização de procedimentos estatísticos. Ademais, procedimentos não oferecidos pelo JASP, como testes de distribuição e validação dos construtos, foram realizados por meio do Microsoft Excel e da linguagem R (https://www.r-project.org), utilizando o software RStudio. Por fim, para gráficos, também foi utilizado o Scimago Graphica (https://www.graphica.app/), uma ferramenta gratuita para visualização de dados.

5.5.2 Métodos Qualitativos

Nesta seção, são descritos os procedimentos de análise aos dados não numéricos coletados durante a pesquisa, assim organizados:

a) Campos de texto do questionário (perguntas 44 a 54, 59 e 61);
b) Conteúdo de texto dos anúncios de vagas de emprego;
c) Conteúdo programático de cursos de formação do cientista de dados, categorizados em bacharelado, tecnológico e cursos livres;

Essencialmente, as análises qualitativas da pesquisa se fundamentam na mineração de texto, que é um tipo de mineração de dados voltado à extração de conhecimento a partir de grandes volumes de texto não estruturado (SUMATHI; SIVANANDAM, 2006, p. 558). Gajzer (2010, p. 219) considera que a principal função

da mineração de texto é sua capacidade em converter documentos de texto, logo não estruturados, em uma organização formal que gera inúmeras possibilidades de análise. A mineração de texto, por meio de abordagens estatísticas, como técnicas de agrupamento e análise fatorial, revela relações ocultas, ou mesmo reforça relações já conhecidas, simplificando a representação do conteúdo semântico presente em grandes volumes de dados (WOLFRAM, 2017, p. 96).

As aplicações da mineração de texto envolvem a recuperação de dados e documentos, identificação de estrutura e conteúdo temático, descoberta da literatura, verificação do estado da arte de um tema ou mesmo serem utilizadas em análises preditivas (SUMATHI; SIVANANDAM, 2006, p. 559). Wolfram (2017, p. 98) afirma que aplicações que combinam mineração de texto, processamento de linguagem natural (*Natural Language Processing* – NLP), *machine learning* e modelagem de tópicos permitem que pesquisas utilizem um corpus maior, formado por textos integrais. Se por um lado essa combinação amplia a capacidade analítica, por outro, ela requer um conjunto de dados grande o suficiente para produzir resultados confiáveis.

A etapa de mineração de texto propriamente dita foi precedida pela apresentação das principais características dos documentos que compõem os *corpora* analisados. Para o estudo dos anúncios, por exemplo, buscou-se extrair informações acerca das vagas oferecidas a cientistas de dados no Brasil. Nesse caso, além da relevância de cada website utilizado, foram verificadas as localidades das vagas, os salários, o nível de escolaridade exigido, além do nível de experiência (Júnior, Pleno ou Sênior). O mesmo tipo de procedimento foi aplicado aos cursos superiores e livres.

Assim, com o perfil dos conteúdos conhecido, pode-se proceder à mineração de texto propriamente dita, dividida em três passos:

- **Limpeza e formatação dos dados:** ações específicas para a análise textual, conforme apresentado por Gajzer (2010, p. 223–224). Primeiramente, os documentos (anúncios) passam pelo processo de "tokenização", a separação das palavras (tokens). Desta forma, cada documento é convertido em uma lista de termos que serão contabilizados. Em seguida, todo conteúdo foi convertido para letras minúsculas, sendo caracteres especiais, números, URLs, dentre outros elementos textuais sem relevância para a análise. Assim, procedeu-se à remoção de *stopwords*, palavras muito frequentes nos idiomas como

artigos, preposições e conjunções, que possuem pouca carga semântica para a análise textual (Data Science for Business: What you need to know about data mining and data-analytic thinkingPROVOST; FAWCETT, 2013, p. 253). Por fim, para evitar variações como plural, foram adotados dois procedimentos de padronização de termos: a) Stemmização: conversão de cada token para um termo menor (radical); b) Lemmatização: processo que busca por meio de dicionário a palavra que deu origem ao termo em questão (BENGFORT; BILBRO; OJEDA, 2018-, p. 72). Por exemplo, os tokens “dados” e “dado” são convertidos para o *stem* “dad”, enquanto o *lemma* destes dois termos é o verbo “dar”. Os resultados mais coerentes foram encontrados com o uso de stems, por isso, as análises se concentram neste formato.

- **Procedimentos pré-análise:** com a remoção de elementos desnecessários e padronização dos termos restantes, é verificada a distribuição de frequência de palavras. Nesta etapa, foi computada a matriz BOW, que verifica a frequência de cada termo para todo o corpus, e o TD-IDF, que normaliza a frequência dos tokens em um documento em relação ao restante do corpus (BENGFORT; BILBRO; OJEDA, 2018-, p. 56–62). Estes procedimentos são preparação para as análises seguintes.
- **Procedimentos de análise:** para identificar padrões nos anúncios coletados, são adotados três tipos de análise: a) N-grama: apresentação dos termos (*stems*) mais frequentes, além das expressões formadas por dois e três stems (RASCHKA, 2016, p. 237); b) Modelagem de tópicos: para extrair os principais temas presentes nos anúncios foi adotada a técnica Latent Dirichlet Allocation (LDA), que pertence à família de modelos probabilísticos e emprega uma abordagem Bayesiana de duas camadas para identificar padrões de coocorrência de palavras, definindo tópicos do corpus (WESSLEN, 2018, p. 1). Para a LDA foi utilizada a biblioteca Gensim (ŘEHŮŘEK; SOJKA, 2011) com o auxílio da biblioteca de visualização pyLDAvis (MABEY, 2018) e; c) Agrupamento: por fim, implementação do algoritmo de agrupamento K-Means por meio da ferramenta Clustering Workbench (CARROT^2 CLUSTERING ENGINE, 2021). O algoritmo K-Means, que é bastante popular, começa

> com um numérico arbitrário de agrupamentos e posiciona as instâncias (documentos) conforme sua proximidade com os centroides dos grupos, sendo que o objetivo final é minimizar a soma dos quadros na estruturada encontrada (BENGFORT; BILBRO; OJEDA, 2018-, p. 103).

Segundo Wesslen (2018), a adoção de modelos de aprendizagem de máquina por pesquisadores das Ciências Sociais tem emergido como uma das principais técnicas para descoberta de variáveis latentes, que antes só poderiam ser medidas sob suposições não testáveis. Neste sentido, esta pesquisa explora esses recursos, algoritmos de *machine learning* para mineração de texto, para identificar os principais requisitos para ser contratado como um cientista de dados. Além disso, o emprego da mineração de texto busca possibilitar a extração de padrões e conhecimento de centenas de documentos cuja análise manual seria mais onerosa.

A linguagem de programação Python foi adotada na etapa de limpeza e formatação dos dados, nos procedimentos pré-análise e, também, nas próprias análises aplicadas. Para as nuvens de palavras, utilizou-se a ferramenta online WordClouds.com e para visualizar os termos no seu contexto original, empregou-se o software livre AntConc (https://www.laurenceanthony.net/software/antconc/).

5.6 SÍNTESE DO CAPÍTULO

Neste capítulo, além da classificação metodológica, foram detalhados os procedimentos relacionados à coleta, tratamento e análise dos dados da pesquisa. Esse conteúdo é assim sintetizado:

- A pesquisa é considerada como aplicada quanto à sua natureza, descritiva em relação aos objetivos e emprega uma abordagem mista, coordenando métodos quantitativos e qualitativos para estudar o problema (CRESWELL, 2010; MARTINS; THEÓPHILO, 2009; SAMPIERI; COLLADO; LUCIO, 2013; SILVA; MENEZES, 2005-).
- Para explorar as competências dos cientistas de dados no Brasil, optou-se por utilizar mais de uma unidade de análise (SAMPIERI; COLLADO; LUCIO, 2013). Assim, além da pesquisa de levantamento com os profissionais pesquisados, foi realizada uma pesquisa documental em

anúncios de vagas de emprego, conteúdos de cursos superiores e cursos livres voltados à Ciência de Dados.

- O instrumento de coleta dados, fundamentado em pesquisas anteriores (HARRIS; MURPHY; VAISMAN, 2013-; HAYES, 2020; KAGGLE, 2020) e reforçado pela revisão da literatura, foi definido com a ajuda de especialistas e acadêmicos que participaram da etapa de pré-teste. Em resumo, o questionário foi organizado em seis seções (Tecnologia, Análise, Entendimento de Negócios, Sociocultural, nível individual e organizacional, além de informações adicionais), sendo disponibilizado no endereço eletrônico http://cienciadedados.info. O período de coleta de dados foi de julho a novembro de 2021 e resultou em 227 respostas válidas.
- Para a coleta dos documentos analisados, a raspagem de dados foi utilizada para os anúncios de vagas de emprego, realizada em seis *websites* distintos, e para os cursos livres, mais especificamente na plataforma Udemy. Os conteúdos dos cursos livres indicados na pesquisa de levantamento foram coletados manualmente, assim como realizado para os cursos de bacharelado e tecnológicos. Dos cursos superiores, identificados no portal e-MEC, foram coletados os textos das páginas institucionais e as matrizes curriculares, quando disponíveis.
- O protocolo de análise quantitativa foi essencialmente direcionado para a aplicação da análise fatorial confirmatória (FIELD; MILES; FIELD, 2012-; HAIR *et al.*, 2014; KORKMAZ; GOKSULUK; ZARARSIZ, 2014). Assim, a análise descritiva dos dados é sucedida pela verificação de condições para realização da análise fatorial, incluindo normalidade dos dados, consistência do instrumento de coleta, além te testes como KMO e de esfericidade de Bartlett. Em seguida, com a definição do método de estimação das cargas fatorial, foi descrita a validação do modelo, de segunda ordem, por meio de indicadores específicos (χ^2, RMSEA, GFI, TLI, CFI, SRMR, CR, AVE e HTMT).
- A análise de dados qualitativos foi detalhada com base nas etapas de mineração de texto: a) limpeza e formatação dos dados; b) procedimentos pré-análise e; c) procedimentos de análise (BENGFORT; BILBRO; OJEDA, 2018-; GAJZLER, 2010; RASCHKA, 2016;

WESSLEN, 2018). Para identificar padrões nos conteúdos textuais da tese, foram utilizadas nuvens de palavras, análise N-grama, modelagem de tópicos (LDA) e agrupamento de documentos (K-Means).

- As ferramentas de análise usadas foram os softwares JASP, Microsoft Excel, R, RStudio e Scimago Graphica para os métodos quantitativos. Para os dados qualitativos, a linguagem Python foi utilizada para raspagem, tratamento e análise de dados. Para o agrupamento dos documentos, foi adotada a ferramenta Clustering Workbench, para nuvem de palavras, o site WordClouds.com e, para visualizar os termos em contexto, foi empregado o software AntConc.

Em suma, este capítulo é destinado a detalhar as escolhas metodológicas dessa pesquisa, classificada como aplicada, descritiva e de abordagem mista. Por isso, a descrição aqui apresentada, além de expor o caminho percorrido para os achados dessa pesquisa, fornece subsídios para futuras pesquisas relacionadas. Nesse sentido, é fundamental para posterior replicabilidade dos procedimentos adotados. Todavia, salienta-se que mais informações acerca dos dados coletados e analisados são apresentadas no próximo capítulo.

6 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Os resultados obtidos pelos procedimentos de análise são apresentados neste capítulo. Primeiramente, são explorados todos os aspectos da pesquisa de levantamento, do perfil dos respondentes ao modelo estatístico resultante da Análise Fatorial Confirmatória (AFC). Em seguida, são expostos os procedimentos de mineração de texto, iniciados pela análise de anúncios de vagas de emprego, seguidos pelo conteúdo educacional, cursos superiores e cursos livres.

6.1 PESQUISA DE LEVANTAMENTO COM PROFISSIONAIS

A pesquisa de levantamento realizada com profissionais da Ciência de Dados corresponde à parte central dessa tese. Além de verificar as competências fundamentadas na literatura, esta parte da pesquisa identifica o perfil dos cientistas de dados atuantes no Brasil. Além de incluir dados demográficos e educacionais, apresenta informações do próprio mercado, como segmentos e portes das organizações que contam com profissionais de dados em seus quadros de colaboradores.

6.1.1 Perfil dos respondentes

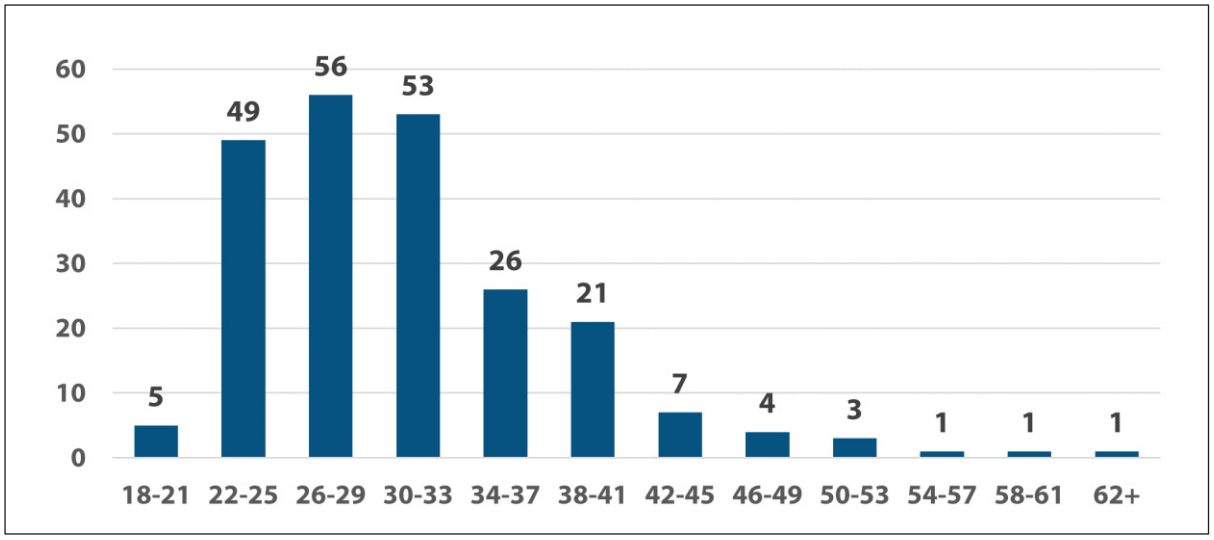
Nesta seção, são descritos os dados de caracterização dos respondentes. Inicialmente, os dados de gênero e idade são apresentados. Em seguida, o perfil educacional, contendo nível educacional e cursos de graduação mais recorrentes, é abordado. Por fim, são expostos os dados referentes à atuação profissional dos respondentes.

Dos 227 respondentes, 184 (81,05%) são homens e 41 (18,06%), mulheres. Um respondente preferiu não responder à pergunta, enquanto outro se idêntica como não-binário. Esse resultado que aponta a Ciência de Dados como um campo majoritariamente masculino vai ao encontro de pesquisas preliminares (BURTCH WORKS, 2019; HARRIS; MURPHY; VAISMAN, 2013-). Dentre essas pesquisas, a mais recente é o levantamento anual realizado pela comunidade Kaggle (2021) que, em sua última versão, obteve 25.974 respostas, sendo 82,2% de homens, 16,2% de mulheres e, consequente, 1,6% de pessoas que preferiram não responder ou não se

identificam com nenhum dos dois gêneros. Ao filtrar os dados de respondentes do Brasil, terceiro país com mais respostas na pesquisa com 751 registros, tem-se um aumento no predomínio masculino da área: 664 (88,41%) são homens, 81 (10,78%), mulheres e 6 respondentes preferem não responder ou são não-binários.

A idade média dos respondentes é de 30,99, com desvio padrão de 7,30, moda de 27 e mediana 30. O profissional mais jovem a responder a pesquisa tem 20 anos, enquanto o mais velho, 62 anos. O agrupamento das idades em intervalos de quatro anos, apresentado na Figura 33, revela que quase um quarto (24,67%) dos profissionais da pesquisa têm entre 26 e 29 anos, seguidos pelos profissionais de 30 a 33 anos (23,35%) e pelos de 22 a 25 anos (21,59%). Nota-se, ainda, que 81,05% dos profissionais têm entre 22 e 37 anos, confirmando que o segmento é dominado por profissionais jovens. Novamente, os valores encontrados para a idade dos profissionais estão em consonância com a pesquisa da comunidade Kaggle (2021).

FIGURA 33 – IDADE DOS RESPONDENTES



FONTE: O autor (2022).

No total, participaram da pesquisa profissionais de dados de 20 estados brasileiros e do Distrito Federal. Com 58 respondentes (25,55%), São Paulo foi o estado mais representativo, seguido pelo Paraná com 30 respostas (13,22%) e Rio de Janeiro, 27 respostas (11,89%). As outras unidades federativas com 10 ou mais respondentes são Minas Gerais (16 respostas), Ceará (15) e Distrito Federal (10). A Tabela 2 apresenta a lista com todos os estados e as principais cidades dos respondentes:

TABELA 2 – ESTADOS E CIDADES DOS RESPONDENTES

Estado	n	%	Cidade	n	%
São Paulo	58	25,55%	Curitiba	24	10,57%
Paraná	30	13,22%	São Paulo	23	10,13%
Rio de Janeiro	27	11,89%	Rio de Janeiro	22	9,69%
Minas Gerais	16	7,05%	Fortaleza	9	3,96%
Ceará	15	6,61%	Brasília	8	3,52%
Distrito Federal	10	4,41%	Belo Horizonte	8	3,52%
Paraíba	9	3,96%	Campinas	7	3,08%
Rio Grande do Norte	8	3,52%	Bauru	6	2,64%
Rio Grande do Sul	8	3,52%	Salvador	6	2,64%
Bahia	7	3,08%	Natal	5	2,20%
Espírito Santo	7	3,08%	Campina Grande	5	2,20%
Santa Catarina	6	2,64%	Belém	5	2,20%
Pará	6	2,64%	Recife	4	1,76%
Pernambuco	5	2,20%	Porto Alegre	4	1,76%
Goiás	4	1,76%	Florianópolis	4	1,76%
Maranhão	4	1,76%	João Pessoa	3	1,32%
Mato Grosso	2	0,88%	Vila Velha	3	1,32%
Alagoas	2	0,88%	São Luís	3	1,32%
Amazonas	1	0,44%	Ribeirão Preto	3	1,32%
Piauí	1	0,44%	Vitória	3	1,32%
Mato Grosso do Sul	1	0,44%	Outras	72	31,72%

FONTE: O autor (2022).

Embora São Paulo seja o estado com mais participantes, a cidade que mais contribuiu para a pesquisa foi a capital do Paraná, Curitiba, com 24 respostas válidas, ou 10,57% da amostra. Essa inversão se justifica pelo local de origem da pesquisa e a consequente influência da rede de contatos do autor. São Paulo, com 23 respostas (10,13%), e Rio de Janeiro, 22 respostas (9,69%), completam o trio de cidades com mais de 20 participações. Na Tabela 2, são exibidas todas as cidades com três ou mais respondentes, sendo que as demais totalizam 72 respostas ou 31,72% da amostra.

Em relação ao nível educacional, apenas um dos respondentes não possui educação formal de nível superior, embora 28 respostas (12,33%) fossem de estudantes com a graduação em andamento. Ou seja, 198 (87,22%) dos respondentes possuem graduação completa. Além disso, 70,04% dos profissionais da pesquisa concluíram ou estão em cursos de pós-graduação. O mestrado é a categoria de curso de pós-graduação com mais respostas, representando 29,07% dos respondentes, seguida pela especialização (28,19%) e, por fim, pelo doutorado

(12,78%), concluído por 17 respondentes (7,48%). A Tabela 3 traz todas as opções apresentadas aos respondentes, com os números de respostas de cada uma.

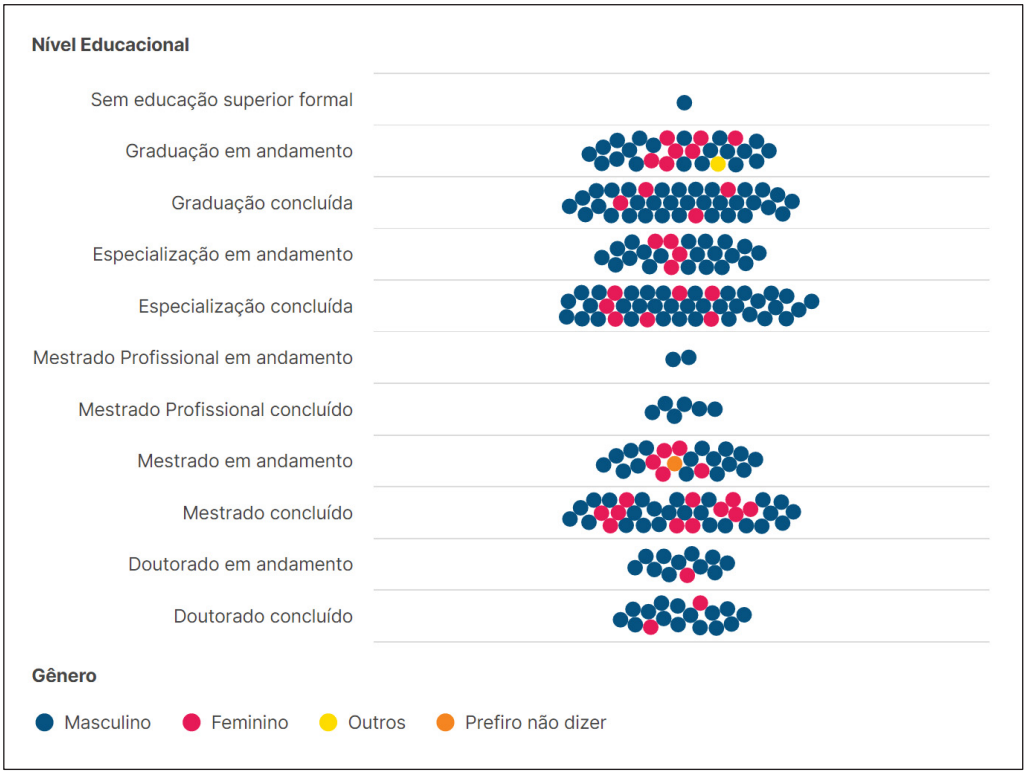
TABELA 3 – NÍVEL EDUCACIONAL DOS RESPONDENTES

Nível	**n**	**%**
Sem educação superior formal	1	0,44%
Graduação em andamento	28	12,33%
Graduação concluída	39	17,18%
Especialização em andamento	23	10,13%
Especialização concluída	41	18,06%
Mestrado em andamento	22	9,69%
Mestrado concluído	36	15,86%
Mestrado Profissional em andamento	2	0,88%
Mestrado Profissional concluído	6	2,64%
Doutorado em andamento	12	5,29%
Doutorado concluído	17	7,49%
Prefiro não responder	0	0,00%
Outros	0	0,00%
TOTAL	**227**	**100,00%**

FONTE: O AUTOR (2022).

A seguir, a Figura 34, apresenta uma visualização da distribuição dos respondentes pelos níveis educacionais apresentados, mostrando também a distinção de gêneros. O nível denominado “Especialização concluída”, referente àqueles profissionais que já concluíram ao menos uma pós-graduação *lato sensu*, é o maior grupo, com 42 respondentes. Por outro lado, o mestrado profissional é uma modalidade com menor adesão, apontada apenas por oito respondentes do gênero masculino, onde seis já concluíram o curso. Nas demais categorias, não houve diferença significativa entre os gêneros, pois a proporção se manteve próxima à representatividade da amostra.

FIGURA 34 – DISTRIBUIÇÃO DOS RESPONDENTES EM NÍVEL EDUCACIONAL E GÊNERO



FONTE: O autor (2022).

Das 227 respostas válidas, 202 informaram qual o curso de graduação que haviam concluído ou estavam cursando. Estatística foi a graduação mais recorrente, com 36 respostas (15,86% do total da amostra), seguido pelo curso de Ciência da Computação, com 30 registros (13,22%). Os cursos Sistemas de Informação (18 respostas, 7,93%), Análise e Desenvolvimento de Sistemas (4,85%), Engenharia da Computação e Economia (ambos com 10 respostas, 4,41%), Gestão da Informação e Engenharia Elétrica (ambos com 9 respostas, 3,96%) e Física (8 respostas, 3,52%) foram os que atingiram ao menos 3,00% da amostra. Graduação em Ciência de Dados foi indicada por seis respondentes (2,64%), dos quais dois ainda estão cursando a faculdade. Para mencionar a interdisciplinaridade da área, ressalta-se a presença de cursos como Música, Enfermagem, Geofísica, Oceanografia, Astronomia, Direito, Letras e outras engenharias, como aeronáutica, petróleo e madeireira. Dentre as instituições de educação superior mencionadas, destacaram-se a Universidade Federal do Paraná com 19 ocorrências (8,37% da amostra), Universidade de São Paulo, 11 respostas (4,85%), Universidade Federal do Ceará, seis respostas (2,65%), e Universidade de Brasília e Universidade Federal do Rio de Janeiro, com cinco respostas cada (2,20%).

Bem como ocorrido para os cursos de graduação, na pós-graduação, também há o predomínio da Computação e da Estatística. Computação foi um termo presente em 21 respostas referentes a mestrado: Ciência da Computação (11), Computação Aplicada (4), Engenharia da Computação (2), Computação (2), Computação Inteligente (1) e Engenharia Eletrônica e Computação (1). Além disso, um respondente indicou doutorado em Ciência da Computação. O curso de Estatística foi indicado em nove mestrados e dois doutorados. Dentre as instituições mais citadas, USP (14), UFPR (7), UFRJ (6), PUC-Minas (6), UFMG (5), UNICAMP (5) e UFG (5) são aquelas que foram indicadas por cinco ou mais respondentes.

Em relação às formações na área de Ciência de Dados, a instituição de ensino mais indicada pelos respondentes foi a Data Science Academy (DSA), com 13 respostas. Conforme indica o próprio nome, a DSA oferece cursos e formações para a Ciência de Dados e áreas afins, como análise de dados, engenharia de dados, inteligência artificial, dentre outros. Em segundo lugar, com nove indicações, está a plataforma Coursera que oferece cursos online para outras áreas além da Ciência de Dados, como Negócios, Línguas, Ciência da Computação, Saúde, Desenvolvimento Pessoal, Artes e Humanidades, Matemática e Lógica. Contudo, o material da Coursera é majoritariamente produzido no idioma inglês e legendado para outras línguas. Dentre as outras instituições indicadas por mais de dois respondentes estão o IGTI, com seis respostas, Udacity e Udemy, com cinco respostas, e PUC-MG e DataCamp, com três e duas indicações, respectivamente.

Quanto às informações profissionais, o tempo de experiência é outra evidência de que a Ciência de Dados é um campo em desenvolvimento, bem como é ocupado por profissionais jovens. A média dessa variável é de 5,87 anos, com desvio padrão de 4,98, moda de 2 e mediana 5. Na amostra, 83 respondentes (37,05%) afirmam ter até três anos de experiência, enquanto 73 dizem ter de quatro a seis anos de experiência (32,59%). Logo, sete em cada 10 respondentes possuem até seis anos de atuação na área. Na Tabela 4, pode-se verificar a distribuição das respostas fornecidas pelos profissionais em relação ao tempo de experiência na área.

TABELA 4 – TEMPO DE EXPERIÊNCIA

Anos	n	%	% Acumulada
1-3	83	37,05%	37,05%
4-6	73	32,59%	69,64%
7-9	25	11,16%	80,80%
10-12	23	10,27%	91,07%
13-15	10	4,46%	95,54%
16-18	3	1,34%	96,88%
19-21	3	1,34%	98,21%
22-24	0	0,00%	98,21%
25-27	2	0,89%	99,11%
28-30	1	0,45%	99,55%
31-33	1	0,45%	100,00%

FONTE: O autor (2022).

Na questão sobre remuneração, não houve predominância de uma faixa salarial específica, como pode ser visto na Tabela 5. Com exceção da primeira faixa apresentada, referente a quem ganha até dois salários-mínimos brasileiros[8], que recebeu 16 respostas (7,05%), as demais obtiveram entre aproximadamente 15 e 19% de respostas. A menor proporção obtida (14 respostas, 6,17%) diz respeito aos respondentes que não quiseram informar a remuneração, além de quatro (1,76%) que não possuem remuneração no momento.

TABELA 5 – REMUNERAÇÃO DOS RESPONDENTES

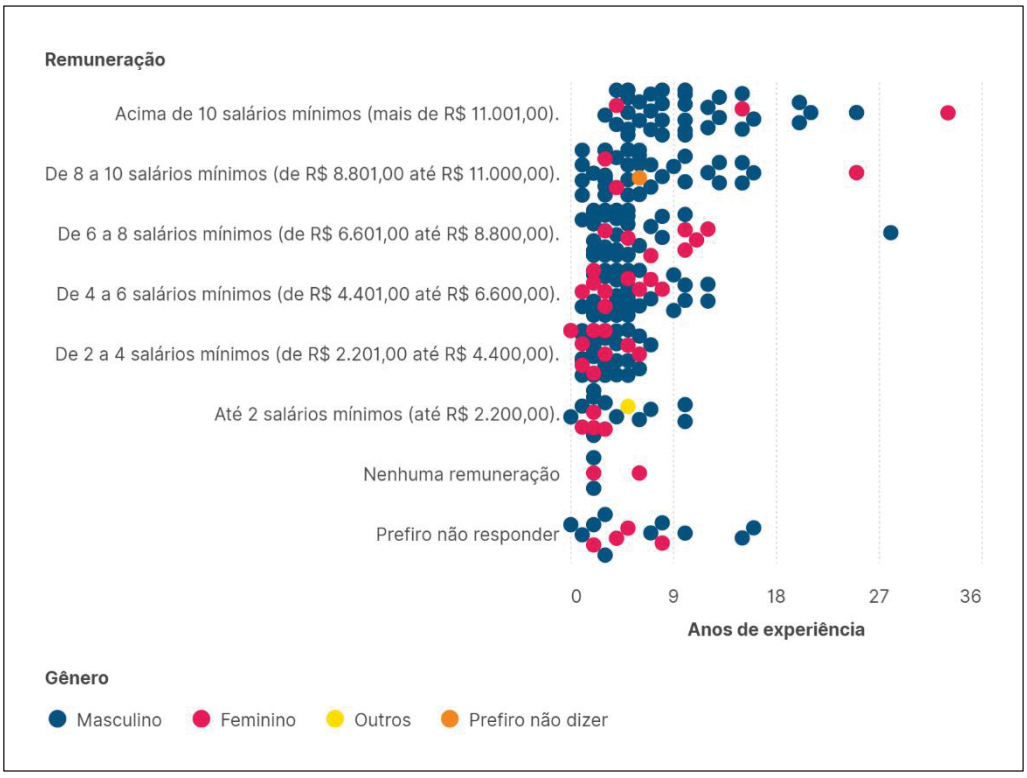
Faixa salarial	n	%	% Acumulada
Até 2 salários-mínimos (até R$ 2.200,00).	16	7,05%	7,05%
De 2 a 4 salários-mínimos (de R$ 2.201,00 até R$ 4.400,00).	38	16,74%	23,79%
De 4 a 6 salários-mínimos (de R$ 4.401,00 até R$ 6.600,00).	44	19,38%	43,17%
De 6 a 8 salários-mínimos (de R$ 6.601,00 até R$ 8.800,00).	35	15,42%	58,59%
De 8 a 10 salários-mínimos (de R$ 8.801,00 até R$ 11.000,00).	39	17,18%	75,77%
Acima de 10 salários-mínimos (mais de R$ 11.001,00).	37	16,30%	92,07%
Prefiro não responder	14	6,17%	98,24%
Nenhuma remuneração	4	1,76%	100,00%

FONTE: O autor (2022).

[8] O valor salário-mínimo mensal brasileiro, durante o ano de 2021, era R$ 1.100,00, equivalente a U$ 194.94, segundo cotação do Banco Central do Brasil realizada em 03 de dezembro de 2021, na qual o dólar americano valia R$ 5,642 (https://www.bcb.gov.br/estabilidadefinanceira/historicocotacoes).

Para verificar a relação entre remuneração e tempo de experiência, além da questão de gênero, foi desenvolvida a Figura 35, onde são observadas algumas questões. Inicialmente, é visível a correlação, ainda que não verificada estatisticamente, entre o tempo de experiência e a remuneração. Nota-se, por exemplo, que dentre os profissionais mais bem remunerados estão aqueles com mais tempo de profissão. Além disso, os respondentes que estão sem remuneração estão no intervalo de profissionais com menos tempo de experiência.

FIGURA 35 – RELAÇÃO ENTRE REMUNERAÇÃO, TEMPO DE EXPERIÊNCIA E GÊNERO



FONTE: O autor (2022).

Por outro lado, percebe-se que enquanto nas faixas menores de remuneração, onde o profissional recebe até oito salários-mínimos, o percentual de mulheres varia entre 20 e 25,00%, nos níveis mais altos, esta relação cai para cerca de 8,00%. Ou seja, para os profissionais que ganham acima de oito salários, há menos mulheres em termos proporcionais. Ainda assim, a resposta que indica o maior tempo de experiência, 33 anos, foi dada por uma respondente do gênero feminino que se encontra na faixa de remuneração mais alta apresentada.

Em relação à função ocupada pelos profissionais, os cientistas de dados representam mais da metade da amostra com 124 respostas (54,63%). Em segundo

lugar, estão os analistas de dados (26 respondentes, 11,45%), seguidos pelos engenheiros de dados (13 respondentes, 5,73%) e analistas de *Business Intelligence* (12 respondentes, 5,29%). Outros 25 profissionais indicaram funções que não estavam entre as opções apresentadas, sendo: desenvolvedor (3 respostas), engenheiro de *machine learning* (2 respostas) e professor (2 respostas) as únicas respostas recorrentes apresentadas.

O último item relacionado a práticas profissionais verifica a adoção de alguma metodologia específica pelos respondentes. Neste caso, a resposta mais recorrente, com 55 ocorrências (24,23%), foi o emprego de metodologias próprias das organizações onde os respondentes atuam. Em seguida, a CRISP-DM, com 54 respostas (23,79%), foi a metodologia de mercado mais citada, seguida por metodologias voltadas ao segmento de atuação (46 respostas, 20,26%). KDD e SEMMA foram as alternativas menos citadas, com oito e uma resposta, respectivamente.

Por fim, para concluir a seção com o perfil dos respondentes, é abordado o porte das organizações nas quais os profissionais atuam. Quase metade dos respondentes (113 respostas, 49,78%) atuam em empresas com mais de 249 colaboradores, a maior opção apresentada no questionário. Em segundo lugar, estão as organizações com o número de colaboradores de 100 a 249 (24 respostas, 10,57%), seguidas pelas empresas de 50 a 99 pessoas (16 respostas, 7,05%). Cinco respondentes (2,20%) trabalham como autônomos, enquanto 13 (5,73%) estão desempregados ou não estão vinculados a nenhuma organização.

Por outro lado, em relação ao tamanho das equipes de profissionais de dados, 81,52% das organizações contam com até 15 profissionais voltados diretamente à Ciência de Dados, sendo que em 55,43%, este número não supera cinco profissionais. O maior contingente deste tipo de profissionais indicado corresponde ao intervalo de 46 a 50 profissionais, apontado por quatro respondentes (2,17%). Assim, verifica-se que as equipes de cientista de dados e cargos análogos não é proporcional ao porte da organização.

Ainda em relação às empresas, o setor mais citado entre os respondentes foi o de Tecnologia, com 28 respostas, ou 12,33% da amostra. Em seguida, têm-se os segmentos de Consultoria (19 respostas, 8,37%), Financeiro (18 respostas, 7,93%), Energia (14 respostas, 6,17%), Público (13 respostas, 5,73%), Pesquisa e Desenvolvimento (11 respostas, 4,85%) e Educação (10 respostas, 4,41%),

compondo o grupo de segmentos com mais de 10 ocorrências. Dentre outras respostas obtidas, estão, Varejo, Marketing, Saúde, Comunicação, Relacionamento com Cliente, Comércio Eletrônico, dentre outros.

Apresentado o perfil dos respondentes, a próxima seção relata as competências dos profissionais da amostra.

6.1.2 Competências da Ciência de Dados

Antes do procedimento de Análise Fatorial Confirmatória (AFC) propriamente dito, é necessário apresentar as principais características dos dados analisados no modelo. Na Tabela 6, é apresentado, resumidamente, o comportamento de distribuição de cada variável, conforme definido no protocolo de análise.

TABELA 6 – ESTATÍSTICAS DESCRITIVAS

		Média	Desv. Padrão	Assimetria	Curtose
TECNOLOGIA	tecAlgoritmos	5,988	2,088	-0,296	-0,298
	tecBancoDeDados	5,440	2,198	-0,138	-0,520
	tecDadosSemiEst	5,930	2,545	-0,516	-0,465
	tecDadosNaoEst	5,238	2,535	-0,064	-0,774
	tecDesenvSoftware	5,608	2,720	-0,375	-0,681
	tecEngDeDados	5,191	2,554	-0,148	-0,757
	tecGestaoDeDados	7,080	2,007	-0,693	0,347
	tecHabHacking	6,310	2,445	-0,745	0,089
	tecAIML	6,403	2,624	-0,534	-0,678
	tecProgramacao	5,539	2,645	-0,315	-0,804
	tecSistemas	4,701	2,642	0,006	-0,706
	tecDadosDistribuidos	4,087	2,609	0,148	-1,007
ANÁLISE DE DADOS	anAnaliseDeDados	7,810	1,832	-0,944	1,245
	anAnalisesPreditivas	6,501	2,498	-0,695	-0,252
	anEstatistica	6,145	2,698	-0,315	-0,696
	anMatematica	6,546	2,336	-0,428	-0,474
	anMineracaoDeDados	6,729	2,470	-0,548	-0,369
	anModelagemGrafos	4,600	2,381	0,114	-0,685
	anOtimizacao	4,658	2,576	0,112	-0,643
	anNLP	5,048	2,590	-0,014	-0,763
	anFormQuestoes	6,507	2,489	-0,774	0,207
	anMedotoCientifico	6,449	2,635	-0,482	-0,507
	anVisualizacao	7,360	1,840	-0,706	0,610
	anRegressao	7,172	2,558	-0,845	0,054
ENTENDIMENTO DE NEGÓCIO	enConhecimento	6,877	2,239	-0,818	0,471
	enDesenvolvimentoDeP	6,019	2,398	-0,434	-0,309
	enGestaoDeProjetos	5,835	2,445	-0,373	-0,416
	enNegocios	5,386	2,625	-0,001	-0,819
	enFinanceiro	4,385	2,671	0,280	-0,698
	enGovernanca	4,809	2,566	0,027	-0,716
	enCompliance	4,688	2,577	-0,001	-0,664
	enLGPD	5,000	2,556	-0,036	-0,667
COMPETÊNCIAS SOCIOCULTURAIS	socInterdisciplinari	7,184	1,933	-0,830	0,989
	socComunicacao	7,249	1,932	-0,885	1,025
	socLidTec	6,385	2,353	-0,572	-0,166
	socLidEst	6,017	2,342	-0,387	-0,132
	socSolucaoProblemas	6,670	2,050	-0,576	0,130
	socColaboracao	8,148	1,909	-1,374	2,680
	socCriatividade	8,209	1,785	-1,341	2,402
	socCuriosidade	8,510	1,767	-1,826	4,392
	socEtica	8,893	1,892	-2,244	5,395
	socDadosSensiveis	8,078	2,116	-1,247	1,003
	socBias	8,056	2,131	-1,257	1,323

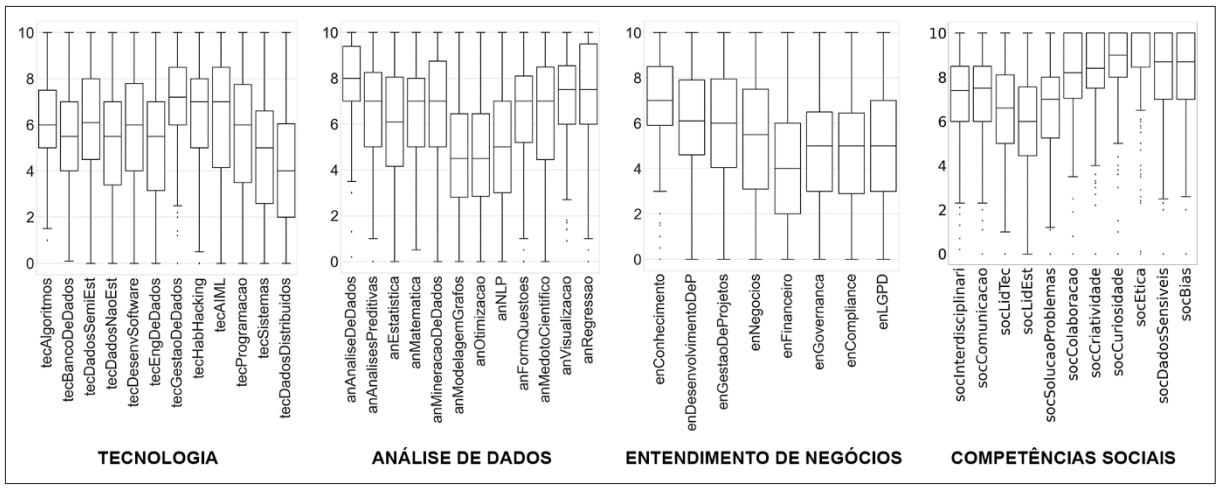
FONTE: O autor (2022).

Além da média e desvio padrão, os valores de assimetria (*skew*) e curtose (*kurtosis*) também são indicadores acerca da distribuição das variáveis. Em distribuições normais, os valores de assimetria e curtose são próximos a zero (FIELD; MILES; FIELD, 2012-), o que não ocorre nos dados coletados. A normalidade univariada das variáveis também foi verificada por meio da aplicação do teste Shapiro-Francia, indicado para amostras com mais de 50 observações (THODE, 2002). Esta

etapa, cujos resultados são apresentados no Apêndice B, confirmou que todas as variáveis da amostra possuem distribuição não-normal. Adicionalmente, a ausência de normalidade multivariada também foi comprovada por meio do teste Mardia, uma vez que os valores de assimetria e curtose do teste apresentam p-valor inferior a 0,05 (assimetria = 21.500,445, p-valor < 0,000 e curtose = 33,269, p-valor < 0,000).

A fim de facilitar a comparação dos resultados das variáveis por meio de análise visual, foram criadas algumas figuras. Inicialmente, na Figura 36, são apresentadas as distribuições no formato de gráfico de caixa onde é possível identificar a presença de *outliers*.

FIGURA 36 – GRÁFICOS DE CAIXAS (BOXPLOT) REFERENTES À DISTRIBUIÇÃO DAS VARIÁVEIS.



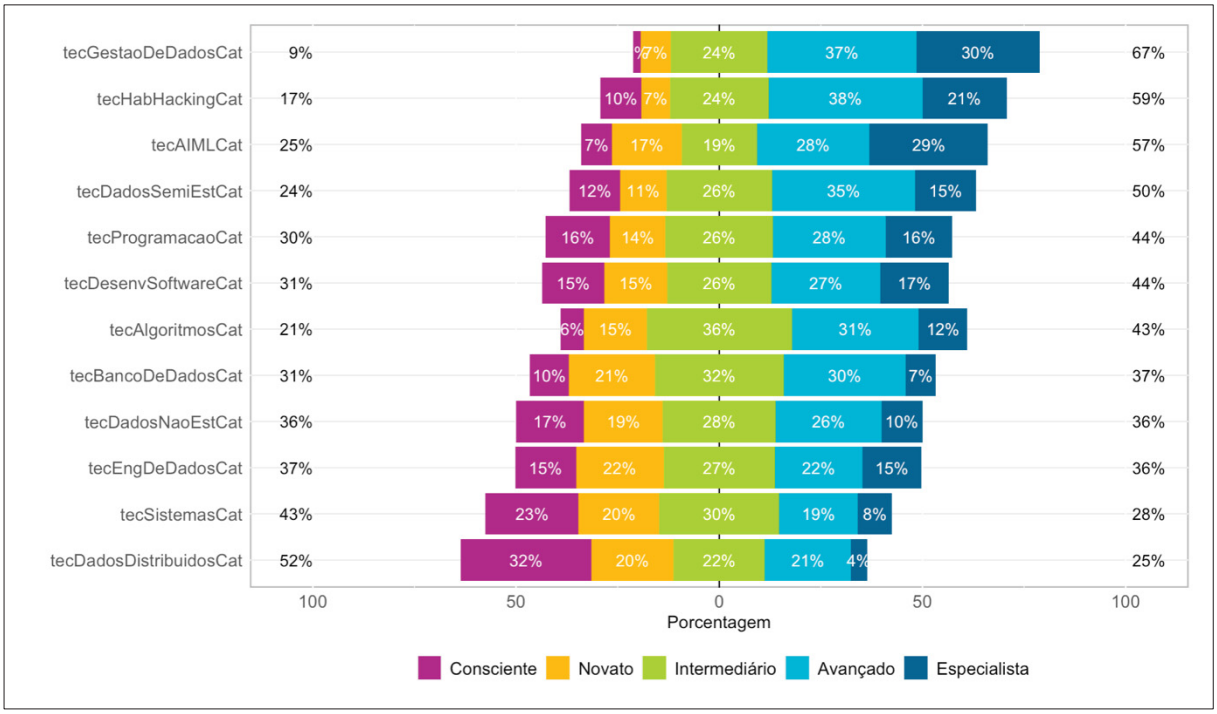
FONTE: O autor (2022).

Visualmente, é possível identificar características particulares entre os grupos de variáveis. Inicialmente, o grupo de Competências Socioculturais é aquele com os valores maiores, especialmente para as variáveis referentes à organização (*socColaboracao*, *socCriatividade*, *socCuriosidade*, *socEtica*, *socDadosSensiveis* e *socBias*). Essa constatação é confirmada pela média geral das variáveis desse grupo: 7,582. O grupo de Análise de Dados apresentou a segunda maior média geral (6,294) e concentrou as maiores médias, excluindo-se as Competências Socioculturais. Por outro lado, o grupo Entendimento de Negócios é aquele que graficamente demonstra os menores valores e menor média geral (5,375). Além disso, quatro das oito variáveis que compõe o grupo apresentam média igual ou menor a cinco (*enLGPD*, *enGovernanca*, *enCompliance* e *enFinanceiro*).

Neste sentido, o grupo de Tecnologia, cuja média geral foi de 5,626, contém a variável com a menor média: *tecDadosDistribuidos*, com valor de 4,087. Além disso, essa questão sobre sistemas de dados distribuídos de alto desempenho foi a mais "desconhecida" pelos respondentes, sendo que 18 profissionais indicaram não terem nenhum conhecimento sobre a questão. Em seguida, vem a questão referente à competência de planejamento financeiro para projetos em Ciência de Dados (*enFinanceiro*), com média de 4,385, do grupo Entendimento de Negócios, onde 15 profissionais indicaram desconhecimento. Outras três variáveis também foram indicadas como desconhecidas por mais de 10 respondentes: administração de sistemas de informação (*tecSistemas*), do grupo de Tecnologia, e otimização (*anOtimizacao*), do grupo de Análise de Dados, ambas com 13 indicações e; competência em garantir a conformidade com leis, normativas e demais regulamentações (*enCompliance*), também do grupo de negócios, com 11 indicações.

Em complemento à análise visual dos níveis de proficiência em relação aos grupos de competência, as figuras a seguir trazem o resultado de uma categorização das respostas. Ou seja, aquele indivíduo que respondeu entre 0 e 2 para um determinado item, é considerado como "Consciente" para o tópico em questão. Consequentemente, uma resposta entre 2 e 4, coloca-o como "Novato(a)", assim sucessivamente, conforme formulação do questionário. Na Figura 37, apresenta-se o grupo de Tecnologia e ordena as competências com mais profissionais de nível avançado e especialista. Ou seja, a primeira competência, Gestão de Dados (*tecGestaoDeDados*), possui 67% dos respondentes nos níveis avançado e especialista. Além disso, apenas 9% dos respondentes são dos níveis consciente ou novato. Por outro lado, para a competência em sistemas de dados distribuídos (*tecDadosDistribuidos*), mais da metade (52%) está nos níveis mais baixos, enquanto apenas 25% se consideram avançados ou especialistas.

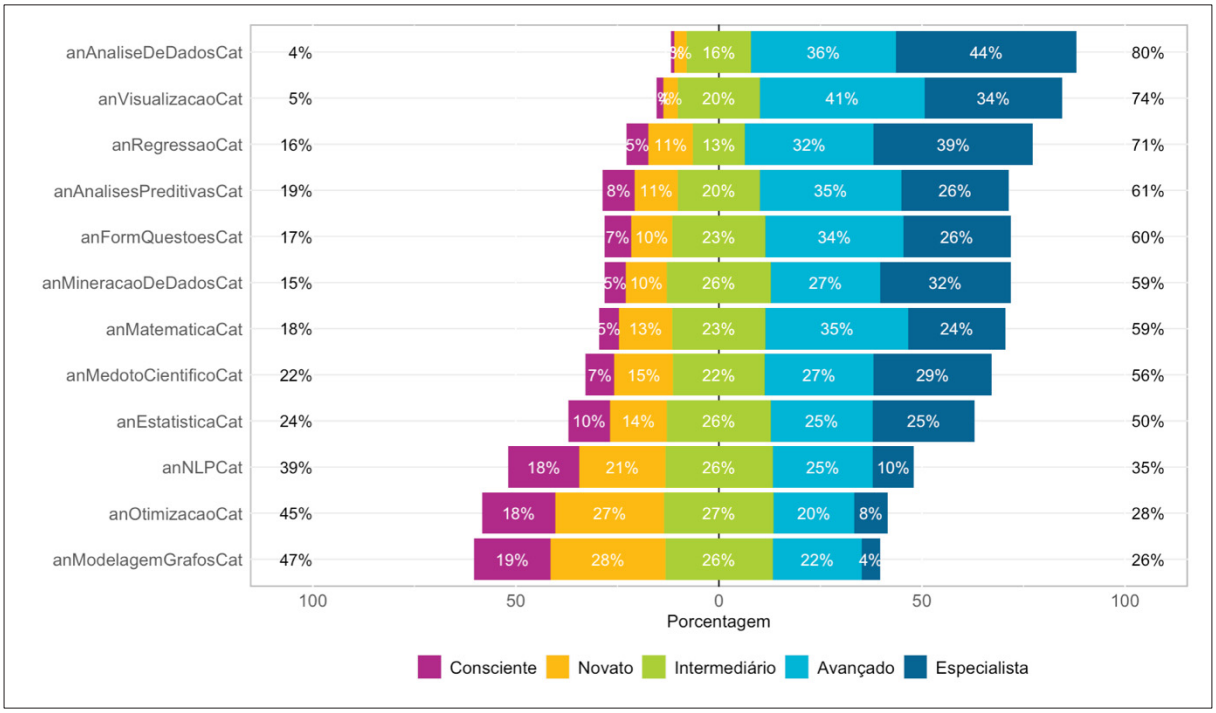
FIGURA 37 – PROFICIÊNCIA DAS COMPETÊNCIAS EM TECNOLOGIA



FONTE: O autor (2022).

Na Figura 38, nota-se que as competências do grupo de Análise de Dados são mais desenvolvidas em relação ao grupo de Tecnologia. Para a variável referente à competência genérica de Análise de Dados (*anAnaliseDeDados*), por exemplo, 80% dos respondentes se consideram nos níveis avançado ou especialista. Este percentual é igual ou maior a 50% para outras oito variáveis do grupo, indicando uma afinidade maior dos respondentes com a análise de dados do que com os tópicos em tecnologia.

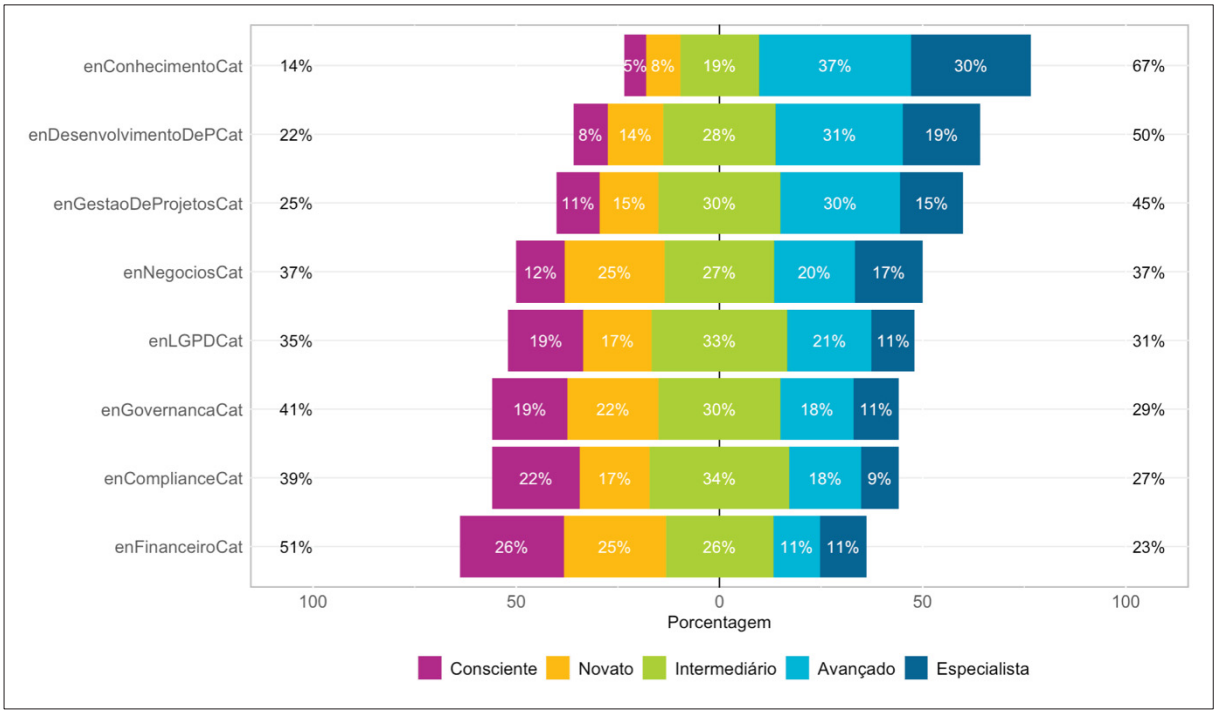
FIGURA 38 – PROFICIÊNCIA DAS COMPETÊNCIAS EM ANÁLISE DE DADOS



FONTE: O autor (2022).

Em compensação, a Figura 39, que representa as competências referentes ao Entendimento de Negócios, demonstra que apenas em dois tópicos a metade ou mais dos profissionais se considera avançado ou especialista: no conhecimento do domínio em que atuam, onde 80% se enquadram nestes níveis, e no projeto e desenvolvimento de novos produtos de dados, com 50% dos respondentes. Em outras quatro variáveis do grupo, o percentual dos níveis mais baixos de proficiência supera os níveis mais altos.

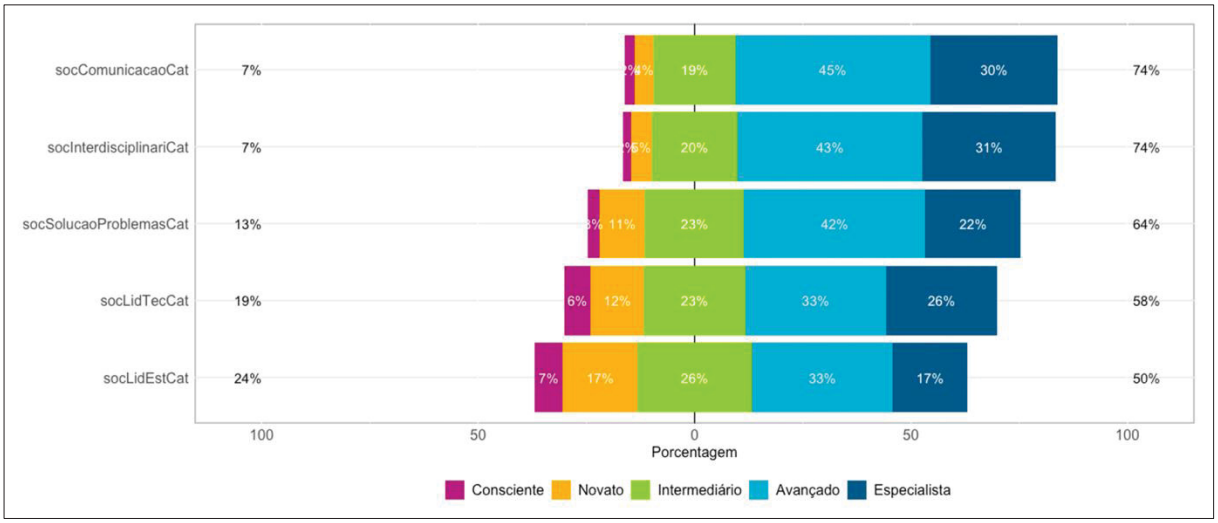
FIGURA 39 – PROFICIÊNCIA DAS COMPETÊNCIAS EM ENTENDIMENTO DE NEGÓCIOS



FONTE: O autor (2022).

Por fim, o grupo das Competências Socioculturais é apresentado em duas figuras (Figura 40 e Figura 41) que correspondem, respectivamente, aos níveis de competência individual e organizacional. Para os indivíduos, as competências socioculturais apresentaram os melhores resultados, onde, para todas as variáveis, os níveis avançados e especialista foram indicados por pelo menos metade dos respondentes. A comunicação e o conhecimento interdisciplinar foram as respostas com melhor desempenho, visto que aproximadamente ¾ dos profissionais se consideram em nível avançado ou especialista. Em seguida, está a competência em solucionar problemas relacionados a dados, com 64,00% de profissionais avançados ou especialistas, seguida pela liderança técnica (58,00%) e pela liderança estratégica (50,00%).

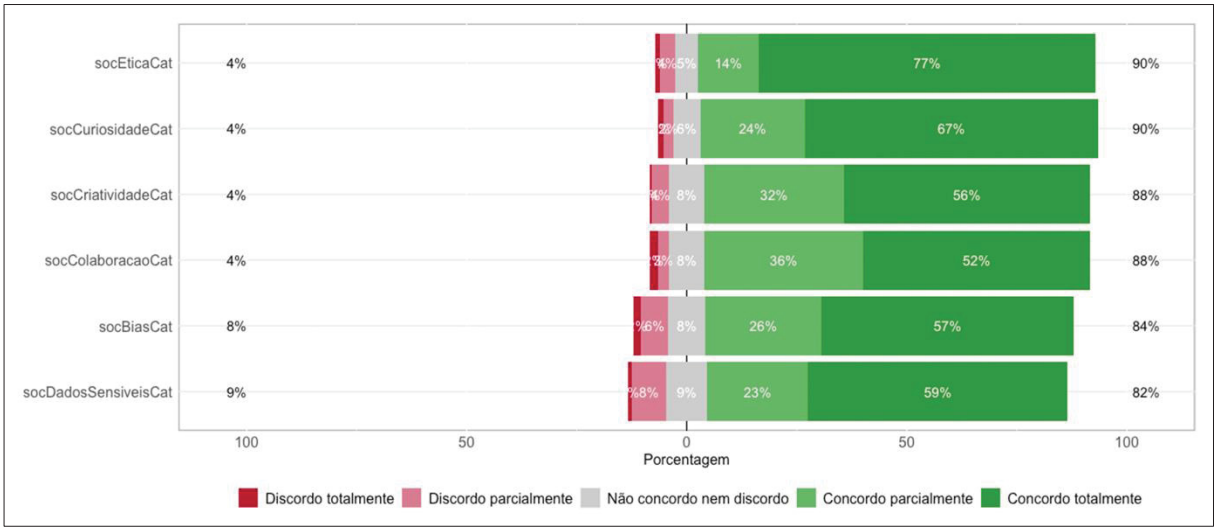
FIGURA 40 – PROFICIÊNCIA DAS COMPETÊNCIAS SOCIOCULTURAIS INDIVIDUAIS



FONTE: O autor (2022).

Novamente, as competências avaliadas no nível organizacional apresentaram resultados muito positivos em relação às demais questões. Aqui, no entanto, as questões não mediram o grau de proficiência dos profissionais, mas sim seu nível de concordância (valor número de 0 a 10) em relação ao seu ambiente organizacional. Dessa forma, para utilizar o mesmo tipo de gráfico, foi necessário agrupar esses valores numéricos em cinco níveis: “Discordo totalmente”, “Discordo parcialmente”, “Não concordo nem discordo”, “Concordo parcialmente” e “Concordo totalmente”. Dessa forma, todos os quesitos apresentados aos respondentes (ética, curiosidade, criatividade, colaboração, preocupação com análises tendenciosas e cuidado com dados sensíveis) apresentaram nível de concordância acima de 82,00%, conforme apresentado na Figura 41.

FIGURA 41 – CONCORDÂNCIA COM AS COMPETÊNCIAS SOCIOCULTURAIS ORGANIZACIONAIS



FONTE: O autor (2022).

Em relação à presença de *outliers* e à ausência de normalidade na distribuição das variáveis, seja uni ou multivariada, é importante frisar que a escolha pelo método de estimação Mínimos Quadrados Ponderados Diagonalmente (DWLS) permite o prosseguimento do protocolo de análise definido. Derivado do método Mínimos Quadrados Ponderados (WLS), estimador destinado a variáveis contínuas em distribuição não-normal, o DWLS utiliza correlação policórica, tratando os dados como ordinais, e também é pouco sensível à não-normalidade dos dados (LI, 2016).

Para concluir a verificação das pré-condições da análise fatorial, foi verificado o Alfa de Cronbach para atestar a confiabilidade da escala utilizada e o resultado foi 0,960, acima do valor aceitável (FIELD; MILES; FIELD, 2012-). Em seguida, para averiguar a adequação da amostra para o procedimento de análise fatorial, foi verificada a estatística Kaiser Meyer Olkin (KMO), cujo valor médio obtido foi 0,935, e realizado o teste de esfericidade de Bartlett cuja significância estatística foi menor que 0,05 (χ^2 = 8093,359, graus de liberdade = 903, p-valor < 0,001). Desta forma, mesmo que estes procedimentos sejam indicados para análises do tipo exploratória, os três indicadores atestam que os dados coletados são adequados à análise fatorial.

Assim, o primeiro modelo testado para a AFC seguiu a mesma estrutura demonstrada na Figura 32, com os construtos e suas respectivas variáveis identificadas na literatura. As cargas fatoriais, bem como o nível de significância de cada indicador, são apresentadas na Tabela 7.

TABELA 7 – CARGAS FATORAIS DO MODELO DE COMPETÊNCIA PARA CIÊNCIA DE DADOS

Fator	Indicador	p-valor	Cargas padronizadas
Tecnologia	tecAlgoritmos	< 0,001	0,753
	tecBancoDeDados	< 0,002	0,701
	tecDadosSemiEst	< 0,003	0,702
	tecDadosNaoEst	< 0,004	0,705
	tecDesenvSoftware	< 0,005	0,683
	tecEngDeDados	< 0,006	0,784
	tecGestaoDeDados	< 0,007	0,828
	tecHabHacking	< 0,008	0,756
	tecAIML	< 0,009	0,756
	tecProgramacao	< 0,010	0,663
	tecSistemas	< 0,011	0,674
	tecDadosDistribuidos	< 0,012	0,692
Análise de Dados	anAnaliseDeDados	< 0,013	0,859
	anAnalisesPreditivas	< 0,014	0,833
	anEstatistica	< 0,015	0,735
	anMatematica	< 0,016	0,638
	anMineracaoDeDados	< 0,017	0,852
	anModelagemGrafos	< 0,018	0,750
	anOtimizacao	< 0,019	0,791
	anNLP	< 0,020	0,725
	anFormQuestoes	< 0,021	0,761
	anMedotoCientifico	< 0,022	0,707
	anVisualizacao	< 0,023	0,749
	anRegressao	< 0,024	0,755
Entendimento de Negócios	enConhecimento	< 0,025	0,790
	enDesenvolvimentoDeP	< 0,026	0,802
	enGestaoDeProjetos	< 0,027	0,918
	enNegocios	< 0,028	0,760
	enFinanceiro	< 0,029	0,748
	enGovernanca	< 0,030	0,826
	enCompliance	< 0,031	0,802
	enLGPD	< 0,032	0,744
Competências Socioculturais	socInterdisciplinari	< 0,033	0,829
	socComunicacao	< 0,034	0,777
	socLidTec	< 0,035	0,767
	socLidEst	< 0,036	0,736
	socSolucaoProblemas	< 0,037	0,913
	socColaboracao	**0,127**	**0,148**
	socCriatividade	**0,167**	**0,117**
	socCuriosidade	0,005	**0,283**
	socEtica	**0,150**	**0,138**
	socDadosSensiveis	**0,715**	**0,028**
	socBias	0,029	**0,191**

FONTE: O autor (2022).

Ao considerar apenas os indicadores de significância e ajustes, este primeiro modelo apresentou resultados aceitáveis. Mesmo que p-valor do χ^2 (1274,44) tenha valor inferior a 0,05, a sua razão pelos graus de liberdade ($gl = 856$) apresentou valor de 1,49 dentro do limite indicado de 2,00 (SCHREIBER *et al.*, 2006). O valor de RMSEA (0,047) também ficou dentro do aceitável, bem como os índices de ajuste

(GFI = 0,959, TLI = 0,982 e CFI = 0,982). No entanto, o valor de SRMR (0,094) ficou acima do limite esperado que é de 0,08.

Além disso, como demonstrado pela cor vermelha na Tabela 7, alguns indicadores não apresentaram significância estatística (*socColaboracao*, *socCriatividade*, *socEtica* e *socDadosSensiveis*) ou apresentaram cargas fatoriais padronizadas abaixo de 0,5 (*socCuriosidade* e *socBias*). Por esta razão, optou-se pela remoção destas variáveis do modelo, uma que itens com baixas cargas podem ser removidos dos modelos reflexivos sem grandes consequências quando o construto mantém o mínimo de três ou quatro indicadores (HAIR *et al.*, 2014), caso do fator Competências Socioculturais.

Juntamente com a remoção das variáveis citadas, foram analisados os 10 índices de modificação que trariam maior efeito para o χ^2, apresentados na Tabela 8. Neste processo, percebeu-se que a variável referente ao domínio de técnicas de inteligência artificial (*tecAIML*) apresentava covariância residual com seis variáveis do fator Análise de Dados. Logo, verificou-se que a variável em questão apresentou maior relação com este grupo de variáveis em relação ao grupo de Tecnologia. Por isso, a variável *tecAIML* foi deslocada do fator Tecnologia para o fator Análise de Dados, decisão apoiada em Hair et al. (2014) que afirma que mudanças entre itens podem levar a um modelo melhor do que muitas alterações nos índices de modificação.

TABELA 8 – 10 ÍNDICES DE MODIFICAÇÃO PRIORITÁRIOS

			Mod. Ind.	EPC
socDadosSensiveis	↔	socBias	24.154	2.111
tecAIML	↔	anAnalisesPreditivas	23.835	2.487
tecAIML	↔	anRegressao	23.143	2.767
socColaboracao	↔	socDadosSensiveis	17.904	1.713
tecAIML	↔	anOtimizacao	17.147	1.888
socCriatividade	↔	socCuriosidade	16.154	1.654
tecAIML	↔	anMatematica	15.959	1.932
tecAIML	↔	anNLP	15.894	1.851
tecAIML	↔	anMineracaoDeDados	15.811	1.964
tecDadosNaoEst	↔	anNLP	15.782	1.799

FONTE: O autor (2022).
NOTA: Índice de modificação (Mod. Ind.): mostra o quanto o valor do qui-quadrado (χ^2) do ajuste geral mudaria se o parâmetro em questão fosse liberado. *Expected parameter change* (EPC): mostra a mudança esperada do próprio parâmetro, quando a alteração for realizada.

Diante destes ajustes (remoção das variáveis sem significância estatística, remoção das variáveis com cargas fatoriais baixa e deslocamento da variável *tecAIML*

para o construto Análise de Dados), foram verificadas novamente as cargas fatoriais, demonstradas na Tabela 9, e os indicadores de ajuste do modelo.

TABELA 9 – CARGAS FATORAIS DO MODELO DE COMPETÊNCIA PARA CIÊNCIA DE DADOS (MODELO AJUSTADO)

Fator	Indicador	p-valor	Cargas padronizadas
Tecnologia	tecAlgoritmos	< 0,001	0,765
	tecBancoDeDados	< 0,001	0,722
	tecDadosSemiEst	< 0,001	0,719
	tecDadosNaoEst	< 0,001	0,720
	tecDesenvSoftware	< 0,001	0,706
	tecEngDeDados	< 0,001	0,810
	tecGestaoDeDados	< 0,001	0,841
	tecHabHacking	< 0,001	0,764
	tecProgramacao	< 0,001	0,683
	tecSistemas	< 0,001	0,701
	tecDadosDistribuidos	< 0,001	0,704
Análise de Dados	anAnaliseDeDados	< 0,001	0,856
	anAnalisesPreditivas	< 0,001	0,840
	anEstatistica	< 0,001	0,740
	anMatematica	< 0,001	0,640
	anMineracaoDeDados	< 0,001	0,854
	anModelagemGrafos	< 0,001	0,750
	anOtimizacao	< 0,001	0,792
	anNLP	< 0,001	0,728
	anFormQuestoes	< 0,001	0,758
	anMedotoCientifico	< 0,001	0,711
	anVisualizacao	< 0,001	0,745
	anRegressao	< 0,001	0,765
	tecAIML	< 0,001	0,766
Entendimento de Negócios	enConhecimento	< 0,001	0,790
	enDesenvolvimentoDeP	< 0,001	0,805
	enGestaoDeProjetos	< 0,001	0,916
	enNegocios	< 0,001	0,763
	enFinanceiro	< 0,001	0,749
	enGovernanca	< 0,001	0,827
	enCompliance	< 0,001	0,802
	enLGPD	< 0,001	0,739
Competências Socioculturais	socInterdisciplinari	< 0,001	0,840
	socComunicacao	< 0,001	0,789
	socLidTec	< 0,001	0,784
	socLidEst	< 0,001	0,751
	socSolucaoProblemas	< 0,001	0,923

FONTE: O autor (2022).

Nesta versão ajustada do modelo, todos os indicadores mantiveram suas significâncias estatísticas, além de terem suas cargas fatoriais levemente aumentadas. Ademais, todos os indicadores de ajuste do modelo também obtiveram resultados superiores. O valor do χ^2, que ainda manteve sua significância estatística, diminuiu para 800,24, representando uma diferença de 474,20, enquanto o número de graus de liberdade (*gl*) passou a 625. Logo, a relação χ^2/gl também ficou menor: 1,28. RMSEA obteve um valor de 0,035, enquanto ficou próximo ao limite aceitável de 0,08. Por fim, todos os índices de ajustes também foram melhorados (GFI = 0,974, TLI = 0,992 e CFI = 0,992). A Tabela 10 traz uma comparação entre os indicadores dos dois modelos testados.

TABELA 10 – COMPARAÇÃO ENTRE OS AJUSTAMENTOS DOS MODELOS TESTADOS

Indicador	Modelo 01 (43 itens)	Modelo 02 (37 itens)
χ^2	1274,441	800,239
Graus de Liberdade (gl)	856	625
p-valor	<0,001	<0,001
χ^2/gl	1,488	1,28
RMSEA	0,047	0,035
GFI	0,959	0,974
TLI	0,982	0,992
CFI	0,983	0,993
SRMR	0,094	0,081
$\Delta\chi^2$	-	474,202

FONTE: O autor (2022).

O resultado obtido pelo modelo ajustado foi considerado satisfatório e permitiu avaliar as cargas fatoriais em relação ao construto de segunda ordem relativo ao conceito de Ciência de Dados. Este fator, também reflexivo, indica que a proficiência dos respondentes em tecnologia, análise de dados, entendimento de negócios e competências socioculturais é explicada pela competência como cientista de dados. A Tabela 11 apresenta os valores das cargas fatoriais padronizadas, onde todas as relações se mostram estatisticamente significantes.

TABELA 11 – CARGAS FATORIAIS DE SEGUNDA ORDEM

Fator	Indicador	p-valor	Cargas padronizadas
Ciência de Dados	Tecnologia	< 0,001	0,729
	Análise de Dados	< 0,001	0,778
	Entendimento de Negócios	< 0,001	0,864
	Competências Socioculturais	< 0,001	0,911

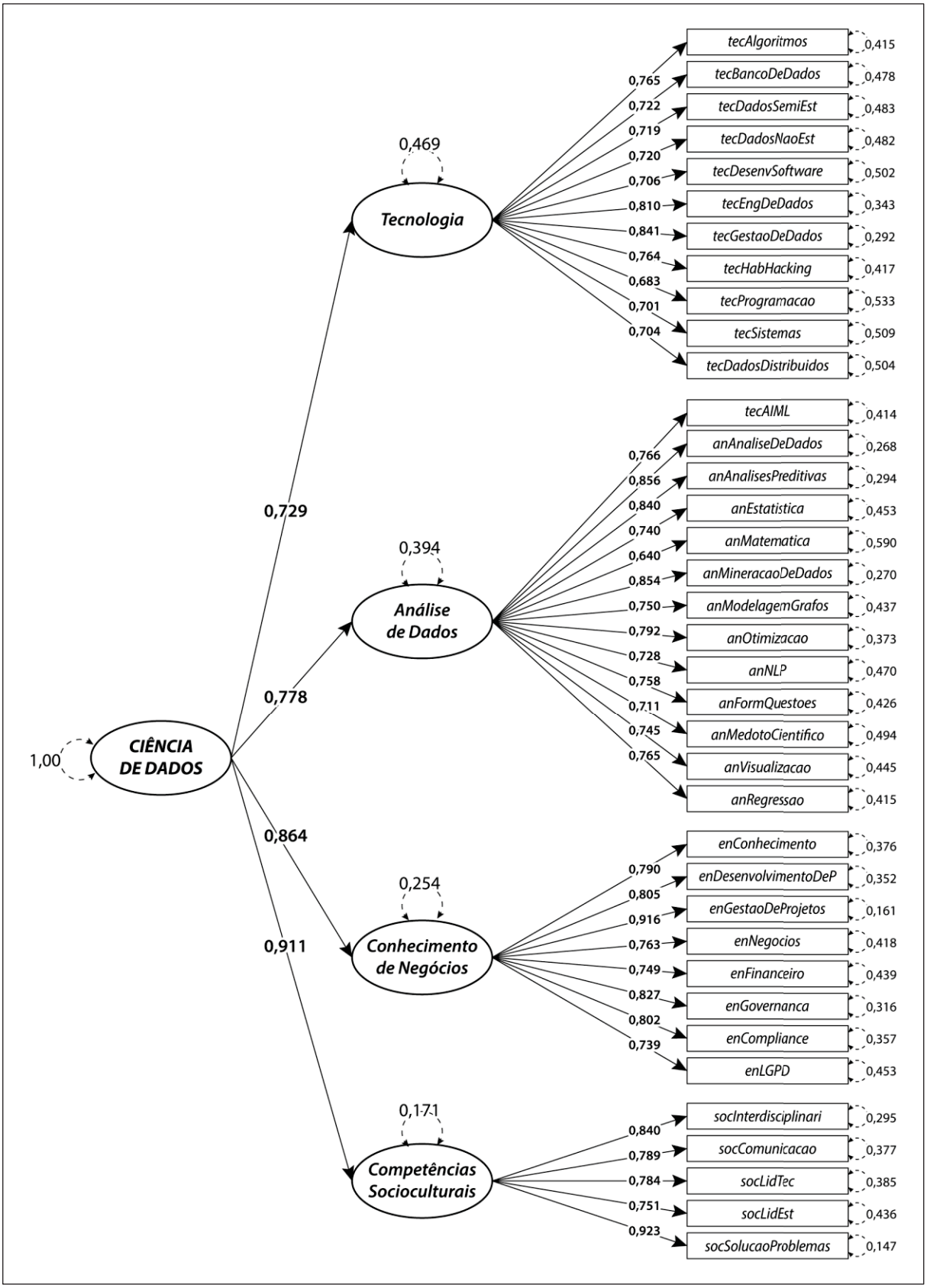
FONTE: O autor (2022).

Como último estágio do protocolo de análise, verificou-se a precisão da mensuração do modelo por meio da validade dos construtos. Ou seja, foi averiguado em que grau um grupo de variáveis mensuradas reflete a variável latente teórica na qual foram designadas (HAIR *et al.*, 2014). Dentre os três procedimentos realizados, o primeiro foi a confiabilidade dos construtos, ou confiabilidade composta (*Composite Reliability* / CR). Para que este indicador seja considerado satisfatório, espera-se valores entre 0,70 e 0,90, porém, em todos os construtos os valores obtidos foram superiores a este limite superior (*Tecnologia* = 0,954, *Análise de Dados* = 0,970, *Entendimento de Negócios* = 0,962 e *Competências Socioculturais* = 0,943). Neste caso, há indícios de redundância entre os indicadores dos construtos, o que reduz sua confiabilidade (HAIR *et al.*, 2019).

Em seguida, foi verificada a Variância Média Extraída (AVE), que mede o quanto um construto converge para explicar a variância de seus indicadores (HAIR *et al.*, 2019). Para este passo, todos os construtos apresentaram valores acima de 0,50, considerados satisfatórios (*Tecnologia* = 0,524, *Análise de Dados* = 0,585, *Entendimento de Negócios* = 0,641 e *Competências Socioculturais* = 0,651). Por fim, a razão Heterotrait-Monotrait (HTMT) foi adotada para avaliar a validade discriminante dos construtos, isto é, o nível em que um construto é empiricamente distinto dos outros construtos no modelo (HAIR *et al.*, 2019; HENSELER; RINGLE; SARSTEDT, 2015). Nesta etapa, todos os cruzamentos entre os fatores apresentaram valores de HTMT satisfatórios (abaixo de 0,90), exceto na relação entre os construtos Enten*dimento de Negócios* e *Competências Socioculturais*. O valor HTMT entre *Tecnologia* e *Análise de Dados* foi 0,592, entre *Tecnologia* e *Entendimento de Negócios*, 0,655 e, *Tecnologia* e *Competências Socioculturais*, 0,647. Já o construto *Análise de Dados* apresentou HTMT de 0,705 em relação ao construto *Entendimento de Negócios* e 0,762, para *Competências Socioculturais*. Porém, entre *Entendimento de Negócios* e *Competências Socioculturais*, o valor de HTMT foi 0,928, o que sugere ausência de validade discriminante entre os construtos (HAIR *et al.*, 2019).

Assim, com todos os passos do protocolo de análise executados, o modelo de competências da Ciência de Dados obtido e validado pelo procedimento de Análise Fatorial Confirmatória é apresentado na Figura 42.

FIGURA 42 – MODELO DA ANÁLISE FATORIAL CONFIRMATÓRIA DAS COMPETÊNCIAS DA CIÊNCIA DE DADOS



FONTE: O autor (2022).

Conforme relatado, o modelo apresentado é válido estatisticamente. Todavia, a manutenção, a remoção ou o deslocamento de variáveis da estrutura apresentada são evento que trazem informações para a pesquisa e demandam reflexão acerca de seu significado. As duas próximas seções se destinam a essa tarefa de discutir os resultados obtidos e apresentar as conclusões desta parte da pesquisa.

6.1.3 Discussão

A Ciência de Dados é um campo jovem, em desenvolvimento e, como tal, apresenta oportunidades e desafios, dentre diversos aspectos a serem explorados e discutidos (CAO, 2017; JORDAN; MITCHELL, 2015; SIEBES, 2018). Os resultados desta etapa da pesquisa vão ao encontro dessa perspectiva, pois mostram os atuantes na área como profissionais jovens, com 81,05% dos respondentes tendo entre 22 e 37 anos de idade. Além disso, com sete em cada 10 profissionais afirmando ter até seis anos de atuação na área, o tempo de experiência também é um indício do quanto a Ciência de Dados ainda tende a se desenvolver no Brasil. Esses achados corroboram com pesquisas anteriores como a *The Burtch Works Study: Salaries of Predictive Analytics Professionals* (BURTCH WORKS, 2019) e *State of Machine Learning and Data Science*, promovida anualmente pela Kaggle (2021).

A pesquisa de levantamento também reforça outros dois aspectos da Ciência de Dados: ser uma área majoritariamente masculina e altamente qualificada. Novamente, como demonstrado pelas pesquisas de Burtch (2019) e da comunidade Kaggle (2021), cerca de 80% dos profissionais da Ciência de Dados são homens. Além disso, Burtch (2019) salienta que esta proporção é ainda mais expressiva nos níveis de gerência, ultrapassando 90%. Em relação à qualificação dos profissionais, os resultados também fortalecem levantamentos anteriores, indicando que a educação formal é extremamente relevante para a atuação na área, pois 87,22% possuem nível superior completo. Os níveis de mestrado e doutorado, que segundo Burtch (2019) estão relacionados a uma maior remuneração, representam juntos 41,85% dos respondentes, ainda que muitos profissionais estejam com curso em andamento.

Por outro lado, chama a atenção o baixo número de respondentes advindos de cursos de graduação específicos para a Ciência de Dados. Enquanto Estatística e Ciência da Computação representam juntos quase 1/3 (66 respostas) dos

profissionais participantes da pesquisa, cursos em Ciência de Dados apresentaram somente seis respostas (2,64%). Novamente, este resultado corrobora com pesquisas anteriores reforçando que os cientistas de dados advêm de outras disciplinas já consolidadas como Estatística, Ciência da Computação, Matemática, Física, Economia (BURTCH WORKS, 2019; HARRIS; MURPHY; VAISMAN, 2013-; LOUKIDES, 2012). Há, portanto, um indicativo de criação de cursos específicos para futuros profissionais da área de Ciência de Dados, como antecipado por Tosic e Beeston (2018), Cao (2019) e, Demchenko, Comminiello e Reali (2019).

Em relação às competências, os resultados apontam que todos os quatro grupos (*Tecnologia*, *Análise de Dados*, *Entendimento de Negócios* e *Competências Socioculturais*) obtiveram média geral superior a cinco. Considerando a escala proposta no questionário, os grupos *Tecnologia* (média = 5,626) e *Entendimento de Negócios* (média = 5,375) apresentaram média equivalente ao nível intermediário de proficiência, enquanto *Análise de Dados* (média = 6,294) e *Competências Socioculturais* (média = 7,582) se enquadraram no nível avançado.

Visto que Estatística e Ciências da Computação foram os cursos de graduação mais recorrentes, constata-se que os níveis de proficiência em Tecnologia e Análise de Dados não apresentaram diferença acentuada. Porém, não se pode afirmar o mesmo em relação aos demais grupos. Como poucos respondentes afirmaram ter graduação na área de administração e gestão de negócios, o grupo de Entendimento de Negócio expôs o pior desempenho. Por outro lado, a mensuração das competências socioculturais, em especial aquelas relacionadas ao contexto organizacional, exibiu o melhor resultado.

De qualquer maneira, os profissionais participantes contemplaram muitos dos requisitos fundamentais a um “bom” cientista de dados, segundo os critérios elencados por Cao (2017, p. 31–32). Dentre eles, destaca-se, principalmente, a alta escolaridade, com ênfase nas áreas de tecnologia, estatística, análise e engenharias. Mas também capacidade de resolver problemas, trabalhando com múltiplos formatos de dados, em grandes *datasets*, adotando diferentes abordagens analíticas. Adicionalmente, as características retiradas do modelo, como colaboração, criatividade, curiosidade, bem como preocupação em relação à ética, dados sensíveis e viés de análise, não estão entre os itens que fazem um bom cientista de Cao (2017). Ainda assim, estas competências são recorrentes entre outros autores da área (BAŠKARADA; KORONIOS, 2017; HARRIS; MURPHY; VAISMAN, 2013-;

KELLEHER; TIERNEY, 2018; SCHOENHERR; SPEIER-PERO, 2015), incluindo o próprio Cao (2019).

Pelos resultados, é possível aferir ainda que os respondentes foram mais "generosos" com as perguntas relacionadas à organização onde atuam, sendo mais críticos com suas próprias competências. Este fato indica o motivo destas variáveis do grupo *Competências Socioculturais* terem sido retiradas do modelo. Uma alternativa é que o comportamento destas variáveis aponta para a presença de um novo construto no modelo.

Novamente, comparando com o modelo *EDISON data science framework* (DEMCHENKO *et al.*, 2016), a presente representação é mais sintética, sem adentrar nos processos da Ciência de Dados, como coleta de dados, identificação de padrões, teste de hipótese, dentre outros. Na presente proposta, a representação da Ciência de Dados é dada pelos quatro grupos de competências identificados na literatura que, na prática, não possuem fronteiras claramente delimitadas. Ressaltando, mais uma vez, o caráter interdisciplinar da atuação do cientista de dados (BRANDT, 2016; HAWAMDEH; CHANG, 2018; LOUKIDES, 2012; PARKS, 2017).

De forma geral, o modelo proposto, testado e validado, contribui para a Ciência de Dados, área que, apesar de promissora e necessária às organizações, carece de formalização de competências e papéis profissionais para seu desenvolvimento (CAO, 2019; EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS)DEMCHENKO; BELLOUM; WIKTORSKI, 2017; SALTZ; GRADY, 2017). Especificamente, o modelo contribui em duas situações ainda problemáticas: 1) formação do cientista de dados: área cuja formação curricular é foco de amplo debate e pesquisa (BAŠKARADA; KORONIOS, 2017; CAO, 2019; CURTY; SERAFIM, 2016; DONOHO, 2017); 2) a contratação destes profissionais: tarefa penosa e onerosa que apresenta dificuldades tanto na seleção quanto na retenção de talentos (DAVENPORT; PATIL, 2012; KAMPAKIS, 2020; REIS; SÁ, 2020).

6.1.4 Conclusões

Esta etapa da pesquisa valida, junto a profissionais brasileiros, os construtos *Tecnologia*, *Análise de Dados*, *Entendimento de Negócios* e *Competências Socioculturais* como um modelo propício à Ciência de Dados. Além disso, descreve a área pesquisada como um campo jovem, formado por indivíduos que atuam há pouco

tempo, cujas formações advêm de outras áreas já consolidadas. Por outro lado, demonstra o surgimento de cursos de nível superior específicos para a Ciência de Dados, apontando uma tendência de crescimento.

É neste contexto, onde a oferta de cursos e de empregos para cientistas de dados é crescente, que o modelo apresentado se torna útil. De forma geral, o delineamento das competências dos cientistas de dados contribui para a formação de novos cursos (planejamento curricular), para a contratação de profissionais e para a composição de equipes de dados. Além disso, esta pesquisa orienta o desenvolvimento de profissionais que estão iniciando ou mesmo migrando para a Ciência de Dados. Desta maneira, os esforços de instituições de ensino, organizações e profissionais podem ser direcionados e otimizados conforme seus objetivos.

Por outro lado, esta pesquisa de levantamento não está livre de limitações. Ainda que o modelo resultante da Análise Fatorial Confirmatória tenha apresentado bom ajustamento, enfatiza-se a configuração final que não se manteve idêntica àquela estipulada inicialmente. Dentre as modificações realizadas, destaca-se o deslocamento da variável referente à competência em Inteligência Artificial e *Machine Learning* (*tecAIML*), que foi retirada do construto *Tecnologia* e passada para o construto *Análise de Dados*. Além disso, também é frisada a remoção das variáveis que não apresentaram significância estatística ou carga fatorial superior a 0,5. Todas as variáveis removidas (*socColaboracao*, *socCriatividade*, *socEtica*, *socDadosSensiveis*, *socCuriosidade* e *socBias*) estavam inseridas no construto *Competências Socioculturais*, direcionadas ao contexto organizacional e não à própria competência do respondente, como as demais questões. Por isso, sugere-se que, em trabalhos futuros, seja verificada a formação de um novo construto que englobe estas competências organizacionais.

Ainda em relação ao modelo estatístico, deve-se analisar a confiabilidade composta dos construtos, onde todos os valores obtidos estiveram acima do recomendado. Neste sentido, uma revisão nas medidas mensuradas auxilia na identificação de redundância e, consequentemente, em uma redução do instrumento de coleta de dados. Igualmente em relação à validade dos construtos, recomenda-se considerar o valor de HTMT encontrado entre os fatores *Entendimento de Negócios* e *Competências Socioculturais*, considerando eventuais alterações no modelo testado.

Ao mesmo tempo, outros tópicos e relações podem ser explorados a partir dos dados coletados. Embora não estivessem no escopo desta pesquisa, o cruzamento

entre as variáveis da seção de informações adicionais poderia revelar novos aspectos dos profissionais que responderam ao questionário. Por exemplo, o grau de educação formal está relacionado à remuneração? E o tempo de experiência? Há diferença significativa entre os gêneros? Ou mesmo se a área de origem dos profissionais implica em diferença na remuneração.

Por fim, ainda que válido (HAIR *et al.*, 2014; MATSUNAGA, 2010), o tamanho da amostra é uma limitação a ser sanada por pesquisas futuras. Sugere-se ainda que se adote uma estratégia de estratificação da amostra por perfis profissionais (cientista de dados, engenheiro de dados, analista de dados, analista de BI, dentre outros) para identificar padrões entre os papéis desenvolvidos na Ciência de Dados. Com isso, o modelo de competência baseado nos fatores *Tecnologia*, *Análise de Dados*, *Entendimento de Negócios* e *Competências Socioculturais* poderia ser validado em outras amostras, reforçando e generalizando os resultados encontrados. Por fim, ainda que se julgue que as características da área no Brasil sejam distintas de outros países, recomenda-se que o modelo seja replicado e validado em pesquisas futuras dentro e fora do território brasileiro.

6.2 MINERAÇÃO DE TEXTO EM ANÚNCIOS DE VAGAS DE EMPREGO

Esta seção, que analisa os anúncios de vagas de emprego para cientistas de dados, é a primeira de três seções que utilizam métodos de mineração de texto a uma coleção de documentos. De maneira geral, as etapas são descrever a amostra, demonstrar os termos mais frequentes, identificar os tópicos abordados nos documentos e procurar por um padrão de agrupamento desses documentos, conforme detalhado na metodologia.

6.2.1 Características da amostra

Antes de descrever os dados em si, apresenta-se a participação de cada *website* na composição do *corpus*. Os valores dos dados coletados e filtrados, conforme critérios apresentados anteriormente, estão contidos na Tabela 12.

TABELA 12 – RESUMO DA COLETA DE DADOS

Base	Expressão	Anúncios	%	Filtrados	%
Indeed	cientista de dados	351	36,26%	250	40,78%
	data scientist	136			
LinkedIn	cientista de dados	217	24,72%	236	38,50%
	data scientist	115			
Catho	cientista de dados	159	17,42%	63	10,28%
	data scientist	75			
Empregos	cientista de dados	45	13,03%	42	6,85%
	data scientist	130			
Vagas	cientista de dados	31	2,31%	15	2,45%
	data scientist	0			
Infojobs	cientista de dados	10	6,25%	7	1,14%
	data scientist	74			
	TOTAL	**1343**	**100%**	**613**	**100%**

FONTE: O autor (2022).

Dos 613 anúncios eleitos para análise, 79,28% foram obtidos em duas únicas fontes: Indeed e LinkedIn, ambas de origem estrangeira. Estes valores confirmam a relevância do site Indeed para a área da Ciência de Dados que já tinha sido apresentada por Kim e Lee (2016). Por outro lado, a empresa Infojobs, de origem espanhola, apresenta a menor contribuição com apenas sete anúncios analisados (1,14%). Em relação à Infojobs, também chama a atenção que dos 84 anúncios coletados inicialmente foram excluídos 77, principalmente, por não conterem em seus títulos “cientista de dados” ou “data scien”. Em relação às empresas de origem brasileira (Catho, Empregos e Vagas), pode-se constatar que juntos compuseram quase 1/5 das vagas analisadas (19,58%).

Outra característica ponderada foi quanto à localidade das vagas referentes aos anúncios coletados. A cidade de São Paulo concentrou quase 1/3 da amostra com 193 anúncios (31,48%). Na segunda posição, com 102 anúncios (16,64%), estão as vagas destinada a trabalho remoto (*home office*), seguidas pela cidade do Rio de Janeiro com 51 anúncios (8,32%). As demais cidades apresentaram menos de 22 anúncios, não ultrapassando 4,00% do total. Em relação aos estados brasileiros, somente São Paulo, com 260 anúncios, corresponde a 42,41% dos documentos analisados, impulsionado naturalmente pela demanda apresentada pela sua capital.

O salário oferecido foi outra informação analisada, porém, apenas 98 anúncios (15,99%) faziam menção a este item, dos quais 62 exibiam a mensagem “A combinar”.

Desta forma, as remunerações puderam ser analisadas em somente 36 anúncios ou 5,87% da amostra. Ainda assim, foi possível extrair alguns dados relevantes: o valor médio mensal oferecido é R$ 6.782,03 (desvio padrão de R$ 542,47), sendo o menor valor apresentado correspondendo à faixa salarial “De R$ 1.001,00 a R$ 2.000,00” e o maior, R$ 15.000,00. Considerando a cotação do Real[9], a moeda brasileira vale 0,20 dólares americanos e implica em uma média salarial mensal de U$ 1,356.40 ou um valor anual de cerca de U$ 16,276.87. Este valor equivale a 1/6 da média do salário anual de um cientista de dados iniciante nos EUA em 2019 (BURTCH WORKS, 2019).

Em relação à experiência exigida, foram analisados os títulos das vagas procurando pelos termos Júnior (ou Jr), Pleno e Sênior. Dos 613 anúncios, 184 (30,02%) mencionavam em seu título ao menos um dos termos procurados. O nível sênior foi o mais procurado com 103 ocorrências (16,80%), seguido pelos profissionais plenos (70 anúncios, 11,42%) e, por fim, pelos profissionais do nível Júnior (11 anúncios, 1,79%). O que indica uma maior demanda por profissionais mais experientes.

Para concluir esta etapa, foram identificados os anúncios que citavam algum nível de escolaridade. Primeiramente, verificou-se que 209 anúncios, ou 34,09% do *corpus*, continham a expressão “graduação” ou “ensino superior”. Se por um lado, 1/3 dos anúncios apontam que os contratantes valorizam a educação superior, por outro, expõe que 2/3 não expressam a educação formal como item fundamental à seleção de candidatos. Já os cursos *stricto sensu* são menos exigidos, sendo que mestrado aparece em 53 anúncios (8,65%) e doutorado, em 38 anúncios (6,20%). Estes valores são muito inferiores aos apresentados por Kim e Lee (2016), que apontam que de 1.240 anúncios relativos a cientista de dados, 645 (52,01%) citam o nível de mestre e 561 (45,20%), o nível de doutoramento como diferencial.

6.2.2 Termos mais frequentes

[9] Segundo cotação do Banco Central do Brasil, realizada em 03 de março de 2022, o valor comercial do dólar dos Estados Unidos da América é R$ 5,05 (https://www.bcb.gov.br/estabilidadefinanceira/historicocotacoes).

Para mostrar uma visão geral da frequência dos dados, a Figura 35 traz a nuvem de palavras dos termos tratados, porém não convertidos em *stems*, sem explicitar qualquer informação numérica.

FIGURA 43 – NUVEM DE PALAVRAS COM OS TERMOS MAIS FREQUENTES PARA ANÚNCIOS DE EMPREGO

FONTE: O autor (2022).

Como esperado, a palavra “dados” é a que mais se destaca por fazer parte de uma das expressões de busca (“cientista de dados”), mas também por compor expressões caraterísticas da área como “manipulação de dados”, “processamento de dados”, “análise de dados”, “*pipeline* de dados”, “ciência de dados”, dentre outras. Entre outras palavras que se destacam, estão “modelos” e “machine”, remetendo ao desenvolvimento de modelos de aprendizagem de máquina; “conhecimento” e “experiência” que se referem ao desenvolvimento prévio de competências na área. A seguir, são apresentados os resultados obtidos pela análise n-grama aplicada aos *stems* gerados. Os 20 termos mais frequentes 1-grama são listados na Tabela 13:

TABELA 13 – 1-GRAMA PARA ANÚNCIOS

#	Termo	F	N	%
1	dad	2911	567	92,50%
2	model	1447	478	77,98%
3	conhec	1441	499	81,40%
4	experienc	1372	481	78,47%
5	desenvolv	1127	451	73,57%
6	analis	986	435	70,96%
7	negoci	938	379	61,83%
8	estatis	823	425	69,33%
9	are	783	406	66,23%
10	learning	703	364	59,38%
11	pesso	648	263	42,90%
12	machin	605	351	57,26%
13	dat	602	291	47,47%
14	tecn	602	327	53,34%
15	empr	601	323	52,69%
16	tim	586	283	46,17%
17	tecnolog	567	297	48,45%
18	python	565	464	75,69%
19	client	559	265	43,23%
20	process	552	309	50,41%

FONTE: O autor (2022).
LEGENDA: F é quantidade de vezes que o termo aparece no *corpus*. N é o número de documentos nos quais o termo aparece. % é a porcentagem do número de documento em relação ao total do *corpus*.

A classificação 1-grama, formada pelos *stems*, confirmou “dado” como o termo mais recorrente nos anúncios, com 2.911 ocorrências e presente em 567 documentos analisados (92,50%). Em seguida, com menos da metade de ocorrência, vem o *stem* “model” com 1.447 aparições em 478 documentos distintos. Já “conhec”, referente a conhecimento, está presente em mais anúncios (499), mas em menor frequência (1.441). De maneira geral, o *ranking* 1-grama vai ao encontro da nuvem de palavras, enfatizando dado, modelo, conhecimento, experiência, desenvolvimento e análise como as palavras presentes em mais de 70% dos documentos analisados. Além disso, percebe-se entre as 20 palavras mais recorrentes, termos relativos ao contexto organizacional como “empresa”, “negócios”, “pessoas”, “time” e “clientes”. Nos anúncios, estes termos são adotados para descrever a própria empresa contratante, mas também para tratar do cotidiano do futuro contratado.

TABELA 14 – 2-GRAMA PARA ANÚNCIOS

#	Termo	F	N	%
1	machin learning	580	348	56,77%
2	cient dad	275	195	31,81%
3	analis dad	272	196	31,97%
4	cienc dad	249	162	26,43%
5	model estatis	218	176	28,71%
6	banc dad	193	156	25,45%
7	desenvolv model	162	130	21,21%
8	dat scienc	160	105	17,13%
9	big dat	158	126	20,55%
10	cienc computaca	153	142	23,16%
11	aprend maquin	139	92	15,01%
12	model predi	134	109	17,78%
13	segur vid	133	133	21,70%
14	inteligenc artific	132	107	17,46%
15	lingu programaca	129	110	17,94%
16	superi complet	127	127	20,72%
17	hom offic	115	105	17,13%
18	programaca python	107	94	15,33%
19	conhec avanc	106	52	8,48%
20	model machin	105	79	12,89%

FONTE: O autor (2022).

A lista com os termos 2-grama (Tabela 14) demonstra a expressão referente a “machine learning” como a mais recorrente entre os anúncios, aparecendo 580 vezes em 348 documentos, ou seja, 56,77% do *corpus*. Em seguida, os *stems* de “cientista de dados” com 275 ocorrências em 195 anúncios (31,81%) e de “análise de dados”, 272 ocorrências, 196 anúncios (31,97%). As outras expressões presentes em pelo menos 1/5 dos anúncios são: “ciência de dados”, “modelo estatístico”, “banco de dados”, “desenvolvimento de modelo”, “big data” e “ciência da computação”. A adoção de expressão aumenta o valor semântico dos itens da classificação, especificando as competências mais recorrentes, como o desenvolvimento de modelos estatísticos, administração de banco de dados e formação em ciências da computação. Neste sentido, outra expressão presente em mais de 20% dos anúncios é “superior completo”, ressaltando a valorização da educação formal, e “seguro de vida”, um dos benefícios mais recorrentes nos anúncios analisados. A seguir, a Tabela 15 apresenta as principais expressões formadas por três *stems*.

TABELA 15 – 3-GRAMA PARA ANÚNCIOS

#	Termo	F	N	%
1	model machin learning	97	77	12,56%
2	grand volum dad	65	53	8,65%
3	ensin superi complet	60	59	9,62%
4	machin learning deep	57	51	8,32%
5	learning deep learning	55	50	8,16%
6	banc dad relac	52	48	7,83%
7	estatis machin learning	50	48	7,83%
8	cienc computaca engenh	47	45	7,34%
9	lingu programaca python	47	40	6,53%
10	desenvolv model estatis	45	43	7,01%
11	plan saud plan	38	38	6,20%
12	saud plan odontolog	38	38	6,20%
13	algoritm machin learning	37	32	5,22%
14	tecn machin learning	34	35	5,71%
15	dad machin learning	33	33	5,38%
16	process lingu natur	31	31	5,06%
17	conhec machin learning	30	30	4,89%
18	medic assistenc odontolog	30	30	4,89%
19	banc dad sql	30	33	5,38%
20	machin learning conhec	29	29	4,73%

FONTE: O autor (2022).

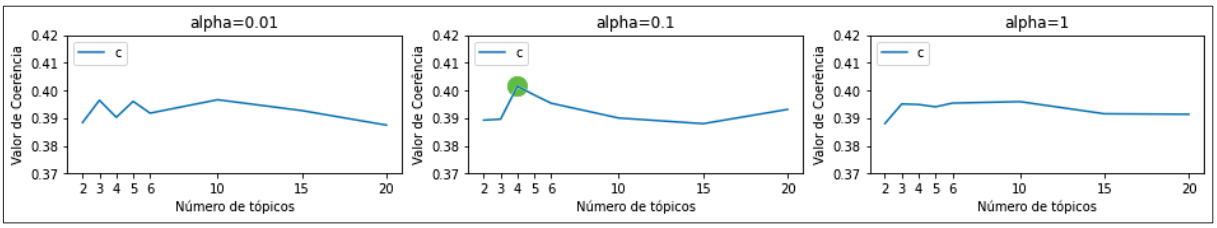
Novamente, o termo mais frequente está relacionado à aprendizagem de máquina. A expressão referente a “modelos de *machine learning*” foi a única presente em mais de 10% do *corpus*, aparecendo em 77 anúncios. Ademais, ressalta-se que “machine learning” está presente em oito expressões 3-grama dentre as 20 mais recorrentes. A segunda expressão mais frequente foi “grande volume de dados”, referente à “matéria-prima” do cientista de dados, seguido por “ensino superior completo”, que ocorre em 9,62% dos anúncios. Mais uma vez, a educação formal está presente entre os termos mais recorrentes, porém em um patamar muito abaixo do que demonstrado em outras pesquisas (CURTY; SERAFIM, 2016; KIM; LEE, 2016).

6.2.3 Modelagem de tópicos

Inicialmente, procedeu-se pela verificação de coerência, medida utilizada para avaliar os modelos gerados, para definir a quantidade de tópicos a serem extraídos. Para isso, fez-se o cruzamento de possibilidades de número de tópicos (2, 3, 4, 5, 6, 10, 15, 20) com três de variações do valor de alfa (0,01, 0,1 e 1), que é um parâmetro

que influencia na definição de tópicos para cada documento. O resultado deste cruzamento pode ser visto na Figura 44:

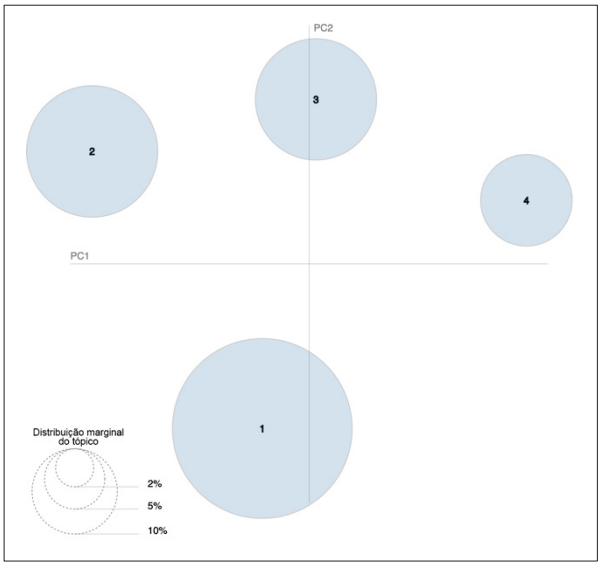
FIGURA 44 – VERIFICAÇÃO DOS VALORES DE COERÊNCIA PARA DIFERENTES MODELOS DOS ANÚNCIOS



FONTE: O autor (2022).

Os valores de coerência obtidos variaram entre 0,387 e 0,401, sendo este maior valor encontrado para a combinação de quatro tópicos e o valor de alfa de 0,10. Uma vez que o valor de coerência pode variar de 0 a 1, onde valores próximos a 0,7 são esperados, a coerência obtida é considerada baixa, mas não impede a utilização da LDA. Assim, com a definição do número de tópicos, procedeu-se à definição do modelo cuja visualização, desenvolvida com a biblioteca LDAVis (SIEVERT; SHIRLEY, 2014), é apresentada na Figura 45:

FIGURA 45 – VISUALIZAÇÃO DOS TÓPICOS COM A LDAVIS PARA ANÚNCIOS

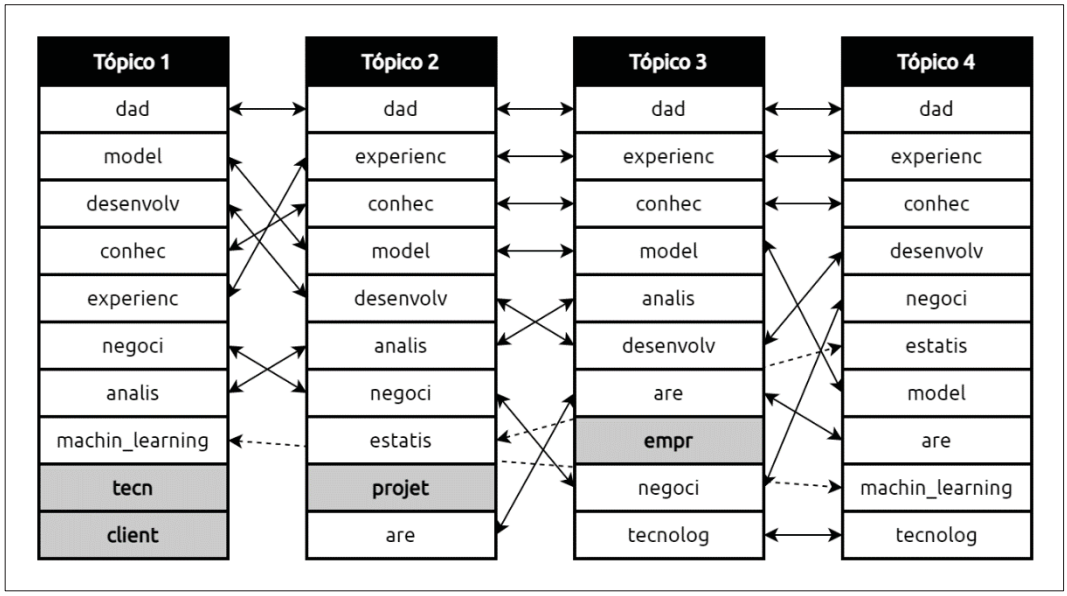


FONTE: O autor (2022).

Em princípio, a visualização apresenta um modelo adequado onde os tópicos estão distantes, não havendo sobreposição, e todos os quatro tópicos, segundo a distribuição marginal, são relevantes para grande parte do *corpus*. Porém, quando se

verifica os 10 principais termos que compõem cada tópico, percebe-se que os termos encontrados são praticamente os mesmos, diferenciando-se apenas pela posição, ou seja, pelo peso que o termo representa a cada tópico. Termos únicos (entre os 10 principais) acontecem apenas nos três primeiros tópicos. No primeiro tópico, os *stems* “tecn”, referente a técnica, e “client”, de cliente, não aparecem nas primeiras posições dos demais tópicos. Para o segundo tópico, o termo exclusivo se refere a “projeto” e para o terceiro, o termo é “empr”, referente a empresa. A Figura 46 traz os 10 principais termos de cada tópico, demonstrando as variações nas posições entre cada tópico definido.

FIGURA 46 – PRINCIPAIS TERMOS PARA CADA TÓPICO DA LDA PARA ANÚNCIOS



FONTE: O autor (2022).

A similaridade encontrada entre os tópicos definidos, que se repetiu em modelos com dois, três, cinco e seis tópicos, tornou incoerente a usual nomeação dos tópicos, vide que todos “tratam” de temas muito parecidos. Em outras palavras, os quatro tópicos tratam do mesmo tema, com poucas e sutis diferenças entre os pesos dos termos principais. Desta forma, julga-se que os resultados obtidos pela LDA pouco contribuem para extrair informações da amostra analisada.

6.2.4 Análise de agrupamento

Com a utilização da ferramenta *online Clustering Workbench* (CARROT[2] CLUSTERING ENGINE, 2021), a partir de uma base de dados já tratada, é possível aplicar algoritmos de agrupamento, dentre eles o K-Means. Os únicos parâmetros alterados na configuração padrão da ferramenta foram o número de *clusters*, definido em quatro para seguir o mesmo número obtido pela LDA, número de rótulos exibidos em cada grupo, definido como 10, e o idioma, alterado para português. Nesta situação, todos os 613 documentos (anúncios) foram agrupados nos seguintes *clusters*:

1. Time, *Data*, **Negócio**, **Novas**, Empresa, Processos, ***Learning***, ***Machine***, **Algoritmos**, **Atuar** *(367 documentos)*
2. Time, Modelos, **Diversidade**, **Crédito**, **Plano**, **Oferecemos**, **Saúde**, **Vida**, **Auxílio**, Máquina *(93 documentos)*
3. **Projetos**, Processos, **Serviços**, **Financeiro**, **Ciência**, Análise, **Experiência**, **Aprendizado**, Habilidades, Máquina *(77 documentos)*
4. *Data*, Modelos, **Tecnologia**, Empresa, **Avançado**, Análise, **Modelagem**, **Área**, **Analytics**, Habilidades *(76 documentos)*

O primeiro *cluster*, que engloba 59,87% do total de anúncios, apresenta exclusividade nos termos “negócio”, “novas” (de “novas técnicas”, “novas ferramentas”, “novas práticas”, “novas soluções”, demonstrado relação com inovação e aprendizagem, conforme apontado por Cao (2019)), “machine”, “learning”, “algoritmos” e “atuar”. Os termos sem negritos são aqueles presentes em mais de um grupo, como “time”, referente aos profissionais já presentes na organização contratante, “data”, dados no idioma inglês, “empresa”, “processos”, dentre outros.

Os outros três agrupamentos estão equilibrados em relação à quantidade de documentos englobados, 15,17%, 12,56% e 12,40%, respectivamente. O segundo agrupamento é marcado por termos relacionados a benefícios que a empresa oferece aos seus funcionários, bem como a ênfase à diversidade. O terceiro, apresenta os termos “Financeiro” e “Ciência” como destaques de exclusividade, enquanto o quarto apresenta “Tecnologia”, “Modelagem” e “Avançado”, termo geralmente utilizado juntamente com “Conhecimento”.

Ainda que haja termos comuns a mais de um *cluster*, percebe-se uma maior distinção entre os grupos pela utilização do algoritmo *K-Means* do que pela aplicação

do LDA. Logicamente, os objetivos das duas técnicas são distintos. Conforme, apontado por Bengfort et al. (2018-, p. 111), *clustering* procura estabelecer grupos em uma coleção de documentos, deixando cada documento em um grupo; enquanto a modelagem de tópico, busca abstrair os principais temas desta coleção, onde um único documento pode abranger mais de um tópico. A conclusão aqui é que o agrupamento se mostra mais definido que a abstração de tópicos.

6.2.5 Considerações preliminares sobre análise dos anúncios

Para identificar os principais requisitos presentes nos anúncios de emprego para cientistas de dados no Brasil, esta seção da pesquisa coletou anúncios em *websites* especializados e empregou técnicas de mineração de texto. As análises preliminares revelaram que as vagas ofertadas se concentram no estado de São Paulo, especialmente em sua capital e que o trabalho na modalidade remota já é a segunda opção encontrada. Além disso, verificou-se que a divulgação da remuneração ofertada não é uma prática adotada pela maioria das empresas e a média salarial obtida está muito abaixo dos valores apresentados por outras pesquisas (BURTCH WORKS, 2019; KAGGLE, 2021). Ainda em relação às características gerais dos anúncios, a pesquisa demonstrou a busca por profissionais mais experientes, de nível sênior. Todavia, a exigência por educação formal, especialmente por mestrado e doutorado, ficou aquém dos resultados apresentados por Kim e Lee (2016) e corroborou com Baumeister, Barbosa e Gomes (2020), que indicam que as companhias procuram por educação formal, mas este não é um tema central nos anúncios de emprego.

Em relação aos procedimentos de mineração de texto, as classificações de 1, 2 e 3-grama destacaram conceitos mais técnicos como *machine learning* (aprendizagem de máquina), modelos estatísticos, análise e ciência de dados, inteligência artificial, ciência da computação, tecnologia, python, banco de dados, modelos preditivos, processamento de linguagem natural, dentre outros. Por outro lado, também destacou conceitos relacionados ao ambiente organizacional, como negócios, pessoas, time, e referente à educação, como ensino superior completo. Além disso, também houve a presença de termos relacionados a características e benefícios, como home office, seguro de vida, plano de saúde e plano odontológico.

As técnicas de modelagem de tópicos (LDA) e agrupamento (K-Means), embora tenham apresentado desempenhos diferentes, reforçaram os termos-chave para a descoberta de temas centrais ou a classificação dos anúncios. Novamente, dados, *machine learning*, análise, experiência, conhecimento, tecnologia, modelagem estatística, entendimento de negócio, projeto, desenvolvimento (software), busca pelo novo são os termos que mais representam os requisitos para a Ciência de Dados. Este resultado vai ao encontro de pesquisas preliminares como Kim e Lee (2016), Meyer (2019), Halwani et al. (2021) e Gottipati et al. (2021), por exemplo. Em contrapartida, diferentemente do que outras pesquisas apontam, não foi possível identificar a frequência de habilidades interpessoais, como comunicação oral e escrita (BAUMEISTER; BARBOSA; GOMES, 2020; CAO, 2019; KIM; LEE, 2016) ou conhecimentos referentes a questões sociais, éticas e legais relacionadas a privacidade e segurança de dados (ANDERSON; MCGUFFEE; UMINSKY, 2014, p. 147; CURTY; SERAFIM, 2016).

Desta forma, julga-se que as ferramentas contribuíram para um esclarecimento acerca do que é solicitado a um candidato a cientista de dados no Brasil. Contudo, para aprimorar o desempenho das técnicas utilizadas, em especial em relação à LDA, há sugestões para pesquisas futuras. Dentre elas, a recomendação inicial é de aumentar o tamanho do *corpus*, visto que aplicações de NLP produzem resultados mais confiáveis diante de conjuntos de dados maiores (WOLFRAM, 2017). Ainda, a incorporação de outros perfis profissionais como engenheiros de dados, analistas de dados, estatísticos e outros. Este procedimento possibilitaria a rotulagem das vagas, permitindo a avaliação da acurácia dos algoritmos de agrupamento e da modelagem de tópicos, uma vez que a tendência aponta para conjuntos diferentes de requisitos entre as vagas.

Por fim, ainda que se julgue as técnicas de mineração de texto apropriadas para o objetivo da pesquisa, sugere-se para pesquisas futuras a aplicação de metodologias mistas, como análise de conteúdo, empregadas em pesquisas preliminares (BAUMEISTER; BARBOSA; GOMES, 2020; KIM; LEE, 2016; MEYER, 2019). Assim, se utilizados os mesmos dados, os resultados poderiam ser comparados e avaliados.

6.3 MINERAÇÃO DE TEXTO EM CURSOS SUPERIORES

Uma vez que as principais características da amostra de cursos superiores foram exploradas na metodologia, bem como o detalhamento do processo para se chegar a 34 documentos a serem analisados, esta seção se destina a apresentar os resultados das análises. Assim, seguindo os mesmos procedimentos adotados para os anúncios de vagas de emprego, o conteúdo dos cursos superiores passou pelas análises n-grama, *clustering* e modelagem de tópicos.

6.3.1 Termos mais frequentes

A nuvem de palavras, apresentada na Figura 47, traz novamente “dados” como o termo de maior destaque, seguido, agora, pelo termo “ciencia”, ganhando relevância se comparado à nuvem dos anúncios de emprego. Por outro lado, termos como “experiencia”, “learning”, “python”, dentro outros, perderam relevância entre os termos mais frequentes no conteúdo dos cursos superiores.

FIGURA 47 – NUVEM DE PALAVRAS COM OS TERMOS MAIS FREQUENTES PARA CURSOS SUPERIORES

FONTE: O autor (2022).

Agora, há o predomínio de termos mais abrangentes como “estatistica”, “analise (de dados)”, “sistema”, “metodos” e preferência da expressão “inteligência artificial” a “machine learning”, por exemplo. Todavia, a nuvem de palavras, que não traz informação numérica, não permite estabelecer relações precisas. Para isso, são adotadas as análises n-grama cuja classificação dos termos únicos (*stems*) é apresentada na Tabela 16:

TABELA 16 – 1-GRAMA PARA CURSOS SUPERIORES

#	Termo	F	N	%
1	dad	928	34	100,00%
2	cienc	293	31	91,18%
3	analis	242	32	94,12%
4	model	216	23	67,65%
5	projet	204	30	88,24%
6	desenvolv	189	29	85,29%
7	profiss	156	32	94,12%
8	estatis	152	31	91,18%
9	programaca	147	30	88,24%
10	sistem	145	24	70,59%
11	inteligenc	140	31	91,18%
12	are	135	32	94,12%
13	metod	132	21	61,76%
14	banc	128	30	88,24%
15	artific	125	28	82,35%
16	process	125	21	61,76%
17	merc	121	30	88,24%
18	tecnolog	121	32	94,12%
19	integr	117	19	55,88%
20	dat	113	31	91,18%

FONTE: O autor (2022).
LEGENDA: F é quantidade de vezes que o termo aparece no *corpus*. N é o número de documentos nos quais o termo aparece. % é a porcentagem do número de documento em relação ao total do *corpus*.

Novamente, o *stem* referente a “dados” foi o mais frequente, presente em todos os conteúdos analisados e com um número de ocorrências três vezes maior que o segundo *stem* mais frequente, “cienc”, presente em 91,18% dos documentos. Essa segunda colocação para o termo referente a ciência foi discrepante do resultado encontrado para os anúncios de vagas de emprego, uma vez que não figurou entre os 20 termos mais frequentes do *corpus* anterior. Além disso, verifica-se que neste

ranking dos termos mais comuns, os anúncios e os cursos superiores compartilham nove *stems*: “dad”, “analis”, “model”, “desenvolv”, “estatis”, “are”, “process”, “tecnolog” e “dat”. Por outro lado, as faculdades trazem novos termos à classificação dos mais frequentes: “projet”, “profiss”, “programaca”, “sistem”, “inteligenc”, “metod”, “banc”, “artific”, “merc” e “integr”.

Neste rol de palavras que adentraram a lista dos termos mais frequentes, destacam-se termos acadêmicos, relacionados a nomes comuns de disciplinas, e comerciais, relacionados à promoção dos cursos. O *stem* “projeto”, quinto termo mais encontrado, com 204 ocorrências em 30 documentos (88,24%), está associado a nomes de disciplinas como “Projeto em Ciência de Dados”, “Projeto Integrador/Integrado”, “Projeto Aplicado”. Já “profiss” (156 ocorrências, em 32 documentos) e “merc” (121 ocorrências, em 30 documentos) aparecem em expressões presentes nos sites institucionais como:

- Pesquisas apontam que **profissionais** de Ciência de Dados são os mais procurados do século XXI.
- ...grande demanda do mercado nacional e internacional por **profissionais** de Data Science, com bons salários e qualidade de vida relacionada a essa profissão.
- Formar **profissionais** com competências sólidas para a área de Ciência de Dados...
- Focado nas necessidades do **mercado**: Considerada uma das **profissões** mais promissoras até 2030, o curso possibilita o emprego de sistemas inteligentes, contribuindo para o avanço da transformação digital.
- Em pleno crescimento, o **mercado** de trabalho para quem se forma no curso de Ciência de Dados está repleto de vagas
- O **mercado** de trabalho para o cientista de dados é promissor, em razão do vertiginoso avanço tecnológico, especialmente da Inteligência Artificial.

Por um lado, a importância destes termos indica a busca das IES em formar profissionais que atendam às demandas do mercado, além de atrair alunos para a

área. Por outro, nota-se que nos cursos superiores há conteúdo que não é necessariamente destacado pelas empresas em seus anúncios para cientistas de dados. Inicialmente, é visto que os termos que compõem a expressão *machine learning* não figuram entre os mais recorrentes. Além disso, a expressão "integr", além de "projeto integrado" ou "projeto integrador", relaciona-se à Matemática, especialmente, em disciplinas como "Cálculo diferencial e integral" e "Aplicações da integral". Outra distinção percebida é que enquanto os cursos superiores destacam a programação de maneira geral, os anúncios reforçam o termo "python", referente a uma linguagem de programação específica muito utilizada na Ciência de Dados.

Para adicionar mais informação à análise, a Tabela 17 traz as expressões mais frequentes formadas por dois *stems*:

TABELA 17 – 2-GRAMA PARA CURSOS SUPERIORES

#	Termo	F	N	%
1	cienc dad	928	30	88,24%
2	inteligenc artific	293	28	82,35%
3	banc dad	242	30	88,24%
4	analis dad	216	23	67,65%
5	dad inteligenc	204	13	38,24%
6	big dat	189	27	79,41%
7	projet integr	156	9	26,47%
8	machin learning	152	21	61,76%
9	estrut dad	147	20	58,82%
10	mineraca dad	145	20	58,82%
11	lingu programaca	140	17	50,00%
12	dat scienc	132	10	29,41%
13	aprend maquin	128	15	44,12%
14	cient dad	125	16	47,06%
15	seri tempor	125	11	32,35%
16	tom decisa	121	13	38,24%
17	lingu natur	121	13	38,24%
18	volum dad	117	10	29,41%
19	red neur	113	12	35,29%
20	process lingu	112	13	38,24%

FONTE: O autor (2022).

A lista 2-grama traz "ciência de dados" como a expressão mais recorrente, o que não aconteceu para os anúncios de emprego. Aqui, nota-se pelo conteúdo coletado, que há citação dos nomes dos próprios cursos, há disciplinas com "ciência de dados" em seus títulos, mas também se percebe que os textos utilizados nas páginas dos cursos explicam o conceito de Ciência de Dados para possíveis futuros

alunos. Por isso, a expressão é a mais comum, com 928 ocorrências, embora não esteja presente em todos os documentos analisados. Em quatro documentos, são priorizadas as expressões “cientista de dados” e “data science” em detrimento a “ciência de dados”.

As próximas três expressões mais recorrentes (“inteligenc artific”, “banc dad” e “analis dad”) também estavam entre as frequentes nos anúncios de emprego, bem como “big dat”, “projet integr”, “machin learning”, “estrut dad”, “mineraca dad”, “lingu programaca”, “dat scienc”, “aprend maquin” e “cient dad”. Dessa forma, se as duas listas compartilham 50% de seus termos, metade das expressões bigramas dos cursos superiores é formada por termos exclusivos. Destas, a mais recorrentes “dad inteligenc”, encontrada 204 vezes, em 13 documentos diferentes, não é um conceito em si, mas a associação de conceitos distintos que comumente aparecem em sequência. Nessa associação, o primeiro conceito, que termina sempre em "dados", é seguido por “inteligência artificial”. Por exemplo, “ciência de dados” e “inteligência artificial”, “banco de dados” e “inteligência artificial”, “estrutura de dados” e “inteligência artificial”.

As outras expressões entre as 20 mais frequentes nos cursos superiores que não estão no ranking dos anúncios de emprego são: “projet integr”, “estrut dad”, “mineraca dad”, “seri tempor”, “tom decisa”, “lingu natur”, “volum dad”, “red neur”, e “process lingu”. Aqui, nota-se que conceitos acadêmicos, como estrutura de dados, séries temporais, mineração de dados, processamento de linguagem natural, não são priorizados na busca por novos profissionais. Por outro lado, evidencia uma preocupação com uma formação mais fundamentada do futuro cientista de dados.

Para finalizar a análise dos termos mais frequentes dos conteúdos nos cursos superiores, a classificação das sequências do tipo 3-grama que aparecem em pelo menos cinco documentos distintos é apresentada na Tabela 18.

TABELA 18 – 3-GRAMA PARA CURSOS SUPERIORES

#	Termo	F	N	%
1	dad inteligenc artific	57	13	38,24%
2	cienc dad inteligenc	49	11	32,35%
3	process lingu natur	24	13	38,24%
4	tecnolog cienc dad	18	7	20,59%
5	analis explor dad	18	11	32,35%
6	grand volum dad	17	7	20,59%
7	banc dad relac	13	7	20,59%
8	calcul difer integr	12	6	17,65%
9	programaca orient objet	10	9	26,47%
10	introduca cienc dad	10	6	17,65%
11	dad big dat	9	8	23,53%
12	graduaca cienc dad	9	6	17,65%
13	dad projet integr	9	6	17,65%
14	aplic cienc dad	8	6	17,65%
15	analis seri tempor	8	6	17,65%
16	form cienc dad	8	5	14,71%
17	analis estatis dad	8	7	20,59%

FONTE: O autor (2022).

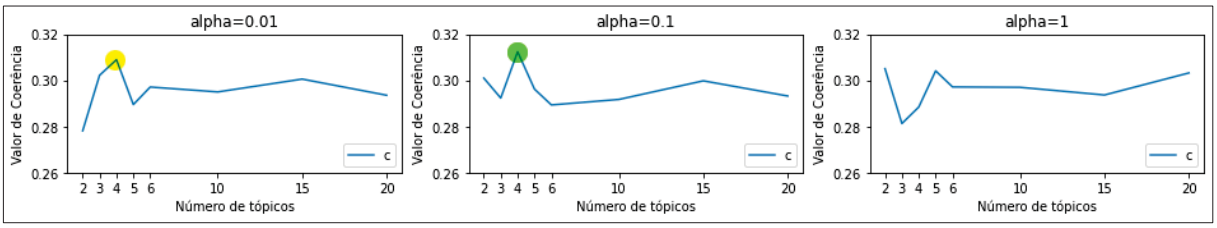
As junções de expressões terminadas em “dados” sucedidas pela expressão “inteligência artificial” formaram os dois trigramas mais frequentes, com 57 ocorrências, em 13 documentos (38,24%), e 49 ocorrências, em 11 documentos (32,35%), respectivamente. Em seguida, tem-se o conceito de processamento de linguagem natural, presente em 13 documentos (38,24%) e 24 ocorrências. Essa expressão, juntamente com “grande volume de dados” e “banco de dados (não) relacionais” formaram os três itens em comum com os anúncios de vagas de emprego, todos presentes em pelo menos 1/5 dos documentos.

Adicionalmente, as outras expressões presentes em mais de 20% do *corpus* são: “tecnolog cienc dad”, “analis explor dad”, “programaca orient objet”, “dad big dat” e “analis estatis dad”. A primeira expressão é referente à modalidade do curso superior em questão: “Curso Superior em Ciência de Dados” ou “Tecnólogo em Ciência de Dados”. As demais expressões se referem a conteúdo (análise exploratória de dados, programação orientada a objetos e análise estatística de dados) e a combinações sequenciais das expressões “dados” e “big data”.

6.3.2 Modelagem de tópicos

Bem como realizado para os anúncios das vagas, para a definição da quantidade de tópicos a serem identificados, foi verificada as coerências dos modelos mediante o cruzamento do número de tópicos (2, 3, 4, 5, 6, 10, 15, 20) com valores de alfa (0,01, 0,1 e 1). Os valores obtidos por estes cruzamentos são apresentados na Figura 48:

FIGURA 48 – VERIFICAÇÃO DOS VALORES DE COERÊNCIA PARA DIFERENTES MODELOS DOS CURSOS SUPERIORES

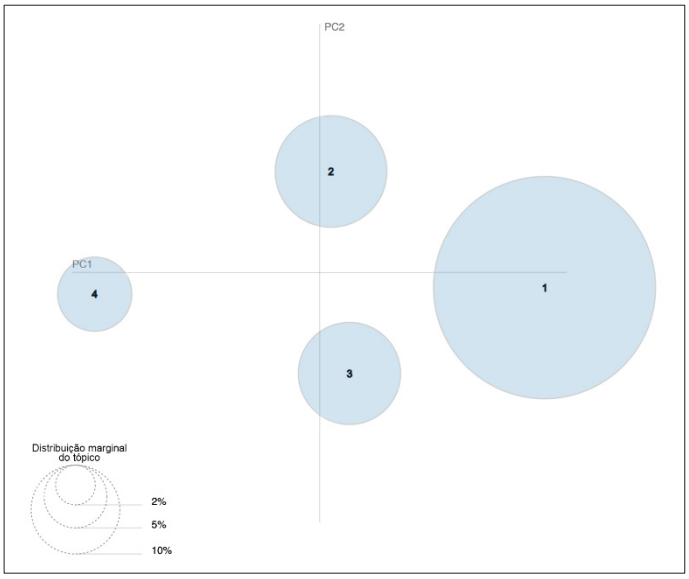


FONTE: O autor (2022).

Os valores de coerência encontrados para os modelos dos cursos superiores foram ainda mais baixos que aqueles obtidos para os anúncios. Uma das explicações plausíveis é o tamanho pequeno do *corpus*, com 34 documentos. O menor valor obtido foi 0,278, para a combinação de dois tópicos e 0,01 para o valor de alfa, enquanto o maior foi 0,312, para quatro tópicos e alfa, 0,10, ponto destacado em verde na Figura 48. Todavia, a visualização dos tópicos dessa combinação por meio da biblioteca LDAVis (SIEVERT; SHIRLEY, 2014), demonstrou haver sobreposição entre dois tópicos, indicando semelhança entre os mesmos.

Desta forma, optou-se pelo segundo maior valor de coerência (0,309), obtido pela combinação de quatro tópicos e valor de alfa igual a 0,01, cuja visualização está na Figura 49:

FIGURA 49 – VISUALIZAÇÃO DOS TÓPICOS COM A LDAVIS DOS CURSOS SUPERIORES

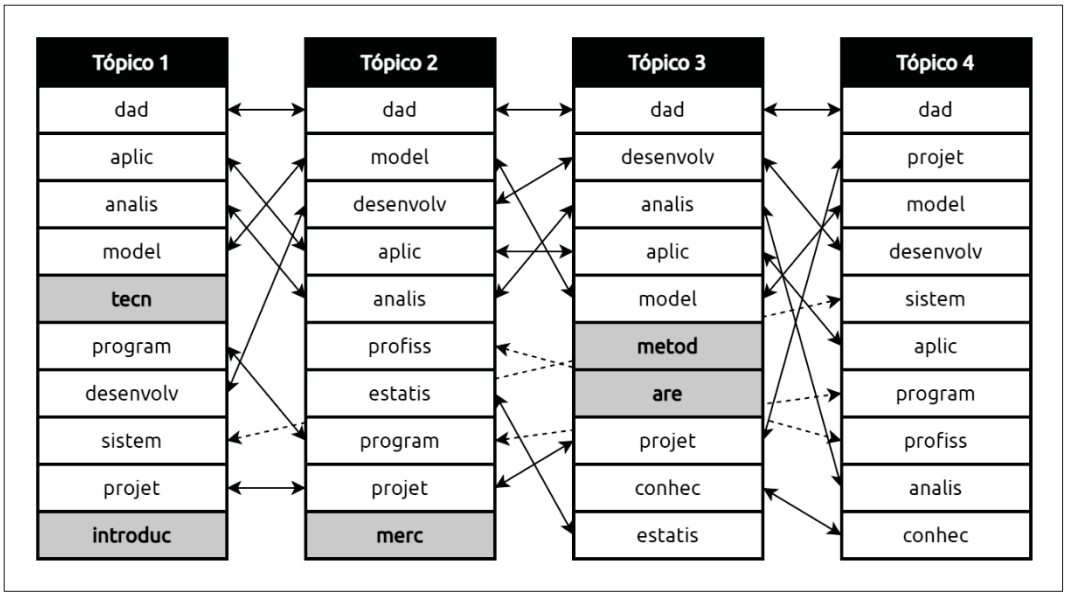


FONTE: O autor (2022).

Agora, com os quatro tópicos distantes uns dos outros, sem nenhuma sobreposição, e diante da distribuição marginal indicando a relevância dos tópicos obtidos para parte considerável dos documentos, tem-se um modelo adequado. No entanto, ao se verificar os 10 principais termos de cada tópico, obtém-se o mesmo resultado do procedimento LDA dos anúncios: muitos termos são compartilhados entre os tópicos, variando o peso e, consequentemente, a posição em que aparecem na lista. Desta forma, não é possível definir um assunto para cada tópico identificado.

No primeiro tópico, apenas os *stems* “tecn”, aqui referente a tecnologia e tecnólogo, e “introduc”, utilizado nos títulos de disciplinas introdutórias como “Introdução à Ciência de Dados”, “Introdução à Ciência da Computação”, “Introdução à Análise Numérica”, dentre outras. O *stem* “merc”, de mercado, é o único exclusivo do segundo tópico, enquanto “metod”, de metodologia, e “are”, referente a área, são específicos do terceiro tópico. Por fim, o último tópico identificado não tem nenhuma particularidade, compartilhando todos os termos com os demais tópicos. A Figura 50 apresenta as listas de termos que definem os tópicos.

FIGURA 50 – PRINCIPAIS TERMOS PARA CADA TÓPICO DA LDA PARA CURSOS SUPERIORES



FONTE: O autor (2022).

Em relação aos anúncios, novamente, o *stem* "dad" é o que tem maior peso em todos os tópicos. Por outro lado, os pesos dos demais termos se distinguiram, não havendo um padrão aparente, como ocorrido anteriormente. Observa-se, ainda, as ausências de termos referentes a *machine learning*, negócios, clientes, empresas e experiência. Há, agora, a ocorrência de termos relacionados a metodologia, introdução (a algum conceito), profissão, sistemas, aplicação e programação. De forma geral, a LDA mais uma vez não trouxe tópicos nitidamente distintos, mas revelou diferenças entre os *corpora* de anúncios de vaga e conteúdo de cursos superiores.

6.3.3 Análise de agrupamento

Para testar o agrupamento dos documentos dos cursos superiores, novamente, foi utilizado o algoritmo K-Means, por meio da ferramenta *online Clustering Workbench* (CARROT^2 CLUSTERING ENGINE, 2021). Os quatros agrupamentos identificados, número definido seguindo o padrão do procedimento LDA, foram:

1. EaD, Science, Digital, Interdisciplinar, Uninassau, *Work*, Informações, Extensão, Coding, Infra *(10 documentos)*
2. Modelos, Aplicações, Testes, Funções, Linear, Introdução, Integrador, Equações, Linguagem, POO *(9 documentos)*

3. Presencial, Integrador, Ciclo, Eletiva, Subtotal, Especializada, Marketing, Livre, *Management*, Security *(8 documentos)*
4. Obrigatória, Economia, CDN, Relacionais, Formação, Mundo, Comunidade, Alta, Qualificados, Avaliação *(5 documentos)*

Observa-se que, diferentemente do ocorrido nos anúncios, não há repetição entre os termos que caracterizam os *clusters*, com exceção ao termo "Integrador", presente no segundo e no terceiro agrupamento. No primeiro *cluster*, que contém 10 documentos (29,41% do *corpus*), destacam-se os termos EaD, *science*, digital, interdisciplinar, informações, dentre outros, incluindo o nome de uma IES (Uninassau). Nenhuma outra instituição foi relevante para determinação dos agrupamentos. O segundo *cluster*, com nove documentos (26,47%), é determinado principalmente por termos como aplicações, testes, funções, equações, linguagem, POO, conceitos relacionados à tecnologia da informação. No terceiro *cluster*, oito documentos (23,52%), misturam-se termos acadêmicos (presencial, integrador, ciclo, eletiva) com outros de gestão (*management*, marketing). Por fim, o quarto agrupamento, com cinco documentos (14,70%), traz conceitos acadêmicos (obrigatória, avaliação), de tecnologia (CND, relacionais) e diversos (mundo, formação, economia).

Percebe-se que não há grande discrepância entre os tamanhos dos *clusters* obtidos. Ou seja, não há o predomínio de um tipo específico de documento. É importante ressaltar, contudo, que o número de documentos é um fator limitante para a utilização de algoritmos de "clusterização". De qualquer forma, julga-se válida sua adoção para fins de comparação entre os conteúdos dos anúncios e dos cursos superiores. Adiante, o mesmo procedimento é aplicado aos cursos livres.

6.3.4 Considerações preliminares sobre cursos superiores em Ciência de Dados

Com o uso de técnicas de mineração de texto, buscou-se identificar padrões e extrair informações relevantes acerca dos conteúdos de cursos superiores sobre Ciência de Dados. Mais especificamente, foi coletado o conteúdo das páginas web e as matrizes curriculares dos cursos, bacharelados e tecnológicos, reconhecidos pelo Ministério da Educação no Brasil. Os resultados foram comparados com aqueles obtidos pelos mesmos procedimentos metodológicos anteriormente aplicados a anúncios de emprego para cientistas de dados.

Primeiramente, os resultados ressaltam a relevância da estatística para os cursos de Ciência de Dados. A recorrência de conceitos como análise de dados, modelo estatístico, séries temporais, bem como, *machine learning*, apontam a estatística como um dos pilares da educação superior da área. Por outro lado, os termos relacionados a tecnologia também estão presentes em todas as análises realizadas. Banco de dados, estruturas de dados, sistemas, programação, orientação a objetos foram alguns dos termos que ressaltam o papel da tecnologia e das Ciências da Computação para a Ciência de Dados.

Neste sentido, os conteúdos dos cursos superiores vão ao encontro de Kauermann e Seidl (2018) que definem a Ciência de Dados como a combinação das abordagens estatística e computacional, capaz de lidar com inúmeros problemas característicos do *Big Data*. Portanto, para esses autores, os currículos para a formação de futuros cientistas de dados devem se fundamentar primordialmente nestas duas disciplinas: Estatística e Ciências da Computação. Tosic e Beeston (2018) concordam com essa perspectiva e adicionam a Matemática e a Administração (*Business*) neste arranjo. No conteúdo analisado, enquanto a Matemática se fez presente por meio de termos como cálculo, diferencial e integral, a Administração é representada pela tomada de decisão, expressão recorrente no conteúdo analisado.

De maneira geral, o conteúdo dos cursos analisados apontou para competências evidenciadas por outros autores. Além das disciplinas Estatística, Ciências da Computação, Matemática e Administração, conceitos como mineração de dados, processamento de linguagem natural, redes neurais são recorrentemente associados às competências do cientista de dados (DEMCHENKO; COMMINIELLO; REALI, 2019; ROMERO; VENTURA, 2017; STODDER, 2018; WASHINGTON DURR, 2018) e aparecem entre os principais termos dos cursos superiores brasileiros.

Contudo, há requisitos tidos como essenciais ao profissional de dados que não foram corroborados pelos resultados. Anderson et al. (2014) afirma que além das competências supracitadas, os currículos para cientistas de dados devem contemplar competências focadas em comunicação, além de reflexão e discussão de questões éticas e sociais relacionadas à Ciência de Dados. Neste sentido, Tosic e Beeston (2018) reforçam o valor da comunicação interdisciplinar e intercultural, além da capacidade de falar em público, bem como da ética nos negócios. Kauermann e Seidl (2018) também destacam a ética dos dados, um guarda-chuva que contempla

questões legais, envolvendo privacidade e segurança, a comunicação e a colaboração como requisitos para os cientistas de dados.

A ética, por exemplo, apresenta 35 ocorrências em 16 documentos diferentes (47,00%), ocupando a 110ª posição entre os termos mais frequentes identificados. Sociedade é um conceito que vem logo atrás, na posição 141, com 30 ocorrências, e privacidade está apenas na 683ª, com cinco ocorrências em três documentos. Dentre outras competências citadas que não figuram entre as principais encontradas, está a comunicação (1056ª, com três ocorrências em dois documentos) e colaboração, que não foi identificada nos conteúdos analisados.

Em comparação aos resultados obtidos para os anúncios de vagas de emprego, é possível afirmar que há o alinhamento, ainda que diferenças sejam percebidas. No conteúdo das IES, percebe-se o aumento da relevância do termo ciência, menor utilização de termos em inglês e adoção de conceitos no lugar de ferramentas. Como citado, enquanto os cursos reforçam a habilidade de programar, sem explicitar a linguagem, as vagas são específicas, citando linguagens como Python, R, por exemplo. De qualquer forma, as listas 2-grama dos anúncios e dos cursos superiores compartilham 50% das 20 expressões mais frequentes, demonstrando o alinhamento entre os dois *corpora*.

De qualquer modo, bem como sugerido para o estudo dos anúncios, uma análise de conteúdo dos cursos traria resultados mais esclarecedores, visto que o contexto dos conceitos pode ser explorado. Quanto ao aumento do tamanho da amostra, considera-se uma sugestão inviável, visto que se trabalhou com a totalidade dos cursos de graduação registrados. Poder-se-ia solicitar mais material às coordenações dos cursos para ampliar o material analisado. Mesmo assim, entende-se que a mineração de texto trouxe um panorama sobre os cursos superiores em Ciência de Dados que pode fundamentar pesquisas futuras.

6.4 MINERAÇÃO DE TEXTO EM CURSOS LIVRES

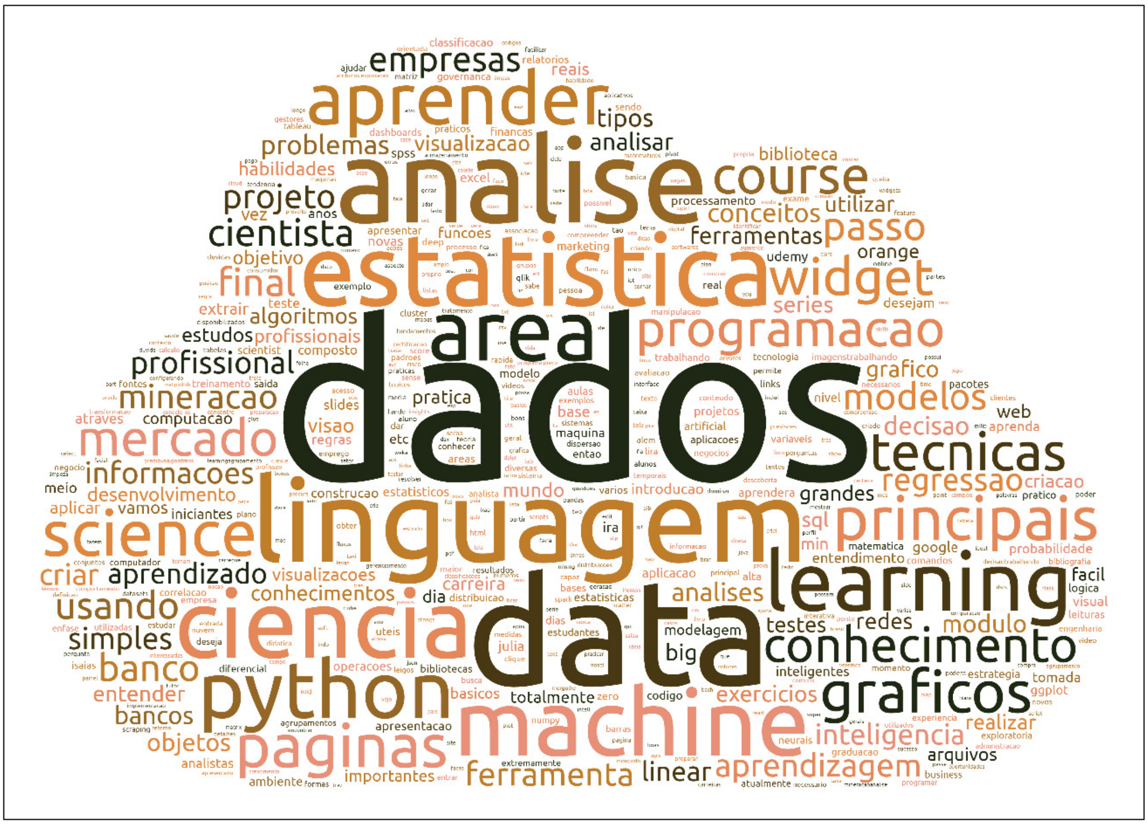
Visto que o processo de coleta e formação de dados do *corpus* referente aos 153 cursos livres foi detalhado na seção de metodologia, pouco há para se descrever sobre a amostra de cursos aqui. Acrescenta-se unicamente que os 142 cursos coletados da plataforma Udemy são oferecidos por 56 professores distintos. Desses, destacam-se os professores Odemir Depieri Jr., graduado e pós-graduado em

Tecnologia, com 22 cursos ofertados, Fernando Amaral, professor de Inteligência Artificial, com 12 cursos, e Marco Aurélio Dias de Oliveira, também da área de Tecnologia, com 11 cursos. Outros quatro professores possuem mais de seis cursos encontrados, quatro desses ofertam quatro cursos e sete ofertam dois cursos. Consequentemente, 38 professores possuem apenas um curso na amostra. Assim, verifica-se que poucos professores produzem muitos cursos, enquanto muitos professores produzem um único curso. A seguir, são explorados os termos mais frequentes dos cursos livres encontrados.

6.4.1 Termos mais frequentes

Pela nuvem de palavras (Figura 51), os principais termos dos documentos referentes aos cursos livres resgataram palavras que haviam se sobressaído nos anúncios, mas não nos cursos superiores. Dentre os termos nesta situação, destacam-se “conhecimento”, “machine”, “learning”, “área”, “mercado” e “python”. Nos cursos superiores, foi identificado mais termos genéricos, como “programação”, no lugar de linguagens específicas e a predominância de expressões em português, como “inteligência artificial”.

FIGURA 51 – NUVEM DE PALAVRAS COM OS TERMOS MAIS FREQUENTES PARA CURSOS LIVRES



FONTE: O autor (2022).

Como novidade, as páginas dos cursos livres trouxeram termos mais práticos na aprendizagem em Ciência de Dados, como “passo”, de passo a passo, “widget”, referente a componentes visuais para montagem de aplicações, “gráficos”, “principais técnicas”, dentre outros. Observa-se, neste caso, um caráter mais pragmático do que o adotado pelos cursos superiores. Para aprofundar análise do conteúdo dos cursos livres, a Tabela 19 apresenta os 20 termos mais frequentes 1-grama:

TABELA 19 – 1-GRAMA PARA CURSOS LIVRES

#	Termo	F	N	%
1	dad	1741	147	96,08%
2	analis	475	121	79,08%
3	aprend	433	125	81,70%
4	estatis	432	104	67,97%
5	dat	418	115	75,16%
6	lingu	349	93	60,78%
7	are	299	101	66,01%
8	graf	271	76	49,67%
9	cienc	271	75	49,02%
10	conhec	271	112	73,20%
11	machin	255	65	42,48%
12	python	228	65	42,48%
13	learning	228	59	38,56%
14	ferrament	219	82	53,59%
15	model	216	54	35,29%
16	princip	207	87	56,86%
17	estud	192	105	68,63%
18	profiss	186	86	56,21%
19	scienc	178	71	46,41%
20	banc	170	47	30,72%

FONTE: O autor (2022).
LEGENDA: F é quantidade de vezes que o termo aparece no *corpus*. N é o número de documentos nos quais o termo aparece. % é a porcentagem do número de documento em relação ao total do *corpus*.

Se o *stem* referente a dado é o mais frequente, presente em 147 dos 153 documentos (96,08%), repetindo os anúncios de emprego e os cursos superiores, o mesmo não ocorre para o segundo termo mais importante. Agora, “analis”, de análise de dados é o segundo com maior número de ocorrências, com 475 aparições em 121 documentos diferentes (79,08%). Por outro lado, se considerada a abrangência, “aprend”, de aprendizagem, é o segundo que aparece em mais documentos: 433 ocorrências em 125 cursos (81,70%).

A lista dos principais termos dos cursos livres apresenta sete termos que não estavam presentes entre os mais frequentes das análises anteriores: “aprend”, de aprendizagem, “lingu”, referente a linguagem, “graf”, de gráficos, “ferrament”, de ferramentas, “princip”, principais, “estud”, de estudantes, estudo e estudar, e, por fim, “scienc”, de ciência em inglês. Ademais, a lista compartilha 10 itens com a lista de anúncios de emprego e nove itens com os cursos superiores. Por fim, seis termos (“dad”, “analis”, “estatis”, “dat”, “are” e “model”) estão presentes nos três *corpora*.

TABELA 20 – 2-GRAMA PARA CURSOS LIVRES

#	Termo	F	N	%
1	machin learning	203	58	37,91%
2	cienc dad	200	58	37,91%
3	analis dad	193	88	57,52%
4	dat scienc	171	71	46,41%
5	banc dad	131	44	28,76%
6	cient dad	106	41	26,80%
7	mineraca dad	73	23	15,03%
8	lingu programaca	65	51	33,33%
9	visualizaca dad	62	43	28,10%
10	inteligenc artific	58	32	20,92%
11	big dat	57	23	15,03%
12	algoritm machin	51	25	16,34%
13	aprend maquin	46	26	16,99%
14	seri tempor	44	15	9,80%
15	bas dad	44	25	16,34%
16	anal dad	43	31	20,26%
17	pass pass	41	33	21,57%
18	regressa line	38	23	15,03%
19	analis estatis	36	28	18,30%
20	lingu jul	35	1	0,65%

FONTE: O autor (2022).

Já a lista com os termos 2-grama, apresentada na Tabela 20, traz a expressão *machine learning* como a mais recorrente, 203 aparições em 58 cursos (37,91%), bem como ocorrido para os anúncios de emprego. Em segundo lugar, “cienc dad”, referente a Ciência de Dados, aparece na mesma quantidade de documentos, porém com três ocorrências a menos. Dentre as 20 expressões mais recorrentes com dois *stems*, oito não aparecem nos rankings dos anúncios e dos cursos superiores: “visualizaca dad” (visualização de dados), “algoritm machin” (algoritmos de *machine learning*), “bas dad” (bases de dados), “anal dad” (análise de dados), “pass pass” (passo a passo), “regressa line” (regressão linear), “analis estatis” (análise estatística) e “lingu jul”, referente à linguagem de programação Julia.

Por outro lado, a lista compartilha 12 termos com os cursos superiores e 10 termos com os anúncios de emprego. Além disso, há 10 expressões que estão presentes entre os termos mais frequentes nos três conjuntos de documentos: “machin learning”, “cient dad”, “analis dad”, “cienc dad”, “banc dad”, “dat scienc”, “big dat”, “aprend maquin”, “inteligenc artific” e “lingu programaca”. Pode-se afirmar, então,

que esses são conceitos-chave tanto para vagas, quanto para cursos para cientistas de dados, seja do tipo superior, seja do tipo livre.

Finalmente, a última lista de expressões 3-grama, que aparecem em pelo menos cinco cursos, é apresentada na Tabela 21:

TABELA 21 – 3-GRAMA PARA CURSOS LIVRES

#	Termo	F	N	%
1	algoritm machin learning	34	25	16,34%
2	flux analis dad	30	16	10,46%
3	are mineraca dad	22	14	9,15%
4	analis explor dad	17	17	11,11%
5	ferrament analis dad	17	16	10,46%
6	analis dad simpl	16	16	10,46%
7	are cienc dad	16	13	8,50%
8	cienc dad machin	16	8	5,23%
9	col distribut pivot	16	16	10,46%
10	cours leig interess	16	16	10,46%
11	featur statistic dat	16	16	10,46%
12	widget col distribut	16	16	10,46%
13	widget featur statistic	16	16	10,46%

FONTE: O autor (2022).

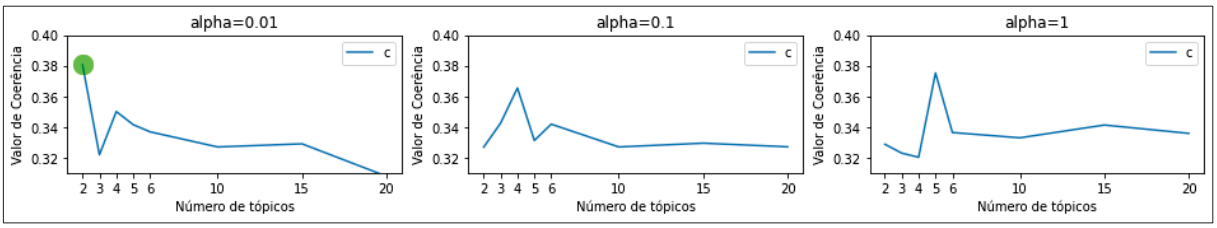
Novamente, o termo mais frequente está relacionado à aprendizagem de máquina: “algoritm machin learning” que aparece 34 vezes em 25 cursos diferentes (16,34%). Esta é a única expressão em comum com o ranking 3-grama dos anúncios de emprego. Outro conceito recorrente é Análise de Dados, presente em três dos 13 itens da lista: fluxo de análise de dados, análise de dados simples e análise de dados exploratória, sendo esta última o único item que também aparece na lista dos cursos superiores. Nota-se, também, a presença de expressões de língua inglesa, como *color*, *widget*, *pivot*, *feature*, *statistics* e *course*. Nestes casos, é possível identificar cursos com descrição parecida, que criam um padrão nos resultados, além de expressões da plataforma, como “Who this course is for” (Para quem é este curso), onde há a definição do público-alvo.

6.4.2 Modelagem de tópicos

Igualmente às análises de anúncios de emprego e cursos superiores, foram verificadas as coerências obtidas pelos modelos que combinam números de tópicos

(2, 3, 4, 5, 6, 10, 15, 20) e valores de alfa (0,01, 0,1 e 1), cujo resultado é apresentado na Figura 52:

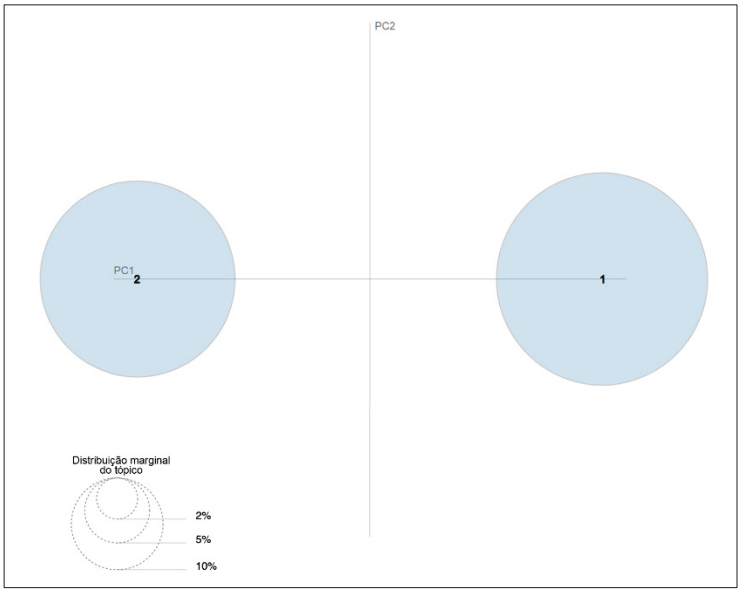
FIGURA 52 – VERIFICAÇÃO DOS VALORES DE COERÊNCIA PARA DIFERENTES MODELOS DOS CURSOS LIVRES



FONTE: O autor (2022).

Por outro lado, diferente dos procedimentos anteriores, o melhor modelo foi obtido (coerência = 0,380) para a combinação de apenas dois tópicos com um valor de alfa de 0,01. O segundo melhor modelo, cuja coerência foi 0,375, deu-se pela combinação de cinco tópicos com alfa definido em 1,00. Dessa forma, optou-se pelo modelo com dois tópicos, reforçado pela visualização da Figura 53:

FIGURA 53 – VISUALIZAÇÃO DOS TÓPICOS COM A LDAVIS DOS CURSOS LIVRES

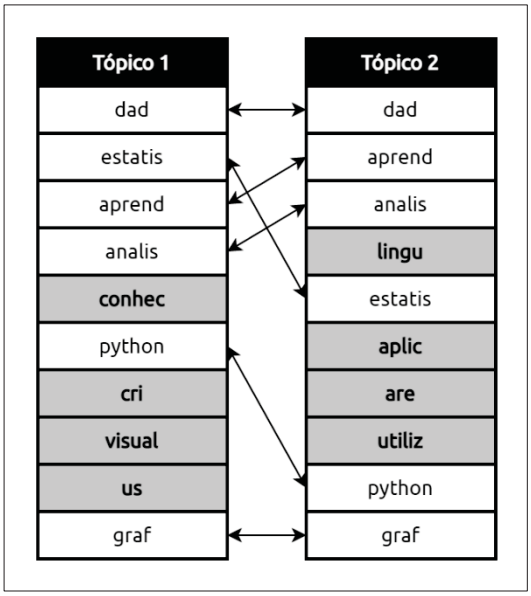


FONTE: O autor (2022).

Tem-se, então, dois tópicos distantes, sem qualquer sobreposição e são relevantes para grande parte dos documentos. Foram verificados, por consequência, os 10 termos com maior peso para a definição destes tópicos. Neste modelo, os dois tópicos compartilham seis termos entre si, onde o primeiro corresponde ao *stem* “dad”,

mesma situação dos anúncios e dos cursos superiores, e o último, é “graf”, referente a gráficos, visualização gráfica e expressões relacionadas. Os outros *stems* compartilhados são “estatis” e “python”, com maior peso para o primeiro tópico, e “aprend” e “analis”, mais importantes ao segundo tópico.

FIGURA 54 – PRINCIPAIS TERMOS PARA CADA TÓPICO DA LDA PARA CURSOS LIVRES



FONTE: O autor (2022).

Dentre os itens exclusivos a cada tópico, no primeiro destacam-se “conhec”, “cri”, de criação, “visual”, de visualização, e “us”, de uso e usar, principalmente. No segundo tópico, os itens específicos destacados são “lingu”, de linguagem (de programação), “aplic”, “are” e “utiliz”, de utilizando. Desta forma, novamente, não é distinta a separação dos dois tópicos. O que se percebe é a relevância do conhecimento e da visualização de dados no primeiro tópico, enquanto o segundo foca no desenvolvimento, por meio de aplicações/aplicativos e linguagens de programação. Por outro lado, por mais que sejam *stems* distintos, “us” e “utilz” funcionam como sinônimos ao passo que descrevem sobre o que os alunos aprenderão durante o curso.

6.4.3 Análise de agrupamento

Para a utilização do algoritmo K-Means, determinou-se dois agrupamentos seguindo o padrão adotado na LDA. Os dois conjuntos de termos obtidos correspondem a:

1. *Machine*, *Learning*, Python, Banco, Ciência, Carreira, Cientista, Processo, Tecnologia, Programação *(81 documentos)*
2. Widget, R, Modelos, Gráficos, Estatística, Series, Passo, Linear, Probabilidade, Distribuição *(72 documentos)*

Percebe-se, pela utilização da ferramenta *online Clustering Workbench* (CARROT^2 CLUSTERING ENGINE, 2021), que os dois grupos apresentam diferenças consideráveis entre si. A primeira é a oposição entre as linguagens de programação Python e R. Python é determinante para o primeiro grupo (81 documentos, 52,94%) que contém termos mais alinhados à tecnologia, como banco de dados, programação, além de *machine learning*. Ademais, ainda estão neste grupo, termos como ciência e cientista, além de carreira e processo, muito relacionado às atividades do cientista de dados. No segundo grupo (72 documentos, 47,06%) que contém a linguagem R, nota-se a predominância de termos relacionados à estatística, como séries temporais, regressão linear, modelos estatísticos, distribuição de dados, além de gráficos e *widgets*.

Dessa forma, diferentemente do ocorrido para os cursos superiores, a técnica de agrupamento demonstrou uma diferenciação mais clara entre os termos determinantes para a formação dos *clusters*. Ou seja, é possível afirmar que os cursos livres podem ser enquadrados em dois grupos distintos, com perfis de conteúdos diferentes.

6.4.4 Considerações preliminares sobre cursos livres em Ciência de Dados

Nesta seção, foram analisados, com técnicas de mineração de texto, o conteúdo de cursos livres na área de Ciência de Dados. Inicialmente, foram coletados 142 cursos do tipo MOOC na plataforma Udemy que tivessem “ciência de dados” ou

“data science” em seus títulos. Adicionalmente, foram adicionados ao *corpus* de análise, 11 cursos indicados pelos respondentes da pesquisa de levantamento.

Dentre as competências ressaltadas no conteúdo analisado, são reforçados conceitos relacionados a Estatística e Tecnologia. As principais expressões são *machine learning* (aprendizagem de máquina), análise (exploratória) de dados, séries temporais, banco de dados, mineração de dados, programação, entre outras, *big data*. Percebe-se uma consonância maior com os cursos superiores do que com os anúncios de emprego. A lista com as 20 expressões 2-grama mais frequentes, por exemplo, compartilha 12 termos com o conteúdo das graduações e 10 termos com os empregos.

Por outro lado, os cursos livres trazem expressões que não estão entre as mais comuns das análises dos anúncios de emprego e cursos superiores. A visualização de dados, recorrentemente citada na literatura (HARDIN *et al.*, 2015; TOSIC; BEESTON, 2018; WASHINGTON DURR, 2018), só foi ressaltada pelos resultados dos cursos livres, juntamente com a habilidade gráfica. As vagas de emprego e cursos superiores não trazem a visualização entre os requisitos primordiais ao cientista de dados. Este resultado, também encontrado por Kim e Lee (2016), é justificado pelos autores pela valorização de linguagens como R e Python, que possuem muitas ferramentas para visualizar dados, ofuscando o termo visualização em si.

Outra expressão que só apareceu nos cursos livres foi a linguagem de programação Julia, um projeto de *open source*, focado em alto desempenho e que tem se popularizado na comunidade de Ciência de Dados, por proporcionar um ecossistema com ferramentas de visualização, manipulação de dados, *machine learning*, dentre outros benefícios (JULIA PROJECT, 2021). Embora tenha sido citada em apenas um documento, a presença deste curso é um indicativo que os cursos livres se adaptam mais rapidamente às novidades tecnológicas.

De maneira geral, os cursos livres trouxeram um aspecto mais pragmático, ressaltando o processo de aprendizagem dos profissionais, com ênfase em tutoriais (passo a passo), ferramentas e soluções finais. Novamente, para pesquisas futuras, outros procedimentos metodológicos, como análise de conteúdo, poderiam extrair novas informações e relações do conteúdo coletado. Mesmo assim, a mineração de texto permitiu identificar padrões e comparar os cursos livres com cursos de graduação, bem como com anúncios de empregos para cientistas de dados.

6.5 SÍNTESE DO CAPÍTULO

Para identificar as competências necessárias ao cientista de dados no Brasil, esta tese utilizou múltiplas fontes de dados. Inicialmente, tem-se a pesquisa de levantamento derivada de pesquisas anteriores (HARRIS; MURPHY; VAISMAN, 2013-; HAYES, 2020; KAGGLE, 2020) e reforçada pela identificação de competências encontradas na literatura. Com esta etapa da pesquisa, foi possível identificar as principais características dos profissionais de dados atuantes no Brasil como gênero, idade, localidade, formação educacional, tempo de experiência e remuneração. Além disso, a aplicação de questionário também permitiu conhecer sobre as organizações que contratam cientistas de dados, revelando o setor de atuação, número de colaboradores e tamanho das equipes de dados.

Embora estas informações sejam parte importante da pesquisa, a investigação sobre as competências, foco principal desta tese, também foi priorizada no questionário. Com a investigação das competências, foi possível verificar a adesão dos itens elencados como importantes na literatura com a realidade dos profissionais. Mais que isso, permitiu o desenvolvimento de um modelo estatístico de competências para a Ciência de Dados (Figura 42), por meio da aplicação do método Análise Fatorial Confirmatória.

Outras duas fontes de dados, anúncios de vagas para cientistas de dados e cursos livres para Ciência de Dados, possuem em comum a forma com que foram obtidos: a raspagem de dados da web. Primeiramente, a análise dos anúncios de vagas de emprego, obtidos a partir de seis *websites* específicos, possibilitou elencar os requisitos mais solicitados pelas empresas que buscam contratar cientistas de dados, além de buscar por evidências de tópicos ou características de agrupamento. Esta parte da pesquisa resultou em um artigo publicado na Revista Ibérica de Sistemas e Tecnologias de Informação (RISTI) sob o título “Requisitos para a Ciência de Dados: analisando anúncios de vagas de emprego com mineração de texto” (GUIMARÃES; MENDES JÚNIOR; FREITAS, 2022). O manuscrito dessa publicação foi submetido em fevereiro de 2022, aceito cerca de três meses depois, em maio.

O mesmo protocolo foi aplicado para o conteúdo de cursos livres, raspado da plataforma Udemy ou coletado manualmente a partir de indicações dos respondentes da pesquisa de levantamento; e para o conteúdo de cursos superiores reconhecidos

pelo MEC, também coletado manualmente. Da mesma forma, procurou-se pelas competências mais recorrentes, por tópicos e por agrupamentos. No entanto, as relações entre todas as análises executadas durante a pesquisa são exploradas no próximo capítulo que faz um cruzamento entre os resultados obtidos.

7 TRIANGULAÇÃO DOS RESULTADOS E SÍNTESE CONCLUSIVA

Neste capítulo, são retomados os principais resultados dos procedimentos adotados na tese, da pesquisa de levantamento à mineração de texto dos cursos livres, passando pelas análises dos anúncios de vagas e cursos superiores. Primeiramente, são repassadas as principais características da amostra da pesquisa, seguidas pela síntese das competências desses profissionais de dados. Posteriormente, esses resultados são confrontados com os resultados das análises textuais.

Esta pesquisa confirma a Ciência de Dados como um campo jovem, composto por trabalhadores jovens, em sua grande maioria homens, com pouco tempo de experiência na área, ainda que altamente qualificados, porém com raros profissionais com diploma de cientistas de dados, propriamente dito (BURTCH WORKS, 2019; FIGURE EIGHT, 2018; KAGGLE, 2021). No Brasil, um dos possíveis motivos para este baixo número de cientista de dados de formação é que todos os cursos registrados no portal e-MEC iniciaram suas atividades a partir de 2018, com exceção do curso da USP que foi iniciado em 2009 com outro nome.

A amostra obtida na pesquisa de levantamento é composta por profissionais de todas as regiões brasileiras, porém com uma evidente predominância da região Sudeste (47,58%), seguida pelas regiões Nordeste (21,15%) e Sul (19,38%). Todavia, uma vez que não é possível estimar a população total de cientistas de dados no Brasil, o que impossibilita uma estratificação precisa, considera-se que todas as regiões do país são participantes da pesquisa.

Sobre a educação formal, a pesquisa de levantamento indica que cursar uma graduação é comum a quase toda a amostra, visto que apenas um respondente não possui e nem está matriculado no ensino superior. Porém, apenas seis respondentes cursaram graduação em Ciência de Dados, reforçando cursos tradicionais, principalmente Estatística e Ciência da Computação, como principais fornecedores de profissionais de dados ao mercado (BAŠKARADA; KORONIOS, 2017; DONOHO, 2017; HARRIS; MURPHY; VAISMAN, 2013-). Por outro lado, as porcentagens encontradas de profissionais com mestrado e doutorado são inferiores a outras pesquisas (BURTCH WORKS, 2019; KAGGLE, 2021).

Em se tratando do ambiente organizacional, a pesquisa indica que a Ciência de Dados é mais comum em grandes empresas, uma vez que metade (49,78%) dos

respondentes trabalham em companhias com mais de 249 colaboradores. Considerando a faixa seguinte, tem-se que 60,35% das empresas participantes da pesquisa de levantamento contam com 100 colaboradores ou mais. Este resultado contrasta com relatório da Kaggle (2021), onde o porte predominante é de empresas com até 49 colaboradores, equivalente a 34,90% da amostra. Há, portanto, indícios de que fora do Brasil, empresas menores já estão mais adiantadas quanto à exploração de dados. No Brasil, verifica-se ainda que as equipes de profissionais que atuam como cientista de dados, ou funções análogas, tendem a não ultrapassam 15 integrantes, visto que 81,52% das respostas ficam dentro desse limite. Dessa forma, tem-se que o tamanho da equipe de dados não está relacionada ao porte da organização.

Ainda sobre as características do mercado de trabalho, a pesquisa indicou que não houve predomínio entre as faixas salariais apresentadas, demonstrando a variabilidade da remuneração dos profissionais da área. Quanto ao quesito remuneração, foi verificado que os salários pagos aos profissionais brasileiros são consideravelmente menores que aqueles pagos em outros países (BURTCH WORKS, 2021). Este resultado vai ao encontro do relatório da Kaggle (2021) que indica o Brasil como o quarto país com melhor média salarial para cientista de dados, atrás dos Estados Unidos, Alemanha e Japão.

Expondo a abrangência da Ciência de Dados, além de seu potencial de progresso nas mais distintas áreas da sociedade (GROSSI *et al.*, 2021), a pesquisa contou com participações de profissionais de mais de 10 segmentos. Se o segmento de tecnologia foi o maior com 28 respostas, também houve participação de organizações do setor público, energético, educação, saúde, marketing, dentre outros. Neste sentido, a pesquisa aponta que o potencial dos dados consiga os mais diversos setores (BOWNE-ANDERSON, 2018; TOSIC; BEESTON, 2018).

Sobre os grupos de competências contemplados no questionário, os resultados destacam os grupos de Tecnologia e Análise de Dados como aqueles fundamentais aos cientistas de dados. Enquanto o grupo de questões sobre Entendimento de Negócios, com média geral de 5,375, é tido como aquele em que os profissionais possuem menos proficiência. Questões sobre *compliance*, governança, LGPD e planejamento financeiro para um projeto de dados, que justas representam metade do grupo de negócio, ficaram com médias inferiores a cinco. Para Tecnologia e Análise de Dados, ambos com 12 variáveis cada, apenas duas questões em cada ficaram com

médias inferiores a 5,00. Em Tecnologia, essas questões foram sobre a administração de sistemas de informações e o processamento de dados distribuídos. Em Análise de Dados, as questões com menores valores foram sobre otimização e modelagem de grafos. Simplificadamente, estas questões indicam as competências com menores níveis de proficiência da amostra.

Em contrapartida, as questões do grupo de Competências Socioculturais apresentaram os melhores resultados, especialmente aqueles itens medidos no âmbito organizacional. Foi visto que colaboração, criatividade, curiosidade, ética, proteção a dados sensíveis e preocupação com análises enviesadas correspondem aos itens com melhores médias, todas acima de 8,00. De maneira geral, tem-se que os respondentes foram mais benevolentes em relação às habilidades interpessoais, ou *soft skills*.

No Quadro 16, é apresentado o cruzamento das variáveis do questionário original com a primeira ocorrência de um termo (*stem*) ou expressão correspondente àquela competência. Dessa forma, tem-se a triangulação entre os resultados da pesquisa de levantamento com os resultados da mineração de texto (anúncios de empregos, cursos superiores e cursos livres). Ao lado de cada termo ou expressão, há uma célula com um número e uma cor. O número indica a posição do termo ou expressão no *ranking* de termos mais recorrentes do *corpus* específico. E a cor indica se refere a uma escala, onde quanto mais próximo do topo, mais verde será a célula; quanto mais longe do topo, mais vermelha ela será.

QUADRO 16 - TRIANGULAÇÃO DOS RESULTADOS

		Anúncios		Cursos superiores		Cursos livres	
Fator	**Variável**	**#**	**n-gram**	**#**	**n-gram**	**#**	**n-gram**
Tecnologia	tecAlgoritmos	50	algoritm	25	algoritm	28	algoritm
	tecBancoDeDados	6	banc dad	3	banc dad	5	banc dad
	tecDadosSemiEstruturados	46	dad estrutur	50	dad estrutur	-	
	tecDadosNaoEstruturados	-		-		-	
	tecDesenvSoftware	96	desenvolv softw	41	engenh softw	218	softw
	tecEngDeDados	53	engenh dad	-		101	engenh dad
	tecGestaoDeDados	352	gesta dad	-		-	
	tecHabHacking	-		-		-	
	tecAIML	1	machin learning	2	inteligenc artific	1	machin learning
	tecProgramacao	45	programaca	9	programaca	32	programaca
	tecSistemas	181	sistem informaca	49	sistem informaca	-	
	tecDadosDistribuidos	-		-		-	
Análise de dados	anAnaliseDeDados	3	analis dad	4	analis dad	3	analis dad
	anAnalisesPreditivas	12	model predi	-		-	
	anEstatistica	8	estatis	8	estatis	4	estatis
	anMatematica	62	matema	49	matema	115	matema
	anMineracaoDeDados	49	mineraca dad	10	mineraca dad	7	mineraca dad
	anModelagemGrafos	-		63	graf	-	
	anOtimizacao	153	otimizaca	119	otimizaca	681	otimizaca
	anNLP	99	lingu natur	18	lingu natur	267	lingu natur
	anFormQuestoes	693	questo	316	questo	-	
	anMedotoCientifico	941	metod cientif	127	cientif	394	cientif
	anVisualizacao	917	visual	922	visual	92	visual
	anRegressao	197	regressa	84	regressa	51	regressa
Entendimento de Negócios	enConhecimento	-		293	domini	-	
	enDesenvolvimentoDeP	134	desenvolv projet	-		-	
	enGestaoDeProjetos	-		-		-	
	enNegocios	7	negoci	44	negoci	85	negoci
	enFinanceiro*	-		48	financ	-	
	enGovernanca*	503	governanc	318	governanc	286	governanc
	enCompliance*	829	regul	1155	regulaca	2709	regul
	enLGPD*	-		1375	lgpd	-	
Sociocultural	socInterdisciplinari	-		169	interdisciplin	1973	interdisciplin
	socComunicacao	119	comunicaca	168	comunicaca	809	comunicaca
	socLidTec	361	lideranc	855	lideranc	-	
	socLidEst	84	estrateg	56	estrateg	131	estrateg
	socSolucaoProblemas	23	problem negoci	61	soluca problem	46	resolv problem
	socColaboracao	79	colabor	538	colabor	1533	colabor
	socCriatividade	583	criat	872	criat	-	
	socCuriosidade	910	curios	-		1445	curios
	socEtica	683	etic	110	etic	-	
	socDadosSensiveis*	-		705	segur	749	seguranc
	socBias*	-		1302	vie	3863	envies

FONTE: O autor (2022).

Em uma visão geral do quadro, é verificada uma maior concentração de elementos em cores verdes nos grupos Tecnologia e Análise de Dados. Ou seja, as competências pertencentes a esses fatores correspondem àquelas mais valorizadas tanto para as organizações que procuram cientistas de dados, quanto para as iniciativas educacionais. Todavia, mesmo dentro desses grupos mais requisitados, há questões presentes no questionário que não foram identificadas nas análises de mineração de texto.

A competência em lidar com dados não estruturados, fundamental para lidar com a variabilidade do *Big Data* (DAMA INTERNATIONAL TECHNICS, 2017; HASSANI *et al.*, 2020; IBM, 2020), não pode ser verificada na mineração de texto, visto que "não" é tida como uma *stopword*, logo é retirada da análise. O mesmo ocorreu para dados semiestruturados, cujo prefixo "semi" é retirado na mineração. Ou seja, as expressões "não estruturado" e "semiestruturado" são tratadas como "estruturado" e, por isso, não aparecem nos *rankings* de termos mais frequentes. Por outro lado, trabalhar como dados estruturados está entre os 50 termos mais comuns dos anúncios de vagas e dos cursos superiores, mas não é uma competência explicitada nos cursos livres.

Habilidades em *hacking*, ou seja, autonomia para encontrar soluções não convencionais para lidar com problemas de dados (CONWAY, 2010; PARKS, 2017), foi outro item presente no questionário que não foi identificado na mineração de texto. Uma das explicações possíveis seria a conotação negativa e a informalidade relacionadas ao termo *hacker*.

A Gestão de Dados, essencial para a integração de múltiplas fontes de dados, de forma confiável, segura e eficiente (DAMA INTERNATIONAL TECHNICS, 2017; DEMCHENKO *et al.*, 2016; HALWANI *et al.*, 2021), só foi identificada na 352ª posição nos anúncios de emprego, ficando de fora do conteúdos educacionais. Por fim, a capacidade em lidar com infraestrutura de *Big Data*, envolvendo computação em nuvem e processamento distribuído (CAO, 2017; SALTZ; GRADY, 2017), também não foi identificada em nenhum dos *corpora* analisados. Assim, verifica-se que trabalhar com sistemas distribuídos, competência que apresentou a menor média (4,087) na pesquisa de levantamento, é uma competência rara entre os profissionais e, mesmo assim, não é priorizada nos anúncios, tão pouco nos cursos superiores e livres.

Em Tecnologia, ainda é possível duas das principais competências dentre as ressaltadas pela mineração de texto. A primeira delas é a capacidade de desenvolver

sistemas de inteligência artificial, em especial do tipo *machine learning*, tendo ficado na primeira posição dos anúncios e dos cursos livres, e na segunda posição dos cursos superiores. Vale ressaltar que no modelo proposto na seção 6.1.2, esta competência mostrou maior afinidade com as variáveis do grupo Análise de Dados, tendo sido deslocada para este fator na versão final do modelo. A outra competência dentre as principais encontradas, é a capacidade de administrar bancos de dados, tendo sido classificada na sexta posição dos anúncios, terceira posição dos cursos superiores e quinta posição dos cursos livres. Por fim, o grupo Tecnologia também reforça a capacidade de programar dos cientistas de dados. No entanto, se nos cursos superiores essa competência está entre as 10 mais recorrentes, ocupando a nona posição, nos cursos livres ela ocupa apenas a 35ª posição e nos anúncios de emprego, somente a 45ª.

O grupo Análise de Dados engloba as demais competências com melhor posicionamento dentre os termos mais frequentes. A principal delas, a competência "genérica" em análise de dados que ocupou a terceira posição, nos anúncios e nos cursos livres, e a quarta posição, nos cursos superiores. Em seguida, está a competência em estatística na oitava posição, dos anúncios e cursos superiores, e quarta posição nos cursos livres. Dentre as outras questões com ocorrência de cores verdes, ou seja, dentre os principais termos mais recorrentes, está a mineração de dados que figura entre as 10 posições dos conteúdos educacionais, mas está apenas na 49ª posição dos anúncios de emprego. Além disso, a capacidade em processar linguagem natural aparece em verde para os cursos superiores, ocupando a 18ª posição, mas em amarelo nos anúncios (99ª posição) e nos cursos livres (267ª). Isso demonstra que essa competência é mais valorizada na educação formal do que para quem contrata um cientista de dados ou oferece um curso livre.

Dentre as questões presentes na seção de Análise de Dados que não foram tão recorrentes na mineração, tem-se a capacidade de realizar análises preditivas. Esta competência não foi identificada nos cursos superiores e livres, mas apareceu na 12ª posição nos documentos dos anúncios de emprego, entre os termos com maior destaque. A modelagem em grafos é outra competência presente em apenas uma das coleções de documentos. A 63ª posição dentre os termos mais frequentes dos cursos superiores indica que o conceito de estruturas em grafos é prioritariamente trabalhado na educação formal, sendo pouco requerido em anúncios ou abordado em cursos livres.

Ainda em relação às competências em análise de dados que compuseram o questionário, mas não se sobressaíram nos procedimentos de mineração, tem-se a capacidade de formular questões relativas a dados e o domínio do método científico. A formulação de questões, presente entre os termos dos anúncios (693ª posição) e dos cursos superiores (316ª posição), não foi identificada nos cursos livres. De forma similar, ainda que menção ao método científico ocorre nos três conjuntos de documentos, em nenhum deles há destaque para essa competência. Nos cursos superiores, o domínio método científico apresenta seu melhor resultado, na 127ª posição. O oposto acontece para os anúncios de emprego, onde a referida competência ocupa a 941ª posição.

Diferentemente do ocorrido para os grupos anteriores, nos cruzamentos das questões de Entendimento de Negócios com os resultados da mineração de texto, somente uma variável atingiu a cor verde para os três *corpora*: desenvolvimento de novos negócios ou capacidade de empreender. Essa competência foi identificada na sétima posição dos anúncios de emprego por meio do *stem* "negoc" onde era utilizado em meio a sentenças que contendo expressões como "conhecimento de negócios", "dados relevantes para o negócio", "solucionar problemas de negócios", "identificar oportunidades de negócios", dentre outras. Percebe-se, assim, uma consonância com outra questão de negócios, o conhecimento de domínio. Todavia, essa segunda competência não foi expressa dessa forma nos anúncios de emprego e nem nos cursos livres, aparecendo apenas nos cursos superiores na 293ª posição.

Dentre outras competências de negócios que não ganharam destaque na mineração, estão o desenvolvimento de novos projetos, que apareceu apenas nos anúncios na 134ª posição, a gestão de projetos, ausente dos três conjuntos de documentos, e o planejamento financeiro de projetos relacionados a dados, que só foi identificado nos cursos superiores na 48ª posição. Além disso, foi verificado que as competências em questões de governança, *compliance* e domínio da LGPD não estão entre os atributos requeridos pelo mercado, tão pouco são valorizadas pelas instituições educacionais, seja formal ou informal. Todos esses itens aparecem no Quadro 16 com cores laranja ou vermelho, indicando que estão entre as últimas posições dos *rankings* analisados. A LGPD, por exemplo, não é citada nos anúncios e nos cursos livres, aparecendo apenas nos cursos superiores e na 1375ª posição.

O último grupo de cruzamento contemplou as competências socioculturais, sob a perspectiva do indivíduo e da organização. Assim como aconteceu para as variáveis

de negócios, apenas uma competência do grupo sociocultural figurou entre os principais termos da mineração de texto: a capacidade de solucionar problemas relacionados a dados. Ainda assim, essa competência esteve na 23ª posição dos anúncios de emprego, 46ª posição dos cursos livres e 61ª posição dos cursos superiores. Embora esse achado corrobore com a categorização do cientista de dados como um resolvedor de problemas (RAWLINGS-GOSS, 2019; WASHINGTON DURR, 2018), as atribuições desse profissional vão além desse rótulo. Para solucionar problemas, o cientista de dados deve compreender as oportunidades antes da solução ser implementada, precisa entender de análise de riscos e engenharia de sistemas, além de comunicar claramente suas descobertas (IBM, 2020).

Dentre as outras competências do grupo, há o destaque para o conhecimento interdisciplinar, presente no conteúdo educacional, mas ausente dos anúncios de emprego. Dessa forma, por mais que a interdisciplinaridade seja inerente à atividade do cientista de dados (ANDERSON; MCGUFFEE; UMINSKY, 2014; GROSSI *et al.*, 2021; HARRIS; MURPHY; VAISMAN, 2013-), esse quesito não é mandatório para as organizações contratantes.

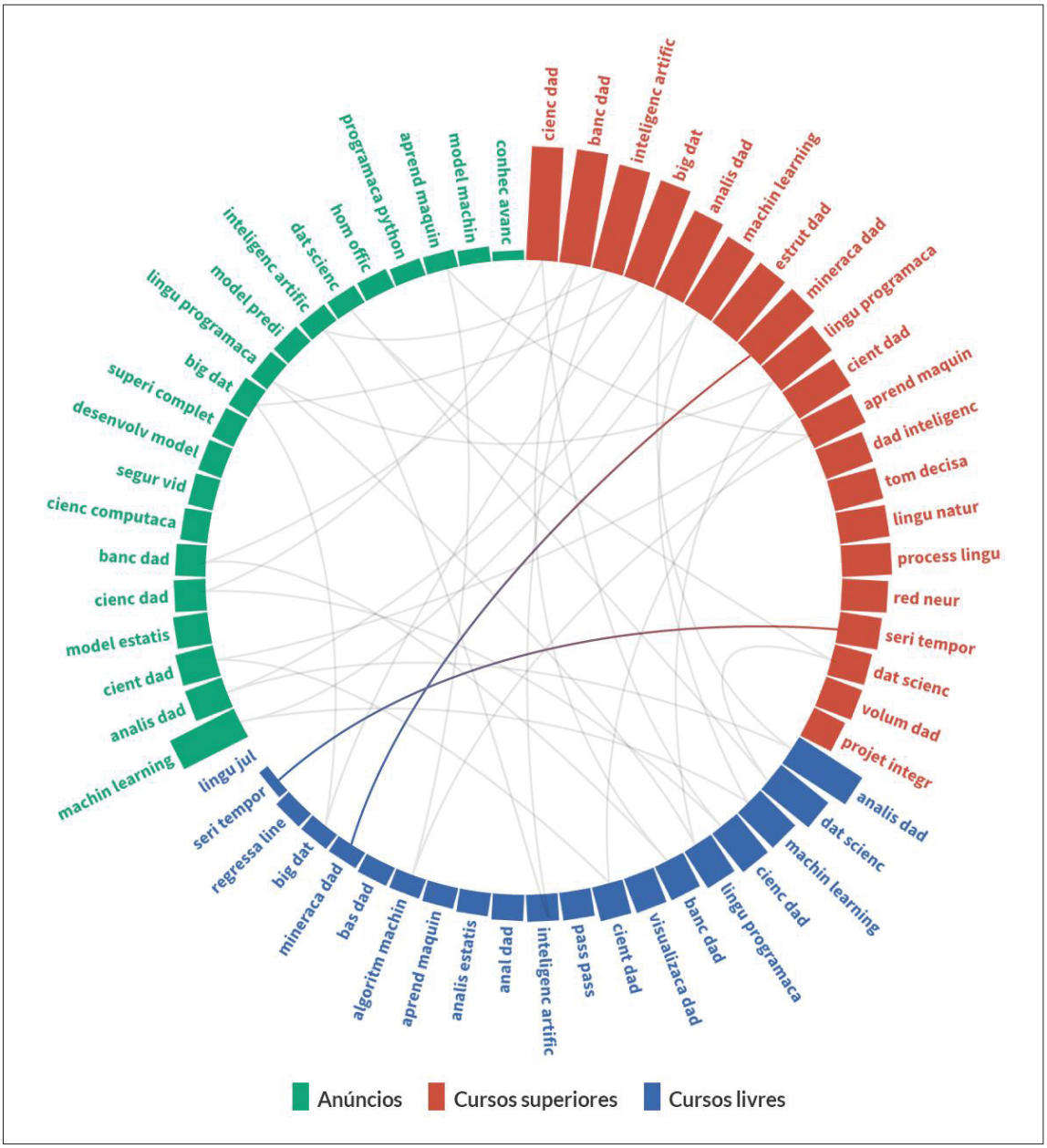
Mesmo a comunicação, uma das habilidades interpessoais mais valorizadas nos processos de recrutamento (MALINSKY, 2022), não obteve destaque nos procedimentos de mineração de texto. Neste sentido, a colaboração, a criatividade e a curiosidade também apresentaram desempenho sem destaque. Questões relacionadas à ética dos dados, que compõem uma discussão aberta e que vem ganhando relevância (FORGÓ *et al.*, 2021; LOUKIDES; MASON; PATIL, 2018), também foram relegadas, segundo o resultado da mineração. Tratamento de dados sensíveis e precaução em relação a análises enviesadas, itens marcados com asteriscos no Quadro 16 por terem sido indicados no pré-teste, não foram identificados nos anúncios de emprego. Neste sentido, a ética não está presente nos cursos livres e apresentou o melhor desempenho nos cursos superiores, figurando na 110ª posição dentre os termos mais frequentes.

De certa maneira, o cruzamento entre a pesquisa de levantamento e as análises de texto reforça o modelo de competência apresentado por meio da análise confirmatória (seção 6.1.2,). Notoriamente, esse cruzamento evidencia as competências em tecnologia e análise de dados como os alicerces da profissão de cientista de dados. Ao mesmo tempo, demonstra que competências em negócios, apesar de valorizadas pela literatura e pelas organizações, ainda não são tão

desenvolvidas por esse profissional. Por fim, as competências socioculturais, igualmente tidas como essenciais para o cientista de dados, não são formalmente vinculadas a esse profissional, especialmente aquelas destacadas em vermelho no Quadro 16. Como resultado, são raras nas descrições de vagas e nos conteúdos educacionais, além de terem sido excluídas do modelo estatístico de competências.

Para concluir a triangulação entre os resultados, são apresentadas na Figura 55 as relações entre os 20 termos mais frequentes, do tipo 2-grama, encontrados nos anúncios de empregos, nos cursos de nível superior e nos cursos livres.

FIGURA 55 – RELAÇÕES ENTRE ANÚNCIOS, CURSOS SUPERIORES E LIVRES



FONTE: O autor (2022).

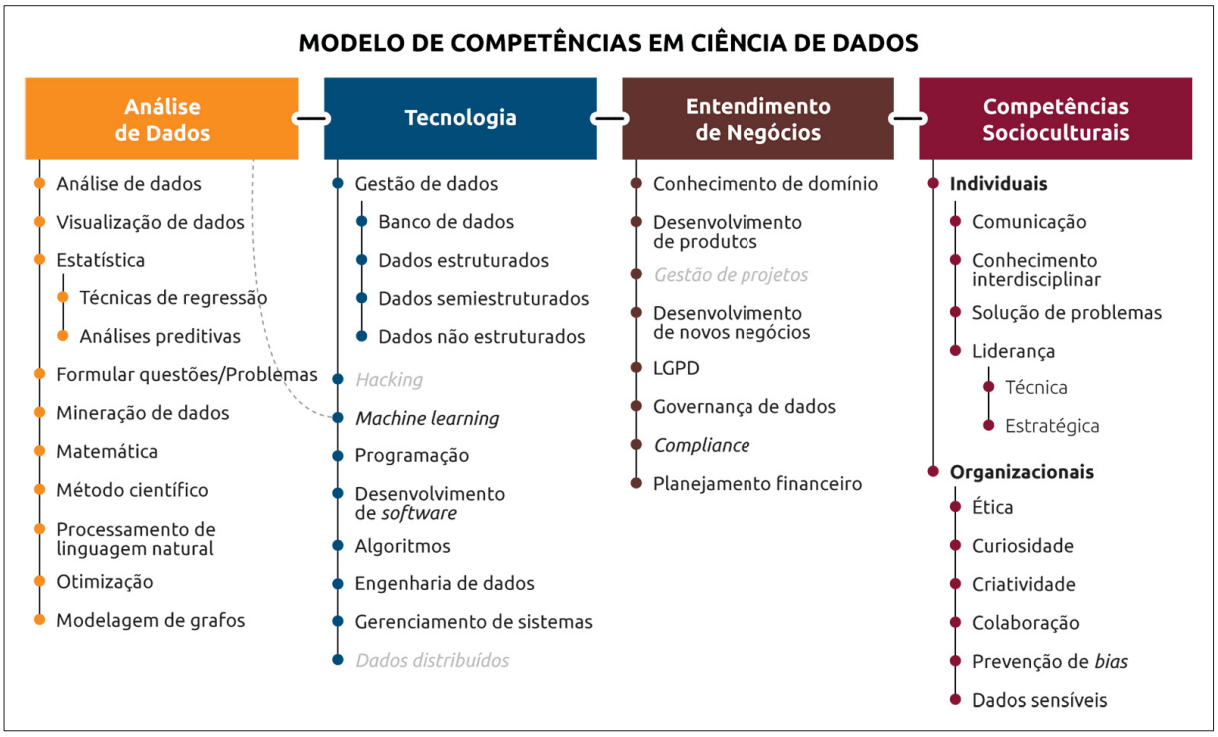
É perceptível que os três conjuntos de documentos compartilham muitos termos entre as 20 expressões n-grama mais frequentes, visto que as linhas em cinza indicam os elementos comuns aos três grupos (anúncios, cursos superiores e cursos livres). Na imagem, há duas expressões que são comuns a apenas dois conjuntos: a) mineração de dados, que ocupa a oitava posição dos cursos superiores e a 16ª posição nos cursos livres; b) séries temporais, 17ª e 19ª posições nos cursos superiores e livres, respectivamente. Há o indicativo, portanto, de que estas expressões não são tão comuns entre os requisitos exigidos na hora de se contratar um cientista de dados, ainda que valorizadas pelos conteúdos organizacionais.

Uma vez que o conteúdo compartilhado é maioria entre os 20 mais frequentes analisados, destaca-se as diferenças entre os conjuntos. O que diferencia o conjunto de anúncios do conteúdo educacional ultrapassa as expressões relacionadas às condições de trabalho, como “seguro de vida” e “home office”, ou relacionadas aos requisitos da vaga, como “superior completo”, “ciência da computação” e “conhecimento avançado”. Para os anúncios, itens como “desenvolvimento de modelo (estatístico, preditivo, de machine learning)” e “programação Python” são elementos que determinam o que é mais requerido para um candidato a cientista de dados.

Nos cursos superiores, os itens característicos estão relacionados a expressões técnicas como “processamento de linguagem natural”, “estrutura de dados” e “redes neurais”. Além disso, há destaque para “volume de dados”, “tomada de decisão”, além da expressão acadêmica “projeto integrado”. O conteúdo dos cursos livres é o único que traz a visualização de dados entre os termos mais recorrentes. Além disso, novamente, tem-se a expressão “passo a passo”, destacando o caráter didático do conteúdo, e conceitos como aprendizado de máquina, big data e análise estatística.

Em suma, por mais que o conteúdo tenha pontos em comum, é possível identificar as particularidades e relacionar ao objetivo específico de cada conjunto de documentos. Além disso, com as relações dos resultados evidenciadas, há subsídios suficientes para a formulação do modelo das competências necessárias para se trabalhar com Ciência de Dados no Brasil, apresentado na Figura 56:

FIGURA 56 – MODELO DE COMPETÊNCIA EM CIÊNCIA DE DADOS



FONTE: O autor (2022).

Para a montagem do modelo, foram tomadas como base as seções e questões definidas para a pesquisa de levantamento. Todavia, a ordem dos elementos foi alterada. O critério de ordenação das competências presentes no modelo foi a porcentagem de respondentes com nível de proficiência avançado ou especialista. A soma dessas duas faixas de proficiência afeta diretamente o valor médio obtido para cada variável do questionário. Por exemplo, 74,00% dos respondentes possuem níveis avançado ou especialista para a competência em visualização de dados, fazendo deste item o segundo no qual os profissionais são mais proficientes. Assim, mesmo que essa competência não esteja entre as principais encontradas pelos processos de mineração de texto, ela foi colocada na segunda posição das competências do grupo Análise de Dados.

A ordem dos grupos também foi alterada em relação ao instrumento de coleta de dados da pesquisa de levantamento. O grupo de Análise de Dados foi priorizado, visto que, excluindo-se as competências socioculturais, foi aquele que apresentou o melhor desempenho. Logo, constata-se que os cientistas de dados estão mais desenvolvidos nas competências desse grupo do que em Tecnologia ou Entendimento de Negócios. Por consequência, espera-se que essas competências também sejam mais requeridas.

No modelo, ainda foram realizados alguns agrupamentos de competências, visto que algumas possuem relação de hierarquia entre si. No grupo de Competências Socioculturais, além da separação de competências individuais e organizacionais, as competências em “Liderança técnica” e “Liderança estratégica” foram agrupadas sobre a competência “Liderança”. Outro agrupamento é verificado no grupo de Tecnologia, onde “Gestão de dados” se configura como uma competência “guarda-chuva” para “Banco de dados”, “Dados estruturados”, “Dados semiestruturados” e “Dados não-estruturados”. O mesmo artifício foi aplicado às competências “Técnicas de regressão” e “Análises preditivas”, incorporadas à competência “Estatística”.

Esse procedimento de agrupamento buscou resolver uma questão de granularidade das competências, apontada no pré-teste do questionário. Isto é, o profissional se tornar proficiente em técnicas de regressão não tem o mesmo peso que se aprofundar em estatística, uma área muito mais ampla que até mesmo engloba a primeira. Por isso, as competências do modelo que englobam outras competências não se restringem aos itens hierarquicamente inferiores, apenas indicam essa relação de hierarquia. Isto é, estatística é muito mais que técnicas de regressão e análises preditivas, bem como gestão de dados é um campo muito mais amplo que banco de dados e estrutura de dados.

Embora outras competências do modelo possuam grande afinidade entre si, apresentando amplas áreas de intersecção, não se julgou pertinente estabelecer uma relação de hierarquia. As competências “Programação” e “Desenvolvimento de software” são exemplos dessa situação. Para se desenvolver um *software*, em princípio, a programação é necessária, mas o processo não se restringe a isso. Há muitos outros processos envolvidos. Do mesmo modo que nem toda ação de programar está relacionada ao desenvolvimento de software. Por isso, estas duas competências foram mantidas separadas no modelo.

Três outros aspectos devem ser comentados em relação ao modelo. O primeiro é a aparência das competências “*Hacking*”, “Dados distribuídos” e “Gestão de projetos” que se apresentam semitransparentes. Este recurso visual foi utilizado para indicar que essas competências não foram encontradas em nenhum dos três processos de mineração de texto (anúncios de vagas de emprego, cursos superiores e cursos livres). Mesmo assim, a manutenção dessas competências no modelo se fundamenta pela presença desse itens na literatura e corroboração dos respondentes

da pesquisa de levantamento, que demonstraram altos níveis de proficiência, especialmente para “Hacking” e “Gestão de projetos”.

Outro recurso gráfico utilizado é linha pontilhada que liga a competência “Machine learning” ao grupo de Análise de Dados. Neste caso, a análise estatística demonstrou que a competência em questão, alocada no grupo de Tecnologia, possui mais afinidade ao grupo Análise de Dados. No modelo, “Machine learning” foi mantida no grupo original, mas ficou demonstrada sua relação com o grupo de afinidade. Por fim, deve-se destacar a manutenção da competência genérica “Análise de dados” que possui o mesmo nome do primeiro grupo do modelo. A manutenção dessa competência foi decorrente de seu desempenho junto aos respondentes da pesquisa de levantamento, indicando que os profissionais de dados se veem como altamente competentes em analisar dados.

Por fim, salienta-se que a proposta apresentada nesta tese corresponde a uma primeira versão de um modelo de competência para a Ciência de Dados. Assim, esse modelo pode sofrer adaptações para o contexto de uso, bem como poderá ser revisado por pesquisas futuras. No próximo capítulo, são apresentadas as contribuições do modelo, tal como da própria tese, bem como as considerações finais da pesquisa.

8 CONSIDERAÇÕES FINAIS

Neste capítulo, está a síntese da pesquisa e apresenta os comentários finais sobre os resultados obtidos pelas análises aplicadas. Além disso, é retomada a questão que guiou o trabalho, além dos objetivos geral e específicos. Na sequência, são apresentadas as limitações da pesquisa, sugestões para trabalhos futuros e, por fim, as principais contribuições da tese.

Inicialmente, reforça-se que, por meio da Análise Fatorial Confirmatória, a pesquisa de levantamento realizada junto a profissionais de dados no Brasil confirmou o modelo de competências para a Ciência de Dados formado por quatro construtos: *Tecnologia*, *Análise de Dados*, *Entendimento de Negócios* e *Competências Socioculturais*. Apesar dos ajustes necessários, essas quatro dimensões identificadas na literatura foram confirmadas pelos procedimentos estatísticos. Paralelamente, a *surve*y também possibilitou o desenho do perfil dos profissionais da Ciência de Dados, confirmando-a como um campo em desenvolvimento, composto por trabalhadores jovens e com pouco tempo de experiência. Além disso, reforçou pesquisas anteriores de que os profissionais que atuam na Ciência de Dados têm formação em disciplinas já tradicionais como Estatística, Ciência da Computação, Matemática, Física, Economia (BURTCH WORKS, 2019; HARRIS; MURPHY; VAISMAN, 2013-; LOUKIDES, 2012). Este resultado era esperado visto que o curso superior mais antigo do Brasil, surgido com o nome Ciência de Dados, teve seu funcionamento iniciado somente em 2018.

A análise dos anúncios de vagas para cientistas de dados consolidou a importância dos algoritmos de *machine learning* e inteligência artificial, mas também de outros conceitos como modelos estatísticos, análise de dados, Python, banco de dados. Demonstrou também que a divulgação do salário não é prática comum aos anunciantes que, por sua vez, buscam profissionais experientes sem priorizarem formações específicas. A exigência por níveis de mestrado e doutorado é baixo se comparado a pesquisas anteriores (KIM; LEE, 2016), concluindo que a formação não é o foco dos anúncios de emprego (BAUMEISTER; BARBOSA; GOMES, 2020). Se por um lado os anúncios englobam com frequência termos do contexto organizacional, relacionados a benefícios empregatícios, por outro, expressões relacionadas a habilidades interpessoais ou à ética dos dados são mais raras.

Os cursos de nível superior reforçaram a relação da Ciência de Dados com a Estatística, quando destaca termos como análise de dados, modelo estatístico, séries temporais, bem como, *machine learning*. É evidenciado que a estatística é um dos principais fundamentos da educação superior da área. Dentre outros conceitos que se destacam no conteúdo dos cursos superiores são Ciências da Computação, Matemática e Administração, além de mineração de dados, processamento de linguagem natural e redes neurais. Todos associados com frequência às competências do cientista de dados, segundo a literatura (DEMCHENKO; COMMINIELLO; REALI, 2019; ROMERO; VENTURA, 2017; STODDER, 2018; WASHINGTON DURR, 2018).

A análise dos termos mais frequentes dos cursos livres demonstrou que este conteúdo está mais alinhado aos cursos superiores do que aos anúncios de emprego, mesmo que grande parte do conteúdo analisado seja compartilhada entre os três conjuntos de dados. Dentre as particularidades dos cursos livres, destaca-se a ênfase à visualização de dados, valorizada pela literatura (HARDIN *et al.*, 2015; TOSIC; BEESTON, 2018; WASHINGTON DURR, 2018), sendo uma das competências mais desenvolvidas segundo a pesquisa de levantamento, mas que não é encontrada nos anúncios e nem nos cursos superiores. Ademais, os cursos livres demonstram um caráter mais pragmático, enfatizando tutoriais (passo a passo), ferramentas e soluções finais.

Revisados alguns dos resultados obtidos, é retomada a questão de pesquisa:

QUAIS AS COMPETÊNCIAS NECESSÁRIAS PARA UM CIENTISTA DE DADOS ATUAR NO BRASIL?

A resposta a essa pergunta pode ser formulada da seguinte forma:

Os resultados da pesquisa sugerem que, para a amostra avaliada, para atuar como cientista de dados no Brasil é necessário possuir competência em análise de dados, com ênfase em modelos de *machine learning*, competência em tecnologia, especialmente gestão de dados e programação, possuir conhecimento interdisciplinar, com foco no domínio de atuação, além de demonstrar interesse em solucionar

problemas e capacidade de comunicar seus resultados para os mais diferentes públicos.

Para fundamentar essa resposta, foram utilizados os resultados obtidos pela pesquisa. Inicialmente, os quatro grupos de competências (*Tecnologia*, *Análise de Dados*, *Entendimento de Negócios* e *Competências Socioculturais*) do modelo proposto estão contemplados. Ademais, foram definidas as competências primordiais para cada grupo. Para Análise de Dados, além da competência com o mesmo nome do grupo, têm-se os modelos estatísticos de *machine learning*, presente entre os mais recorrentes em todos os conjuntos de dados utilizados. Para Tecnologia, os destaques são a gestão de dados que, segundo o modelo, engloba banco de dados e estrutura de dados, e programação, também recorrente para anúncios, conteúdo educacional, além de bastante desenvolvida pelos respondentes da pesquisa de levantamento.

Dentre as competências de Entendimento de Negócios, grupo que apresentou as menores médias da pesquisa de levantamento, destaca-se conhecer bem o domínio de atuação, seja educação, saúde, financeiro, marketing etc. Ademais, ressalta-se o conhecimento interdisciplinar do cientista de dados, característico dos profissionais com formato em "T", que facilita o trabalho em equipes cujos integrantes possuem diferentes formações e repertório. Para concluir a resposta, é destacada a comunicação, principal competência do grupo sociocultural e fundamental para que os resultados obtidos pelos cientistas de dados sejam compreendidos pelas organizações.

Para complementar a resposta à questão de pesquisa, é retomado o objetivo geral que foi analisar as competências necessárias para a atuação de cientistas de dados no Brasil e os objetivos específicos:

a) **Investigar o perfil e as competências dos cientistas de dados atuantes no Brasil:** as informações que buscam definir o perfil dos cientistas de dados atuantes no Brasil foram obtidas por meio da pesquisa de levantamento. A definição das questões do instrumento de coleta de dados é apresentada na seção 5.3.2, onde o Quadro 11 contém as variáveis utilizadas para o perfil dos profissionais participantes. Os resultados, expostos na seção 6.1.1, confirmam a predominância no número de homens (81,05%) em relação a mulheres (18,06%), além de profissionais jovens, com média de idade de

30,99 (desvio padrão = 7,30, moda = 27, mediana = 30), com pouco de tempo de experiência na área (média = 5,87 anos, 69,64% possuem até seis anos de experiência). Além disso, a pesquisa aponta que o estado de São Paulo concentra ¼ dos respondentes, mas a cidade com maior número de respostas foi Curitiba, com 10,57%. A educação foi outro item investigado, apontando que o nível superior é quase unanimidade entre os profissionais, visto que apenas um respondente não possui e não está cursando nenhuma graduação. Outro indicativo do alto nível educacional dos profissionais é que 70,04% concluíram ou estão cursando uma pós-graduação, incluindo mestrado, doutorado e especialização. Por fim, foi investigada a remuneração, onde nenhuma faixa salarial apresentada demonstrou predominância em relação às demais. Ainda assim, pode-se verificar que maior tempo de experiência está relacionado a maiores salários. O mapeamento das competências também foi realizado por meio da pesquisa de levantamento. As questões referentes às competências, disponíveis na seção 5.3.2 (Quadro 6, Quadro 7, Quadro 8, Quadro 9 e Quadro 10), somam 43 variáveis, organizadas em quatro grupos (*Tecnologia*, *Análise de Dados*, *Entendimento de Negócios* e *Competências Socioculturais*). Em suma, o grupo de Competências Socioculturais apresentaram os maiores níveis de proficiência, logo a maior média (7,582). A segunda maior média (6,294) foi obtida pelo grupo de Análise de Dados, seguido pelo de Tecnologia (5,626). Por fim, o grupo com menor desempenho era composto pelas competências de Entendimento de Negócios, com média geral de 5,375). No geral, percebe-se que nas Competências Socioculturais, relacionadas às chamadas *soft skills*, os respondentes tendem a se autoavaliarem de maneira menos rigorosa. Além disso, neste grupo estão as competências organizacionais, também muito bem avaliadas pelos respondentes. Assim, considera-se que o grupo de Análise de Dados é aquele mais desenvolvido, dentre aqueles diretamente relacionados à Ciência de Dados. Na sequência, estão as competências tecnológicas, seguidas pelas competências em negócios, as mais carentes de desenvolvimento.

b) **Identificar os requisitos em anúncios de vagas de emprego para cientistas de dados:** para reconhecer os requisitos presentes nos anúncios

de vagas de emprego, foi utilizado o protocolo definido na seção 5.5.2, mais especificamente, adotaram-se procedimentos de mineração de texto, incluindo termos mais frequentes, modelagem de tópicos e agrupamento. A frequência dos termos foi a técnica que melhor evidenciou os requisitos necessários para um candidato a cientista de dados, conforme demonstrado na seção 6.2. Dentre as expressões mais frequentes, estão *machine learning*, modelos estatísticos, análise de dados, inteligência artificial, ciência da computação, tecnologia, python, banco de dados, modelos preditivos, processamento de linguagem natural, entre outros. Com as técnicas utilizadas, foi possível identificar padrões de predominâncias de competência nos anúncios, permitindo a comparação com os conjuntos de dados dos cursos superiores e livres.

c) **Identificar os tópicos abordados em cursos de nível superior e cursos livres para cientistas de dados:** o conteúdo dos cursos superiores e cursos livres foram submetidos ao mesmo protocolo aplicado aos anúncios, sendo os resultados apresentados nas seções 6.3 e 6.4, respectivamente. Para os cursos superiores, nota-se a relevância da Estatística para os cursos de Ciência de Dados, visto que conceitos como análise de dados, modelo estatístico, séries temporais, *machine learning*, fundamentam a educação superior da área. Embora o conteúdo dos cursos livres apresente muitos termos em comum com os cursos superiores, para este grupo de documentos, percebe-se um aspecto mais prático, focado no ensino técnico de ferramentas.

d) **Comparar os requisitos demandados pelo mercado com o conteúdo abordado na educação formal e livre:** no Capítulo 7, encontra-se o cruzamento dos três conjuntos de dados submetidos à mineração de texto (anúncios, cursos superiores e cursos livres) com as competências identificadas na literatura que fundamentam o instrumento de coleta de dados da pesquisa de levantamento. Assim, além da comparação entre os três *corpora*, as competências mensuradas junto aos profissionais passaram por um processo de validação. Como resultado, tem-se que há uma convergência entre os anúncios, os cursos superiores e os cursos livres,

embora cada conjunto de dados possua características e padrões próprios. Ademais, as competências da pesquisa de levantamento também apresentaram resultado satisfatório, visto que apenas três competências (Habilidades em *Hacking*, Sistemas de dados distribuídos e Gestão de projetos) não foram citadas em nenhum dos três conjuntos de dados.

e) **Elaborar um modelo com as competências necessárias à Ciência de Dados no Brasil:** com os resultados da pesquisa de levantamento e dos processos de mineração de texto, pode-se elaborar o modelo com as competências necessárias para atuar com Ciência de Dados no contexto brasileiro. Portanto, o modelo, apresentado na Figura 56 do Capítulo 7, é resultado de todos os conjuntos de dados e das técnicas utilizadas na pesquisa.

Desta forma, julga-se que os objetivos específicos definidos foram concluídos e, por consequência, foi cumprido também o objetivo geral:

Analisar as competências necessárias para a atuação de cientistas de dados no Brasil.

A seguir, são apresentadas as contribuições da pesquisa.

8.1 CONTRIBUIÇÕES DA PESQUISA

Essa pesquisa de tese visa contribuir com seis atores envolvidos em diferentes níveis com a Ciência de Dados: (1) profissionais que estão migrando ou já trabalham com Ciência de Dados; (2) organizações que buscam por cientistas de dados; (3) instituições de ensino que trabalham para a formação de cientista de dados; (4) comunidade acadêmica; (5) o autor e; (6) Programa de Pós-Graduação em Gestão da Informação (PPGGI).

Para os profissionais que estão iniciando como cientista de dados, seja por uma migração profissional ou pelo início da vida profissional, o modelo fornecido é uma referência de trilha de estudo e aprendizagem, além de uma ferramenta de autoavaliação. Ressaltando que o modelo de competências proposto não foca em

ferramentas específicas, diminuindo a possibilidade de se tornar obsoleto. Além disso, ao responder a pesquisa de levantamento, o respondente poderá comparar seus resultados com a média registrada no histórico de respostas por meio de um sistema que está em desenvolvimento.

Contratar um cientista de dados é uma tarefa complexa, tanto pela dificuldade em encontrar profissionais qualificados, quanto pela dificuldade de manter o profissional na organização (DAVENPORT; PATIL, 2012; REIS; SÁ, 2020). Os resultados da pesquisa ajudam as organizações que precisam buscar profissionais nessa área, seja para qualificar seus anúncios, seja para avaliar adequadamente as competências que realmente são necessárias para a atuação de um cientista de dados. Outra contribuição importante para essa situação, é o alinhamento de expectativas em relação ao que um cientista de dados pode entregar para as organizações.

Para as instituições de ensino, tanto de nível superior quanto da educação informal, a pesquisa traz relevantes contribuições para a montagem de currículos e conteúdos fundamentais aos futuros cientistas de dados. Por exemplo, visto que o grupo de Entendimento de Negócios foi o que apresentou o pior desempenho do modelo, não valeria o esforço das IES reforçarem esse conteúdo em suas grades? Ademais, a falta de ênfase em questões ligadas a ética dos dados, privacidade dos dados, bem como precaução de análises tendenciosas, é sinal de alerta para as instituições de ensino. Especialmente, visto que esses temas deveriam estar presentes na vida do cientista de dados desde a sua formação.

A contribuição para a comunidade acadêmica se dá pela iniciativa de preencher a lacuna identificada e apresentada na justificativa desta pesquisa (Seção 1.3). Conforme relatado, embora o interesse acerca da Ciência de Dados tenha aumentado nos últimos anos, as publicações científicas não acompanharam esse crescimento, tanto que nenhuma tese sobre o tema foi encontrada. Dessa forma, a pesquisa, além da própria tese, busca contribuir por meio de publicações em periódicos, como ocorrido com artigo publicado na RISTI (GUIMARÃES; MENDES JÚNIOR; FREITAS, 2022). Além disso, a pesquisa fornece subsídios para que outros pesquisadores se utilizem do instrumento de coleta de dados em trabalhos futuros.

Para o autor, a pesquisa contribui para seu amadurecimento como pesquisador, visto que, como doutorando, teve que aumentar sua autonomia, assumindo as consequências pelas decisões tomadas. Além disso, a pesquisa se

mostrou um desafio teórico, técnico e metodológico, exigindo momentos de superação e impulsionando sua evolução como acadêmico.

Por último, a pesquisa visa contribuir para o PPGGI não somente com o depósito em seu acervo de teses, mas, principalmente, na publicação de artigos. Conforme informado na seção de justificativa, ainda que nenhum trabalho de conclusão do PPGGI tenha abordado a Ciência de Dados como objeto de estudo, entende-se que compreender a atuação do cientista de dados no processo de extração de conhecimento a partir dos dados está de acordo com a proposta do programa. Ademais, destaca-se que os elementos característicos desse processo se relacionam à gestão da informação e são de interesse aos mais diversos segmentos, sejam eles educacionais, governamentais, negócios, serviços ou indústrias.

8.2 LIMITAÇÕES

Em primeiro lugar, destaca-se o recorte geográfico definido para a pesquisa de levantamento. Ainda que a clarificação do conceito da Ciência de Dados, bem como do cientista de dados e suas competências, não seja de interesse apenas para o cenário Brasil, há motivos para esta delimitação. Inicialmente, expandir a pesquisa para múltiplos países e idiomas implicaria em procedimentos técnicos, adaptações idiomáticas e cuidados adicionais nos processos de análise. Ademais, conforme declarado na justificativa, entende-se que o cenário brasileiro carece de mais pesquisas e, consequentemente, publicações focadas em suas particularidades.

Outra limitação apresentada na pesquisa de levantamento foi o desequilíbrio entre as funções dos respondentes. Das 227 respostas válidas, 124 profissionais se identificam como cientistas de dados (54,63%) e 26 respondentes se identificam como analistas de dados (11,45%). Todas as demais funções (engenheiro de dados, analistas de BI, engenheiro de *machine learning*) não atingiram 10,00% da amostra. Com esta diferença nos tamanhos dos grupos, não foi possível investigar as particularidades e distinções nas competências de cada função. Ainda em relação à pesquisa de levantamento, conforme já relatado, ainda que válido (HAIR *et al.*, 2014; MATSUNAGA, 2010), o tamanho da amostra pode ser considerado uma limitação que pode ser melhorado em pesquisas futuras.

Para os procedimentos de mineração de texto, a limitação está na ausência de cursos de pós-graduação, seja *stricto sensu*, seja *lato sensu*. Sabe-se que muitos

profissionais atuando com Ciência de Dados possuem formação deste nível, conforme descrido na seção 6.1.1. Todavia, a quantidade desse tipo de curso é elevada, tanto quanto a diversidade dos portais que contêm o material de interesse. Por isso, optou-se por retirar esses cursos da pesquisa

Por fim, embora também já relatado na devida seção, considera-se uma limitação que a configuração inicial do modelo conceitual das competências da Ciência de Dados (Figura 32) não tenha se confirmado com todas as variáveis na Análise Fatorial Confirmatória. Ainda assim, com os devidos ajustes, remoção e deslocamento de variáveis, obteve-se um modelo estatístico satisfatório, com bom ajustamento.

8.3 SUGESTÕES PARA TRABALHOS FUTUROS

Dentre as sugestões para trabalhos futuros, destaca-se, inicialmente, a validação do modelo proposto junto a especialistas por meio de entrevistas. Dessa forma, profissionais experientes e outros pesquisadores podem contribuir para que o modelo seja aprimorado. Além disso, a fim de se obter mais respostas, sugere-se serem estabelecidas parcerias com órgãos profissionais, como a Associação Brasileira de Ciência de Dados (https://abracd.org/).

Outras sugestões, que já foram levantadas, para a pesquisa de levantamento envolvem estratificar a amostra por perfis profissionais (cientista de dados, engenheiro de dados, analista de dados, analista de BI, dentre outros). Dessa forma, cada papel profissional poderia ser analisado individualmente e comparado com os demais. Assim, amostras futuras poderiam reforçar e, até mesmo, generalizar os resultados do modelo estatístico formado pelos fatores Tecnologia, Análise de Dados, Entendimento de Negócios e Competências Socioculturais. Todavia, é recomendado a trabalhos futuros que verifiquem a existência de um quinto fator referente às competências organizacionais, variáveis retiradas do modelo estatístico aqui apresentado. Acrescente-se, a adaptação e divulgação da pesquisa de levantamento para profissionais de outros países proporcionariam a possibilidade comparar a realidade brasileira com as demais.

Para os procedimentos de mineração de texto, ou análise qualitativa, as recomendações aqui retomadas são o aumento dos tamanhos dos *corpora*, uma vez que aplicações de processamento de linguagem natural são mais confiáveis em

conjuntos de dados maiores (WOLFRAM, 2017). Além disso, coleta de anúncios de outros perfis como estatísticos, engenheiro de dados, analistas, dentre outros, possibilitaria a comparação e distinção das competências para cada papel.

Por fim, outra recomendação, já mencionada, é a adoção de metodologias mistas, como análise de conteúdo, utilizada em pesquisas da área (BAUMEISTER; BARBOSA; GOMES, 2020; KIM; LEE, 2016; MEYER, 2019). Novos métodos de análise, aplicados sobre o mesmo conteúdo, possibilitariam a comparação e avaliação dos resultados definidos para esta tese.

REFERÊNCIAS

ALVARENGA, A. T. de et al. Histórico, fundamentos filosóficos e teórico-metodológicos da interdisciplinaridade. In: PHILIPPI JR., A.; SILVA NETO, A. J. (org.). **Interdisciplinaridade em ciência, tecnologia & inovação**. Barueri: Manole, 2011. p. 983.

ALVES-MAZZOTTI, A. J.; GEWANDSZNAJDER, F. **O método nas ciências naturais e sociais**: pesquisa quantitativa e qualitativa. 2. ed. São Paulo: Thompson, 1998.

ANDERSON, P. et al. An undergraduate degree in data science: Curriculum and a decade of implementation experience. **SIGCSE 2014 - Proceedings of the 45th ACM Technical Symposium on Computer Science Education**, p. 145–150, 2014.

ANDERSON, P.; MCGUFFEE, J.; UMINSKY, D. Data science as an undergraduate degree. **SIGCSE 2014 - Proceedings of the 45th ACM Technical Symposium on Computer Science Education**, p. 705–706, 2014.

ANDREU-PEREZ, J. et al. Big Data for Health. **IEEE Journal of Biomedical and Health Informatics**, v. 19, n. 4, p. 1193–1208, 2015.

ANTONS, D. et al. The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. **R and D Management**, v. 50, n. 3, p. 329–351, 2020.

AZEVEDO, A.; SANTOS, M. F. KDD, SEMMA and CRISP-DM: a parallel overview. In: , 2008, Amsterdam. **IADIS European Conference on Data Mining 2008**. Amsterdam: IADIS, 2008. p. 182–185.

BACHELARD, G. **A formação do espírito científico**: contribuição para uma psicanálise do conhecimento. Rio de Janeiro: Contraponto, 1996.

BALBINO, J. N. **Lições aprendidas**: o potencial das contribuições do uso de dados abertos da CAPES para a gestão de programas stricto sensu da área interdisciplinar. 2021. 243 f. - Universidade Federal do Paraná, 2021.

BANAFA, A. **What is Data Science?**. 2014. Disponível em: https://works.bepress.com/ahmed-banafa/15/. Acesso em: 9 ago. 2020.

BAŠKARADA, S.; KORONIOS, A. Unicorn data scientist: the rarest of breeds. **Program**, v. 51, n. 1, p. 65–74, 2017.

BAUMEISTER, F.; BARBOSA, M. W.; GOMES, R. R. What is required to be a data scientist? Analyzing job descriptions with centering resonance analysis. **International Journal of Human Capital and Information Technology Professionals**, v. 11, n. 4, p. 21–40, 2020.

BAZZO, W. A. **Ciência, Tecnologia e Sociedade**: e o conceito da educação tecnológica. 4. ed. Florianópolis: Editora da UFSC, 2014.

BEDREGAL-ALPACA, N.; ARUQUIPA-VELAZCO, D.; CORNEJO-APARICIO, V. Técnicas de Data Mining para extraer perfiles comportamiento académico y predecir la deserción universitaria. **Revista Ibérica de Sistemas e Tecnologias de Informação**, v. E27, p. 592–604, 2020. Disponível em: https://www.proquest.com/scholarly-journals/técnicas-de-data-mining-para-extraer-perfiles/docview/2385757429/se-2. Acesso em: 9 ago. 2020.

BENGFORT, B.; BILBRO, R.; OJEDA, T. **Applied Text analysis with Python**: Enabling Language-Aware Data Products with Machine Learning. Sebastopol: O'Reilly Media, Inc., 2018-. ISSN 1098-6596.

BERKELEY SCHOOL OF INFORMATION. **What is Data Science?**. 2020. Disponível em: https://ischoolonline.berkeley.edu/data-science/what-is-data-science/. Acesso em: 23 set. 2020.

BONNELL, J.; OGIHARA, M.; YESHA, Y. Challenges and Issues in Data Science Education. **Computer**, v. 55, n. 2, p. 63–66, 2022. Disponível em: https://ieeexplore.ieee.org/document/9714096/. Acesso em: 14 out. 2022.

BÖRNER, K. et al. Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. **Proceedings of the National Academy of Sciences of the United States of America**, v. 115, n. 50, p. 12630–12637, 2018.

BOWNE-ANDERSON, H. What Data Scientists Really Do, According to 35 Data Scientists. **Harvard Business Review**, p. 2–6, 2018. Disponível em: https://hbr.org/2018/08/what-data-scientists-really-do-according-to-35-data-scientists. Acesso em: 14 out. 2022.

BRANDT, P. S. **The emergence of the data science profession**. 2016. 362 f. - Columbia University, 2016. Disponível em: https://doi.org/10.7916/D8BK1CKJ. Acesso em: 14 out. 2022.

BRASIL. **Lei No 13.709/2018 - Lei Geral de Proteção de Dados Pessoais (LGPD)**. 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 1 jun. 2022.

BRASIL. **Lei No 13.853/2018**. 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 1 jun. 2022.

BREIMAN, L. Statistical modeling: The two cultures. **Statistical Science**, v. 16, n. 3, p. 199–215, 2001.

BROWN, T. A. **Confirmatory Factor Analysis for Applied Research**. 2. ed. New York: The Guilford Press, 2015.

BRUNI, A. L. **Estatística Aplicada à Gestão Empresarial**. 4. ed. São Paulo: Atlas, 2013.

BURTCH WORKS. **The Burtch Works Study**: Salaries of Data Science & Analytics Professionals. Evanston: Burtch Works Executive Recruiting, 2021. Disponível em:

https://www.burtchworks.com/big-data-analyst-salary/big-data-career-tips/the-burtch-works-study/. Acesso em: 14 out. 2022.

BURTCH WORKS. **The Burtch Works Study**: Salaries of Predictive Analytics Professionals. Evanston: Burtch Works Executive Recruiting, 2019.

CAMARGO, M. D. **Plano de desenvolvimento organizacional a partir do mapeamento de competências individuais**. 2013. 142 f. - Universidade Federal do Paraná, Curitiba, 2013.

CAO, L. Data Science: A Comprehensive Overview. **ACM Computing Surveys**, v. 50, n. 3, p. 1–42, 2017. Disponível em: https://dl.acm.org/doi/10.1145/3076253. Acesso em: 14 out. 2022.

CAO, L. Data Science: Profession and Education. **IEEE Intelligent Systems**, v. 34, n. 5, p. 35–44, 2019.

CARROT2 CLUSTERING ENGINE. **Clustering Workbench**. 2021. Disponível em: https://search.carrot2.org/#/workbench. Acesso em: 14 out. 2022.

CASTELLO, F. **Cientistas de dados**: proposta de um modelo conceitual considerando sua definição, sua formação, suas habilidades e as ferramentas que utilizam. 2021. 195 f. - Universidade de São Paulo, 2021. Disponível em: https://www.fea.usp.br/administracao/eventos/mestrado-cientistas-de-dados-proposta-de-um-modelo-conceitual-considerando-sua. Acesso em: 14 out. 2022.

CHALMERS, A. F. **O que é ciência afinal?** 2. ed. São Paulo: Editora Brasiliense, 1993.

CHANDRASEKARAN, S. **Becoming a Data Scientist**: Curriculum via Metromap. 2013. Disponível em: http://nirvacana.com/thoughts/2013/07/08/becoming-a-data-scientist/. Acesso em: 1 jun. 2020.

CHAPMAN, P. et al. **CRISP-DM 1.0: Step-by-step data mining guide**. U. S. A.: SPSS, 2000.

CHEN, Y. et al. Big data analytics and big data science: a survey. **Journal of Management Analytics**, v. 3, n. 1, p. 1–42, 2016. Disponível em: http://dx.doi.org/10.1080/23270012.2016.1141332.

CHEN, J. et al. Fundamentals of Data Science for Future Data Scientists. In: HAWAMDEH, S.; CHANG, H.-C. (org.). **Analytics and Knowledge Management**. Boca Raton: CRC Press, 2018. p. 167–194.

CHIBENI, S. S. **O que é ciência?** Campinas: Departamento de Filosofia - IFCH - Unicamp, 2004. Disponível em: https://www.unicamp.br/~chibeni/textosdidaticos/ciencia.pdf. Acesso em: 14 out. 2022.

CHONG, M.; CHANG, H.-C. Social Media Analytics. In: HAWAMDEH, S.; CHANG, H.-C. (org.). **Analytics and Knowledge Management**. London: CRC Press, 2018. p. 215–240.

CHYUNG, Y.; SWANSON, I. **Evidence-Based Survey Design**: The Use of Sliders. 2019. Disponível em: https://www.td.org/insights/evidence-based-survey-design-the-use-of-sliders. Acesso em: 14 out. 2022.

CISCO IT INSIGHTS. **Preparing for the Data Science-Driven Era Cisco IT Insights**. 2016. Disponível em: https://www.cisco.com/c/en/us/solutions/collateral/enterprise/cisco-on-cisco/i-dc-09022015-preparing-for-data.html. Acesso em: 1 jun. 2020.

CLELAND, C. E. Historical science, experimental science, and the scientific method. **Geology**, v. 29, n. 11, p. 987–990, 2001.

CLEVELAND, W. S. Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics. **International Statistical Reviews**, v. 69, n. 1, p. 21–26, 2001.

COMMITTEE ON ENVISIONING THE DATA SCIENCE DISCIPLINE. **Data Science for Undergraduates**. Washington: The National Academies Press, 2018. E-book. Disponível em: https://www.nap.edu/catalog/25104/data-science-for-undergraduates-opportunities-and-options. Acesso em: 14 out. 2022.

CONWAY, D. **The Data Science Venn Diagram**. 2010. Disponível em: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram. Acesso em: 24 ago. 2020.

COOK, A. **Data Us**e: An analysis of the impact of survey scales. 2013. Disponível em: https://www.quirks.com/articles/data-use-an-analysis-of-the-impact-of-survey-scales. .

CRESWELL, J. W. **Projeto de Pesquisa**: método qualitativo, quantitativo e misto. 3. ed. Porto Alegre: Artmed Editora, 2010.

CROWDFLOWER. **Data Science Report 2016**. São Francisco: CrowdFlower, 2016. Disponível em: https://visit.figure-eight.com/data-science-report.html. Acesso em: 14 out. 2022.

CROWDFLOWER. **Data Scientist Report 2015**. São Francisco: CrowdFlower, 2015. Disponível em: https://visit.figure-eight.com/2015-data-scientist-report. Acesso em: 14 out. 2022.

CROWDFLOWER. **Data Scientist Report 2017**. São Francisco, 2017. p. 14. Disponível em: https://visit.figure-eight.com/rs/416-ZBE-142/images/data-scientist-report-dec.pdf. Acesso em: 14 out. 2022.

CUNHA, R. **Procuram-se cientistas de dados**. 2018. Disponível em: https://www.linkedin.com/pulse/procuram-se-cientistas-de-dados-rodrigo-cunha/. Acesso em: 1 jun. 2020.

CURTY, R. G.; SERAFIM, J. D. S. A formação em ciência de dados: uma análise preliminar do panorama estadunidense. **Informação & Informação**, v. 21, n. 2, p. 307–331, 2016.

DAMA INTERNATIONAL TECHNICS. **DAMA-DMBOK**: Data Management Body of Knowledge: 2nd Edition. 2. ed. New Jersey: Basking Ridge, 2017.

DATA SCIENCE ASSOCIATION. **Data Science Code of Professional Conduct**. 2014. Disponível em: https://www.datascienceassn.org/code-of-conduct.html. Acesso em: 14 out. 2022.

DAVENPORT, M. G. et al. **Data Driven**: What students need to succeed in a rapidly changing business world. London: PricewaterhouseCoopers LLP, 2015.

DAVENPORT, T. H.; PATIL, D. J. **Data scientist**: The sexiest job of the 21st century. Harvard Business Review, v. 90, n. 10, p. 5, 2012. Disponível em: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century. Acesso em: 14 out. 2022.

DEMCHENKO, Y. et al. **Data Science Body of Knowledge**. Eindhoven: European Commission, 2019. E-book. Disponível em: http://edison-project.eu/sites/edison-project.eu/files/filefield_paths/edison_cf-ds-release2-v08_0.pdf. Acesso em: 14 out. 2022.

DEMCHENKO, Y. et al. **EDISON data science framework**: A foundation for building data science profession for research and industry. Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom, v. 0, n. Dtw, p. 620–626, 2016.

DEMCHENKO, Y. **EDISON Data Science Framework**: Part 4 . Data Science Professional Profiles (DSPP). Amsterdam: EDISON, 2017.

DEMCHENKO, Y.; BELLOUM, A.; WIKTORSKI, T. **EDISON Data Science Framework**: Part 1. Data Science Competence Framework (CF-DS). Amsterdam: EDISON, 2017. Disponível em: https://edison-project.eu/data-science-competence-framework-cf-ds/. Acesso em: 14 out. 2022.

DEMCHENKO, Y.; BELLOUM, A.; WIKTORSKI, T. **EDISON Data Science Framework**: Part 3. Data Science Model Curriculum (MC-DS). Amsterdam: EDISON, 2017. Disponível em: http://creativecommons.org/licenses/by/4.0/. Acesso em: 14 out. 2022.

DEMCHENKO, Y.; COMMINIELLO, L.; REALI, G. Designing customisable data science curriculum using ontology for data science competences and body of knowledge. **ACM International Conference Proceeding Series**, p. 124–128, 2019.

DHAR, V. Data science and prediction. **Communications of the ACM**, v. 56, n. 12, p. 64–73, 2013.

DONOHO, D. 50 Years of Data Collection. **Journal of Computational and Graphical Statistics**, v. 26, n. 4, p. 745–766, 2015.

DONOHO, D. 50 Years of Data Science. **Journal of Computational and Graphical Statistics**, v. 26, n. 4, p. 745–766, 2017. Disponível em: https://doi.org/10.1080/10618600.2017.1384734. Acesso em: 14 out. 2022.

DURAND, T. Forms of Incompetence. **Theory Development for Competence-Based Management**, v. 33, n. 0, p. 69–95, 2000.

ECO, U. **Como se faz uma tese**. 26. ed. São Paulo: Perspectiva, 2016.

EDISON PROJECT. **EDISON**: building the data science profession. 2017. Disponível em: https://edison-project.eu/edison/edison-project/. Acesso em: 7 set. 2020.

ERDFELDER, E. et al. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. **Behavior Research Methods**, v. 41, n. 4, p. 1149–1160, 2009.

ESCO. **Cientista de Dados**. 2020. Disponível em: http://data.europa.eu/esco/occupation/258e46f9-0075-4a2e-adae-1ff0477e0f30. Acesso em: 27 set. 2020.

ESCO. **What is ESCO**. 2020. Disponível em: https://ec.europa.eu/esco/portal/howtouse/21da6a9a-02d1-4533-8057-dea0a824a17a. Acesso em: 19 set. 2020.

EUBANKS, C. **Three Lessons CrossFit Taught Me About Data Science**. 2016. Disponível em: https://blogs.gartner.com/christi-eubanks/three-lessons-crossfit-taught-data-science/. Acesso em: 10 jun. 2020.

EUROPEAN COMMISSION. **Final results of the European Data Market study measuring the size and trends of the EU data economy**. London, 2017. Disponível em: https://digital-strategy.ec.europa.eu/en/library/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy. .

EUROPEAN E-COMPETENCE FRAMEWORK 3.0. **A common European Framework for ICT Professionals in all industry sectors**. 2014. Disponível em: https://www.ecompetences.eu/e-cf-3-0-download/. Acesso em: 20 ago. 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37–53, 1996.

FAZENDA, I. C. A. **Interdisciplinaridade**: história, teoria e pesquisa [livro eletrônico]. 16. ed. Campinas: Papirus Editora, 2017.

FAZENDA, I. C. A.; TAVARES, D. E.; GODOY, H. P. **Interdisciplinaridade na pesquisa científica** [livro eletrônico]. Campinas: Papirus Editora, 2018.

FIDLER, F.; WILCOX, J. Reproducibility of Scientific Result. In: ZALTA, E. N. (org.). **The Stanford Encyclopedia of Philosophy.** Stanford: Metaphysics Research Lab, Stanford University, 2021. E-book. Disponível em: https://plato.stanford.edu/archives/sum2021/entries/scientific-reproducibility/. Acesso em: 14 out. 2022.

FIELD, A.; MILES, J.; FIELD, Z. **Discovering Statistics using R**. London: Sage, 2012-. ISSN 15393704.

FIGURE EIGHT. **Data Scientist Report 2018**. São Francisco: Figure Eight, 2018. Disponível em: https://visit.figure-eight.com/WC-2018-Data-Scientist-Report_.html. Acesso em: 14 out. 2022.

FINDDATALAB.COM. **Web scraping and social media**: the past, the present and the legal question. 2020. Disponível em: https://finddatalab.medium.com/web-scraping-and-social-media-the-past-the-present-and-the-legal-question-6d886f840904. Acesso em: 1 mar. 2021.

FINZER, W. The Data Science Education Dilemma. **Technology Innovations in Statistics Education**, v. 7, n. 2, p. 1–9, 2013. Disponível em: https://escholarship.org/uc/item/7gv0q9dc.

FLEURY, M. T. L.; FLEURY, A. Construindo o Conceito de Competência. **RAC**, n. Edição Especial 2001, p. 183–196, 1991.

FORGÓ, N. et al. An ethico-legal framework for social data science. **International Journal of Data Science and Analytics**, v. 11, n. 4, p. 377–390, 2021. Disponível em: https://doi.org/10.1007/s41060-020-00211-7. Acesso em: 14 out. 2022.

FRANZKE, A. S. et al. **Internet Research**: Ethical Guidelines 3.0. Dublin: The Association of Internet Research, 2019. Disponível em: https://aoir.org/reports/ethics3.pdf. Acesso em: 14 out. 2022.

GAJZLER, M. Text and data mining techniques in aspect of knowledge acquisition for decision support system in construction industry. **Technological and Economic Development of Economy**, v. 16, n. 2, p. 219–232, 2010.

GERINGER, S. **Data Science Venn Diagram v2.0**. 2014. Disponível em: http://www.anlytcs.com/2014/01/data-science-venn-diagram-v20.html. Acesso em: 24 ago. 2020.

GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GIL, A. C. **Métodos e Técnicas de Pesquisa Social**. 6. ed. São Paulo: Atlas, 2008.

GILES, J. Internet encyclopaedias go head to head. **Nature**, v. 438, n. 15, p. 900–901, 2005. Disponível em: http://www.nature.com/articles/438900a. Acesso em: 14 out. 2022.

GOOGLE CLOUD. **Guia sobre Análise de Dados e Aprendizado de Máquina para CIO**. Mountain View: Google Inc., 2017.

GOOGLE INC. **Google Trends**. 2020. Disponível em: https://trends.google.com.br/trends/?geo=BR. Acesso em: 30 jun. 2020.

GOTTIPATI, S.; SHIM, K. J.; SAHOO, S. Glassdoor job description analytics - Analyzing data science professional roles and skills. **IEEE Global Engineering Education Conference**, EDUCON, v. 2021-April, n. April, p. 1329–1336, 2021.

GRANVILLE, V. **Developing analytic talent**: Becoming a data scientist. Indianapolis: John Wiley & Sons, Inc., 2014.

GRAY, J. Jim Gray on eScience: A transformed scientific method. In: HEY, T.; TANSLEY, S.; TOLLE, K. (org.). **The Fourth Paradigm: Data-Intensive Scientific Discovery**. Washington: Microsoft Research, 2007. p. xvii–xxxi. E-book. Disponível em: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/. Acesso em: 14 out. 2022.

GROSSI, V. et al. **Data science**: a game changer for science and innovation. International Journal of Data Science and Analytics, v. 11, n. 4, p. 263–278, 2021. Disponível em: https://doi.org/10.1007/s41060-020-00240-2. Acesso em: 14 out. 2022.

GUIMARÃES, A. J. R.; MENDES JÚNIOR, R.; FREITAS, M. do C. D. **Requisitos para a ciência de dados**: analisando anúncios de vagas de emprego com mineração de texto. Revista Ibérica de Sistemas e Tecnologias de Informação, v. 2022, n. 46, p. 54–70, 2022. Disponível em: http://www.risti.xyz/issues/risti46.pdf. Acesso em: 14 out. 2022.

HAIR, J. F. **Análise Multivariada de Dados**. 6. ed. Porto Alegre: Bookman, 2009.

HAIR, J. F. et al. **Multivariate Data Analysis**. 7. ed. New Jersey: Pearson, 2014.

HAIR, J. F. et al. When to use and how to report the results of PLS-SEM. **European Business Review**, v. 31, n. 1, p. 2–24, 2019.

HALL, P.; PHAN, W.; WHITSON, K. **The Evolution of Analytics**: Opportunities and Challenges for Machine Learning in Business. Sebastopol: O'Reilly Media, Inc., 2016.

HALWANI, M. A. et al. Job qualifications study for data science and big data professions. **Information Technology & People**, 2021. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/ITP-04-2020-0201/full/html. Acesso em: 14 out. 2022.

HARDIN, J. et al. Data Science in Statistics Curricula: Preparing Students to "Think with Data". **American Statistician**, v. 69, n. 4, p. 343–353, 2015.

HAREL, D. On visual formalisms. **Communications of the ACM**, v. 31, n. 5, p. 514–530, 1988.

HARRIS, H. D.; MURPHY, S. P.; VAISMAN, M. **Analysing the Analyzers**: An Introspective Survey of Data Scientists and Their Work. Sebastopol: O'Reilly Media, Inc., 2013-. ISSN 1098-6596.

HASSANI, H. et al. Text mining in big data analytics. **Big Data and Cognitive Computing**, v. 4, n. 1, p. 1–34, 2020.

HAWAMDEH, S.; CHANG, H.-C. **Analytics and Knowledge Management**. Boca Raton: CRC Press, 2018.

HAYASHI, C. **What is Data Science?** Fundamental Concepts and a Heuristic Example. p. 40–51, 1998.

HAYES, B. **Data Science Survey**. 2020. Disponível em: https://loyaltywidget.com/limesurvey/index.php?sid=42831. Acesso em: 15 dez. 2020.

HAYES, B. **Demystifying Data Science For All**. 2017. Disponível em: http://businessoverbroadway.com/2017/10/29/demystifying-data-science-for-all/. Acesso em: 1 jun. 2020.

HAYES, B. **Optimizing your Data Science Team**: A Survey of Data Professionals. Concessão: 2015.

HENSELER, J.; RINGLE, C. M.; SARSTEDT, M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. **Journal of the Academy of Marketing Science**, v. 43, n. 1, p. 115–135, 2015.

HEPBURN, B.; ANDERSEN, H. Scientific Method. In: ZALTA, E. N. (org.). **The Stanford Encyclopedia of Philosophy**. Stanford: Metaphysics Research Lab, Stanford University, 2021. E-book. Disponível em: https://plato.stanford.edu/archives/sum2021/entries/scientific-method/.

IBM. **The Data Science Skills Competency Model**. New York: IBM Corporation, 2020. Disponível em: https://www.ibm.com/downloads/cas/7109RLQM. Acesso em: 14 out. 2022.

INDEED. **Indeed**: Job search. 2020. Disponível em: https://www.indeed.com. Acesso em: 22 ago. 2020.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Estatísticas do Cadastro Central de Empresas 2018**. Rio de Janeiro: IBGE, 2020.

INTERNATIONAL FEDERATION OF CLASSIFICATION SOCIETIES. **International Federation of Classification Societies Newsletter**. 1996. Disponível em: http://ifcs.boku.ac.at/site/lib/exe/fetch.php?media=newsletter_archive:ifcs-newsletter-13.pdf. Acesso em: 3 set. 2020.

ISCHOOLS. **ISchools**: Leading and Promoting the Information Field. 2020. Disponível em: https://ischools.org. Acesso em: 22 ago. 2020.

JAPIASSU, H. **Interdisciplinaridade e patologia do saber**. Rio de Janeiro: Imago editora, 1976.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255–260, 2015.

JULIA PROJECT. **The Julia Programming Language**. 2021. Disponível em: https://julialang.org/. Acesso em: 14 out. 2022.

KAGGLE. **Kaggle**: Your Machine Learning and Data Science Community. São Francisco: Kaggle, 2020. Disponível em: https://www.kaggle.com. Acesso em: 24 ago. 2020.

KAGGLE. **Kaggle's State of Data Science and Machine Learning 2019**. São Francisco: Kaggle, 2019. Disponível em: https://www.kaggle.com/kaggle-survey-2019. Acesso em: 14 out. 2022.

KAGGLE. **State of Machine Learning and Data Science 2021**. São Francisco: Kaggle, 2021. Disponível em: https://www.kaggle.com/kaggle-survey-2021. Acesso em: 14 out. 2022.

KAMPAKIS, S. **The Decision Maker's Handbook to Data Science**: A Guide for Non-Technical Executives, Managers, and Founders. 2. ed. Berkeley, CA: Apress, 2020. E-book. Disponível em: http://link.springer.com/10.1007/978-1-4842-5494-3.

KAUERMANN, G.; SEIDL, T. Data Science: a proposal for a curriculum. International **Journal of Data Science and Analytics**, v. 6, n. 3, p. 195–199, 2018. Disponível em: https://doi.org/10.1007/s41060-018-0113-2. Acesso em: 14 out. 2022.

KELLEHER, J. D.; TIERNEY, B. **Data science**. London: MIT Press, 2018.

KENDALL, M. G. **The advanced theory of statistics**. London: Charles Griffin & Company, 1945.

KIM, J. Y.; LEE, C. K. An empirical analysis of requirements for data scientists using online job postings. **International Journal of Software Engineering and its Applications**, v. 10, n. 4, p. 161–172, 2016.

KOENIGSFELD, J. P. et al. Developing a competency model for private club managers. **International Journal of Hospitality Management**, v. 31, n. 3, p. 633–641, 2012.

KOLASSA, S. **The New Data Scientist Venn Diagram**. 2014. Disponível em: https://datascience.stackexchange.com/questions/2403/data-science-without-knowledge-of-a-specific-topic-is-it-worth-pursuing-as-a-ca/2406#2406. Acesso em: 13 set. 2020.

KORKMAZ, S.; GOKSULUK, D.; ZARARSIZ, G. MVN: An R package for assessing multivariate normality. **R Journal**, v. 6, n. 2, p. 151–162, 2014.

LANTZ, B. **Machine Learning with R**. 2. ed. Birmingham: Packt Publishing, 2015.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015.

LEEK, J. **The key word in "Data Science" is not Data, it is Science**. 2013. Disponível em: https://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/. Acesso em: 1 jul. 2020.

LEY, C.; BORDAS, S. P. A. What makes Data Science different? A discussion involving Statistics2.0 and Computational Sciences. **International Journal of Data Science and Analytics**, v. 6, n. 3, p. 167–175, 2018. Disponível em: https://doi.org/10.1007/s41060-017-0090-x. Acesso em: 14 out. 2022.

LI, C. The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. **Psychological Methods**, v. 21, n. 3, p. 369–387, 2016. Disponível em: http://doi.apa.org/getdoi.cfm?doi=10.1037/met0000093. Acesso em: 14 out. 2022.

LIMESURVEY. **LimeSurvey Community Edition**. 2020. Disponível em: https://community.limesurvey.org/. Acesso em: 14 out. 2022.

LINDEN, A. et al. **Staffing Data Science Teams**: Map Capabilities to Key Roles. 2018. Disponível em: https://www.gartner.com/en/documents/3888468. Acesso em: 1 set. 2020.

LINDEN, A. et al. **Staffing Data Science Teams**. 2015. Disponível em: https://www.gartner.com/en/documents/3086717/staffing-data-science-teams. Acesso em: 1 set. 2020.

LIU, Z. et al. A competency model for clinical physicians in China: A cross-sectional survey. **PLoS ONE**, v. 11, n. 12, p. 1–17, 2016.

LOUKIDES, M. **What is data science?** The future belongs to the companies and people that turn data into products. Sebastopol: O'Reilly Media, Inc., 2012. E-book. Disponível em: https://www.oreilly.com/data/free/files/what-is-data-science.pdf. Acesso em: 14 out. 2022.

LOUKIDES, M.; MASON, H.; PATIL, D. J. **Ethics and Data Science**. Sebastopol: O'Reilly Media, Inc., 2018.

MABEY, B. **PyLDAvis Documentation**. 2018. Disponível em: https://pyldavis.readthedocs.io/_/downloads/en/stable/pdf/. Acesso em: 1 jun. 2021.

MALHOTRA, N. K. **Pesquisa de Marketing**: Uma Orientação Aplicada. 7. ed. Porto Alegre: Bookman, 2019.

MALINSKY, G. **93% of employers want to see soft skills on your resume**: here are 8 of the most in-demand ones. 2022. Disponível em: https://www.cnbc.com/2022/07/13/in-demand-soft-skills-to-put-in-your-resume.html. Acesso em: 14 out. 2022.

MANTELERO, A.; VACIAGO, G. Legal aspects of information science, data science, and Big Data. In: DEHMER, M.; EMMERT-STREIB, F. (org.**). Frontiers in Data Science**. Boca Raton: CRC Press, 2017. p. 1–47.

MARCONI, M. de A.; LAKATOS, E. M. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas, 2003.

MARÔCO, J.; GARCIA-MARQUES, T. Qual a fiabilidade do alfa de Cronbach? **Laboratório de Psicologia**, v. 4, n. 1, p. 65–90, 2006. Disponível em: http://repositorio.ispa.pt/handle/10400.12/133. Acesso em: 14 out. 2022.

MARR, B. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. **Forbes**, p. 1–5, 2018. Disponível em: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-

every-day-the-mind-blowing-stats-everyone-should-read/#1cf0b32860ba%0Ahttps://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-. Acesso em: 14 out. 2022.

MARTINS, G. de A.; THEÓPHILO, C. R. **Metodologia da investigação científica para Ciências Sociais Aplicadas**. 2. ed. São Paulo: Atlas, 2009.

MATSUNAGA, M. How to factor-analyze your data right: do's, don'ts, and how-to's. **International Journal of Psychological Research**, v. 3, n. 1, p. 97–110, 2010. Disponível em: https://revistas.usb.edu.co/index.php/IJPR/article/view/854.

MATTER, U. **Data Science in Business/Computational Social Science in Academia?** 2013. Disponível em: http://giventhedata.blogspot.com/2013/03/data-science-in-businesscomputational.html. Acesso em: 23 ago. 2020.

MAYO, M. **The Data Science Puzzle, Explained**. 2016. Disponível em: https://www.kdnuggets.com/2016/03/data-science-puzzle-explained.html. Acesso em: 8 set. 2020.

MCCLELLAND, D. C. Testing for competence rather than for "intelligence". **The American psychologist**, v. 28, n. 1, p. 1–14, 1973.

MCCRACKIN, M. **What is science?** 2017. Disponível em: https://balticeye.org/en/eutrophication/elemental/what-is-science/. Acesso em: 1 set. 2020.

MEMON, M. A. et al. Sample Size for Survey Research: Review and Recommendations. **Journal of Applied Structural Equation Modeling**, v. 4, n. 2, p. i–xx, 2020.

METWALLI, S. A. **A Learning Path To Becoming a Data Scientist**. 2020. Disponível em: https://towardsdatascience.com/a-learning-path-to-becoming-a-data-scientist-56c5c2e8ae3f. Acesso em: 2 out. 2020.

MEYER, M. A. Healthcare data scientist qualifications, skills, and job focus: A content analysis of job postings. **Journal of the American Medical Informatics Association**, v. 26, n. 5, p. 383–391, 2019.

MICHIGAN INSTITUTE FOR DATA SCIENCE. **Data Science Initiative**. 2015. Disponível em: https://midas.umich.edu/dsi/. Acesso em: 30 jun. 2020.

MINER, D. **Scrape data from any website with 1 Click**. 2021. Disponível em: https://dataminer.io/. Acesso em: 1 jun. 2022.

MORIN, E. **Introdução ao pensamento complexo**. Tradução de Eliane Lisboa. Porto Alegre: Sulina. Porto Alegre: Sulina, 2005.

MORIN, E. Problemas de uma epistemologia complexa. In: **O problema epistemológico da complexidade**. 2. ed. Lisboa: Publicações Europa-América, 2002. p. 134.

NAUR, P. **Concise survey of computer methods**. Sweden: Petrocelli Books, 1974.

NIST BIG DATA PUBLIC WORKING GROUP. **NIST Big Data Interoperability Framework**: Volume 1, Definitions. Concessão: out. 2015.

PACHECO, R. C. D. S.; TOSTA, K. C. B. T.; FREIRE, P. D. S. Interdisciplinaridade vista como um processo complexo de construção do conhecimento: uma análise do Programa de Pós-Graduação EGC / UFSC. **Revista Brasileira de Pós-Graduação**, v. 7, n. 12, p. 136–159, 2010. Disponível em: http://www2.capes.gov.br/rbpg/images/stories/downloads/RBPG/Vol.7_12/7_ARTIGO.pdf. Acesso em: 14 out. 2022.

PANDEY, P. **Geek Girls Rising**: Myth or Reality!. 2019. Disponível em: https://www.kaggle.com/parulpandey/geek-girls-rising-myth-or-reality. Acesso em: 4 set. 2020.

PARKS, D. M. D. **Defining Data Science and Data Scientist**. 2017. 68 f. - University of South Florida, 2017. Disponível em: https://scholarcommons.usf.edu/etd/7014.

PATIL, D. J. Building Data Science Teams. **Science**, p. 1–25, 2011.

PIATETSKY-SHAPIRO, G. **CRISP-DM, still the top methodology for analytics, data mining, or data science projects**. Disponível em: <https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>. Acesso em: 1 set. 2020.

PIETSCH, W. Causation, probability, and all that: Data science as a novel inductive paradigm. In: DEHMER, M.; EMMERT-STREIB, F. (org.). **Frontiers in Data Science**. 1. ed. Boca Raton: CRC Press, 2017. p. 329–353. E-book. Disponível em: https://www.taylorfrancis.com/books/9781498799331/chapters/10.1201/9781315156408-11. Acesso em: 14 out. 2022.

PIETY, P. J.; HICKEY, D. T.; BISHOP, M. J. Educational data sciences - Framing emergent practices for analytics of learning, organizations, and systems. **ACM International Conference Proceeding Series**, p. 193–202, 2014.

POWER, D. J. Data science: supporting decision-making. **Journal of Decision Systems**, v. 25, n. 4, p. 345–356, 2016. Disponível em: http://dx.doi.org/10.1080/12460125.2016.1171610. Acesso em: 14 out. 2022.

PROVOST, F.; FAWCETT, T. Data Science and its Relationship to Big Data and Data-Driven Decision Making. **Big Data**, v. 1, n. 1, p. 51–59, 2013.

PROVOST, F.; FAWCETT, T. **Data Science for Business**: What you need to know about data mining and data-analytic thinking. Sebastopol: O'Reilly Media, Inc., 2013.

PWC. **Investing in America's data science and analytics talent**. London: PricewaterhouseCoopers LLP, 2017. Disponível em: https://www.pwc.com/us/en/publications/assets/investing-in-america-s-dsa-talent-bhef-and-pwc.pdf. Acesso em: 14 out. 2022.

RASCHKA, S. **Python Machine Learning**: Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics. Birmingham: Packt Publishing, 2016.

RAWLINGS-GOSS, R. **Data Science Careers, Training, and Hiring**: A Comprehensive Guide to the Data Ecosystem: How to Build a Successful Data Science Career, Program, or Unit. Los Angeles: Springer, 2019.

ŘEHŮŘEK, R.; SOJKA, P. Gensim-python framework for vector space modelling. **NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic**, v. 3, n. 2, 2011. Disponível em: https://radimrehurek.com/gensim/index.html. Acesso em: 14 out. 2022.

REIS, L. C. R.; SÁ, M. I. da F. e. Big Data: Um novo campo de atuação para bibliotecários. **Prisma.com**, n. 41, p. 231–250, 2020.

ROMERO, C.; VENTURA, S. Educational data science in massive open online courses. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 7, n. 1, 2017.

RUSSOM, P. **TDWI Best Practice Report: Big Data AnalyticsOctober**. Woodland Hills, CA: Transforming Data With Intelligence (TDWI), 2011. Disponível em: http://faculty.ucmerced.edu/frusu/Papers/Conference/2012-sigmod-glade-demo.pdf. Acesso em: 14 out. 2022.

SALTZ, J. S.; GRADY, N. W. The ambiguity of data science team roles and the need for a data science workforce framework. **Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017**, v. 2018-Janua, p. 2355–2361, 2017.

SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, M. del P. B. **Metodologia de pesquisa**. 5. ed. Porto Alegre: Penso, 2013.

SAS INSTITUTE INC. **SAS® Enterprise MinerTM 14.3: Reference HelpSAS**. Cary, NC, USA: SAS Institute Inc., 2017.

SCHOENHERR, T.; SPEIER-PERO, C. Data science, predictive analytics, and big data in supply chain management: Current state and future potential. **Journal of Business Logistics**, v. 36, n. 1, p. 120–132, 2015.

SCHREIBER, J. B. et al. Reporting structural equation modeling and confirmatory factor analysis results: A review. **Journal of Educational Research**, v. 99, n. 6, p. 323–338, 2006.

SEROUSS, Y. **What is Data Science?** 2014. Disponível em: https://yanirseroussi.com/2014/10/23/what-is-data-science/. Acesso em: 5 set. 2020.

SIEBES, A. Data science as a language: challenges for computer science—a position paper. International **Journal of Data Science and Analytics**, v. 6, n. 3, p. 177–187, 2018. Disponível em: https://doi.org/10.1007/s41060-018-0103-4. Acesso em: 14 out. 2022.

SIEVERT, C.; SHIRLEY, K. LDAvis: A method for visualizing and interpreting topics. In: **Proceedings of the workshop on interactive language learning, visualization, and interfaces**. Baltimore: Association for Computational Linguistics, 2014. p. 63–70. Disponível em: https://aclanthology.org/W14-3110.pdf. Acesso em: 14 out. 2022.

SILVA, E. L. da; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4. ed. Florianópolis: UFSC, 2005-. ISSN 1517-9702. Disponível em: http://www.mendeley.com/research/metodologia-da-pesquisa-e-elaborao-de-dissertao-4a-edio-revisada-e-atualizada/. Acesso em: 14 out. 2022.

SINTEF. **Big Data, for better or worse**: 90% of world's data generated over last two years. 2013. Disponível em: https://www.sciencedaily.com/releases/2013/05/130522085217.htm. Acesso em: 14 out. 2022.

SMITH, M. J. de. **Statistical Analysis Handbook**: A comprehensive handbook of statistical concepts, techniques and software tools. London: The Winchelsea Press, 2014. E-book. Disponível em: https://www.statsref.com/StatsRefSample.pdf. Acesso em: 14 out. 2022.

SOUZA, A. C. de; ALEXANDRE, N. M. C.; GUIRARDELLO, E. de B. Psychometric properties in instruments evaluation of reliability and validity. **Epidemiologia e servicos de saude: revista do Sistema Unico de Saude do Brasil**, v. 26, n. 3, p. 649–659, 2017.

STARK, H.; HAWAMDEH, S. Relating Big Data and Data Science to the Wider Concept of Knowledge Management. In: **Analytics And Knowledge Management**. Boca Raton: CRC Press, 2018. p. 141–166.

STATISTA. **Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes)**. 2022. Disponível em: https://www.statista.com/statistics/871513/worldwide-data-created/. Acesso em: 14 out. 2022.

STODDER, D. **BI and Analytics in the Age of AI and Big DataTransforming Data With Intelligence (TDWI) Best Practice Report, Fourth Quarter**. Woodland Hills, CA: Transforming Data With Intelligence (TDWI), 2018.

STODDER, D. **Chasing the Data Science Unicorn**. 2015. Disponível em: https://tdwi.org/articles/2015/01/06/chasing-the-data-science-unicorn.aspx. Acesso em: 20 set. 2020.

SUMATHI, S.; SIVANANDAM, S. N. **Introduction to Data Mining and its Applications**. Berlin: Springer, 2006.

TAYLOR, D. **Battle of the Data Science Venn Diagrams**. 2016. Disponível em: https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html. Acesso em: 10 jun. 2020.

THODE, H. C. **Testing For Normality**. Boca Raton: CRC Press, 2002. E-book. Disponível em: https://www.taylorfrancis.com/books/9780203910894. Acesso em: 14 out. 2022.

TIERNEY, B. **Data Science Is Multidisciplinary**. 2012. Disponível em: https://oralytics.com/2012/06/13/data-science-is-multidisciplinary/. Acesso em: 24 ago. 2020.

TOSIC, P. T.; BEESTON, J. Designing undergraduate data science curricula: A computer science perspective. **ASEE Annual Conference and Exposition, Conference Proceedings**, v. 2018-June, 2018.

UDEMY. **Learn about Udemy culture, mission, and careers**. 2022. Disponível em: https://about.udemy.com/. Acesso em: 1 jun. 2022.

VENN, J. I. On the diagrammatic and mechanical representation of propositions and reasonings. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v. 10, n. 59, p. 1–18, 1880.

VIEIRA, S. **Como escrever uma tese**. 6. ed. São Paulo: Atlas, 2008.

WALKER, M. A. The professionalisation of data science. **International Journal of Data Science**, v. 1, n. 1, p. 7, 2015.

WALLER, M. A.; FAWCETT, S. E. Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. **Journal of Business Logistics**, v. 34, n. 2, p. 77–84, 2013.

WANG, Y. F. Constructing career competency model of hospitality industry employees for career success. **International Journal of Contemporary Hospitality Management**, v. 25, n. 7, p. 994–1016, 2013.

WASHINGTON DURR, A. K. **A Text Analysis of Data-Science Career Opportunities and US iSchool Curriculum**. 2018. 114 f. - University of North Texas, 2018. Disponível em: https://digital.library.unt.edu/ark:/67531/metadc1404565/. Acesso em: 14 out. 2022.

WESSLEN, R. **Computer-Assisted Text Analysis for Social Science**: Topic Models and Beyond. 2018. Disponível em: http://arxiv.org/abs/1803.11045. Acesso em: 14 out. 2022.

WHITTAKER, Z. **Web scraping is legal, US appeals court reaffirms**. 2022. Disponível em: https://techcrunch.com/2022/04/18/web-scraping-legal-court/. Acesso em: 18 abr. 2022.

WIKIPEDIA.ORG. **Ciência de Dados**. 2020. Disponível em: https://pt.wikipedia.org/wiki/Ciência_de_dados. Acesso em: 8 set. 2020.

WIKIPEDIA.ORG. **Data Science**. 2020. Disponível em: https://en.wikipedia.org/wiki/Data_science. Acesso em: 8 set. 2020.

WOLFRAM, D. A pesquisa bibliométrica na era do big data: Desafios e oportunidades. **Bibliometria e cientometria no Brasil: infraestrutura para avaliação da pesquisa científica na era do Big Data**, 2017.

WU, C. F. J. **Statistics = Data Science?** Michigan: University of Michigan, 1997. Disponível em: https://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf. Acesso em: 14 out. 2022.

XIA, Y.; YANG, Y. RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. **Behavior Research Methods**, v. 51, n. 1, p. 409–428, 2019.

APÊNDICE 1– COMPETÊNCIAS DO QUESTIONÁRIO APLICADO

(continua)

Grupo	Competência	Autores
Tecnologia	Algoritmos	Harris, Murphy e Vaisman (2013), Demchenko, Belloum e Wiktorski (2017a), Hayes (2017), Parks (2017), Stodder (2018), Meyer (2019).
	Banco de Dados	Dhar (2013), Cao (2017), Hayes (2017), Kelleher e Tierney (2018), Demchenko et al. (2019).
	Dados semi estruturados	Dhar (2013), Harris, Murphy e Vaisman (2013), Eubanks (2016), Hayes (2017), Parks (2017), ESCO (2020b).
	Dados não-estruturados	Dhar (2013), Harris, Murphy e Vaisman (2013), Granville (2014), Chen et al. (2016), Hayes (2017), IBM (2020).
	Desenvolvimento de software	Patil (2011), Harris, Murphy e Vaisman (2013), Geringer (2014), NIST (2015), Cao (2017), Rawlings-Goss (2019).
	Engenharia de Dados	Eubanks (2016), Demchenko, Belloum e Wiktorski (2017a), Linden (2018), Cao (2019).
	Gestão de Dados	Loukides (2012), Tierney (2012), Harris, Murphy e Vaisman (2013), Linden (2015, 2018), Eubanks (2016), Baškarada e Koronios (2017), Cao (2017, 2019), Demchenko, Belloum e Wiktorski (2017a), Hayes (2017), Parks (2017), Rawlings-Goss (2019), ESCO (2020b).
	Habilidades em Hacking	Conway (2010), Eubanks (2016), Parks (2017), Hawamdeh e Chang (2018), Kampakis (2020).
	Inteligência Artificial/Machine Learning	Conway (2010), Tierney (2012), Dhar (2013), Harris, Murphy e Vaisman (2013), Geringer (2014), NIST (2015), Eubanks (2016), Cao (2017), Demchenko, Belloum e Wiktorski (2017a), Hayes (2017), Parks (2017), Kelleher e Tierney (2018), Linden (2018), Kampakis (2020).
	Programação	Harris, Murphy e Vaisman (2013), Kolassa (2014), Davenport (2015), Linden (2015), Eubanks (2016), Cao (2017, 2019), Hayes (2017), Parks (2017), Burtch (2019).
	Sistemas de informação	Harris, Murphy e Vaisman (2013), Linden (2015), NIST (2015), Cao (2017), Demchenko, Belloum e Wiktorski (2017a), Hayes (2017).
	Sistemas de dados distribuídos	Dhar (2013), Harris, Murphy e Vaisman (2013), Cao (2017), Hayes (2017), Saltz e Grady (2017), Kelleher e Tierney (2018).

(continua)

Grupo	Competência	Autores
Análise de Dados	Análise de dados	Serouss (2014), Davenport (2015), NIST (2015), Demchenko (2016), Cao (2017, 2019), Parks (2017), Linden (2018), Burtch (2019), Rawlings-GOSS (2019), Kampakis (2020).
	Análises preditivas	Harris, Murphy e Vaisman (2013), Eubanks (2016), Baškarada e Koronios (2017), Hayes (2017), Burtch (2019).
	Estatística	Conway (2010), Tierney (2012), Dhar (2013), Harris, Murphy e Vaisman (2013), Geringer (2014), Kolassa (2014), NIST (2015), Eubanks (2016), Kim e Lee (2016), Baškarada e Koronios (2017), Cao (2017), Hayes (2017), Parks (2017), Kelleher e Tierney (2018), Linden (2018), ESCO (2020b), Kampakis (2020).
	Matemática	Conway (2010), Dhar (2013), Harris, Murphy e Vaisman (2013), Geringer (2014), Eubanks (2016), Hall, Phan e Whitson (2016), Baškarada e Koronios (2017), Cao (2017), Hayes (2017), Parks (2017), Linden (2018).
	Mineração de dados	Tierney (2012), Walker (2015), Baškarada e Koronios (2017), Cao (2017), Demchenko, Belloum e Wiktorski (2017a), Hayes (2017), ESCO (2020b).
	Modelagem em grafos	Harris, Murphy e Vaisman (2013), Hayes (2017)
	Otimização	Dhar (2013), Harris, Murphy e Vaisman (2013), Hayes (2017).
	Processamento de Linguagem Natural	Demchenko, Belloum e Wiktorski (2017a), Hayes (2017), Parks (2017).
	Formular Questões/Problemas	Dhar (2013), Saltz e Grady (2017), Linden (2018), Cao (2019).
	Método científico	Tierney (2012), Harris, Murphy e Vaisman (2013), Baškarada e Koronios (2017), Cao (2017), Hayes (2017), Parks (2017), Kelleher e Tierney (2018).
	Visualização de Dados	Conway (2010), Harris, Murphy e Vaisman (2013), Matter (2013), Geringer (2014), Donoho (2015), Demchenko et al. (2016), Eubanks (2016), Cao (2017, 2019), Demchenko, Belloum e Wiktorski (2017a), Hayes (2017), Rawlings-Goss (2019).
	Técnicas de Regressão	Hardin et al. (2015), Demchenko et al. (2019), IBM (2020), Gottipati et al. (2021).

(continua)

Grupo	Competência	Autores
Entendimento de Negócios	Conhecimento de Domínio	Conway (2010), Tierney (2012), Harris, Murphy e Vaisman (2013), Geringer (2014), Linden (2015, 2018), NIST (2015), Eubanks (2016), Hall, Phan e Whitson (2016), Kim e Lee (2016), Baškarada e Koronios (2017), Cao (2017), Demchenko, Belloum e Wiktorski (2017a), Hayes (2017), Kelleher e Tierney (2018).
	Desenvolvimento de produto	Patil (2011), Harris, Murphy e Vaisman (2013), Hayes (2017), PwC (2017).
	Gestão de projetos	Kim e Lee (2016), Hayes (2017), Durr (2018), Cao (2019).
	Desenvolvimento de novos negócios	Tierney (2012), Harris, Murphy e Vaisman (2013), Provost e Fawcett (2013a), Kolassa (2014), NIST (2015), Hayes (2017), Hawamdeh e Chang (2018), Stodder (2018), Linden (2018), Demchenko et al. (2019), Rawlings-Goss (2019).
	Planejamento financeiro	Harris, Murphy e Vaisman (2013), Parks (2017), Stodder (2018), Demchenko et al. (2019).
	Governança de dados	Eubanks (2016), Hayes (2017), Hawamdeh e Chang (2018), Kelleher e Tierney (2018).
	Compliance	Eubanks (2016), Hayes (2017), Mantelero e Vaciago (2017), Parks (2017), Kelleher e Tierney (2018), Demchenko et al. (2019).
	LGPD	Mantelero e Vaciago (2017), Demchenko et al. (2019), Rawlings-Goss (2019).

(conclusão)

Grupo	Competência	Autores
Competências Socioculturais	Conhecimento interdisciplinar	Loukides (2012), Cao (2017, 2019), Parks (2017).
	Comunicação	Patil (2011), Loukides (2012), Tierney (2012), Kolassa (2014), Davenport (2015), Linden (2015, 2018), Schoenherr e Speier-Pero (2015), Eubanks (2016), Kim e Lee (2016), Baškarada e Koronios (2017), Cao (2017, 2019), Demchenko, Belloum e Wiktorski (2017a), Hayes (2017), Parks (2017), Saltz e Grady (2017), Kelleher e Tierney (2018), Rawlings-Goss (2019).
	Liderança técnica	Linden (2015, 2018), Cao (2019), Rawlings-Goss (2019).
	Liderança estratégica	Linden (2015, 2018), Cao (2019), Rawlings-Goss (2019).
	Solução de problemas	Loukides (2012), Tierney (2012), Granville (2014), Cao (2017, 2019), Rawlings-Goss (2019)
	Colaboração	Harris, Murphy e Vaisman (2013), Linden (2018), Cao (2019).
	Criatividade	Patil (2011), Harris, Murphy e Vaisman (2013), Linden (2015), Cao (2017).
	Curiosidade	Patil (2011), Tierney (2012), Linden (2015), Power (2016), Parks (2017), Linden (2018).
	Ética	Anderson et al. (2014), Piety, Hickey e Bishop (2014), Kelleher e Tierney (2018), Cao (2019).
	Dados Sensíveis	DAMA International Technics (2017), Stodder (2018), Franzke et al. (2019), Forgó et al. (2021).
	Bias	Provost e Fawcett (2013a), Data Science Association (2014), Hardin et al. (2015), Chen et al. (2016),

APÊNDICE 2– QUESTIONÁRIO APLICADO

PESQUISA DE DOUTORADO

Competências em Ciência de Dados

Pesquisadores responsáveis: André José Ribeiro Guimarães (Doutorando, andrejose@ufpr.br), Prof.ª Dr.ª Maria do Carmo Duarte Freitas (Orientadora) e Prof. Dr. Ricardo Mendes Junior (Coorientador).

Este instrumento de coleta de dados é parte da pesquisa de doutorado, vinculada ao Programa de Pós-Graduação em Gestão da Informação (PPGGI) da Universidade Federal do Paraná (UFPR), que investiga as competências do(a) cientista de dados no contexto educacional e profissional brasileiro.

Objetivo deste questionário: identificar as competências dos(as) cientistas de dados que atuam no Brasil.

Método: Por meio de um levantamento bibliográfico, foram definidos quatro grupos de competências associadas a este profissional: a) **Tecnologia**, b) **Análise de Dados**, c) **Entendimento de Negócios** e d) **Competências Sociais**. Estes grupos correspondem às primeiras seções do questionário, onde você indicará seu nível de proficiência conforme a escala:

0 a 2	>2 a 4	>4 a 6	>6 a 8	>8 a 10
Consciente Você sabe da existência ou possui um entendimento básico de técnicas e conceitos.	**Novato(a)** Você tem o nível de experiência obtido em sala de aula e/ou cenários experimentais ou como estagiário no trabalho. Espera-se que você precise de ajuda ao executar tarefas relacionadas a esta competência.	**Intermediário(a)** Você conclui com êxito tarefas nesta competência, conforme solicitado. Eventualmente, solicita a ajuda de um especialista. Porém, geralmente, você consegue executar este tipo de demanda de forma independente.	**Avançado(a)** Você executa as ações associadas a esta competência sem assistência. Na sua organização, você certamente é reconhecido como "uma pessoa para perguntar" quando surgem perguntas difíceis sobre este tópico.	**Especialista** Você é conhecido como um especialista neste tópico. Você fornece orientação, soluciona problemas e responde a perguntas relacionadas a essa área de especialização e ao campo em que a competência é utilizada.

A penúltima seção também aborda as competências sociais, porém sob uma perspectiva organizacional. Por fim, a última seção apresenta questões adicionais de caracterização que busca coletar dados educacionais, sociais e profissionais.

Público alvo: Profissionais atuantes no Brasil cuja principal demanda profissional é **lidar com dados**, independentemente do nível de experiência, setor ou função.

Contrapartida: Em troca da sua participação, você receberá o relatório executivo que destacará as principais conclusões da pesquisa. Para isso, basta fornecer seu endereço de e-mail no final do questionário. Ademais, este relatório, que destacará suas competências frente aos demais profissionais do mercado, servirá como ferramenta de autoavaliação para seu desenvolvimento profissional na área de dados. Por fim, outro benefício aos interessados pela área é a disponibilização dos dados, formatados e anonimizados, na Base de Dados Científicos da Universidade Federal do Paraná (BDC/UFPR), possibilitando novas análises e abordagens independentes, promovidas pela comunidade.

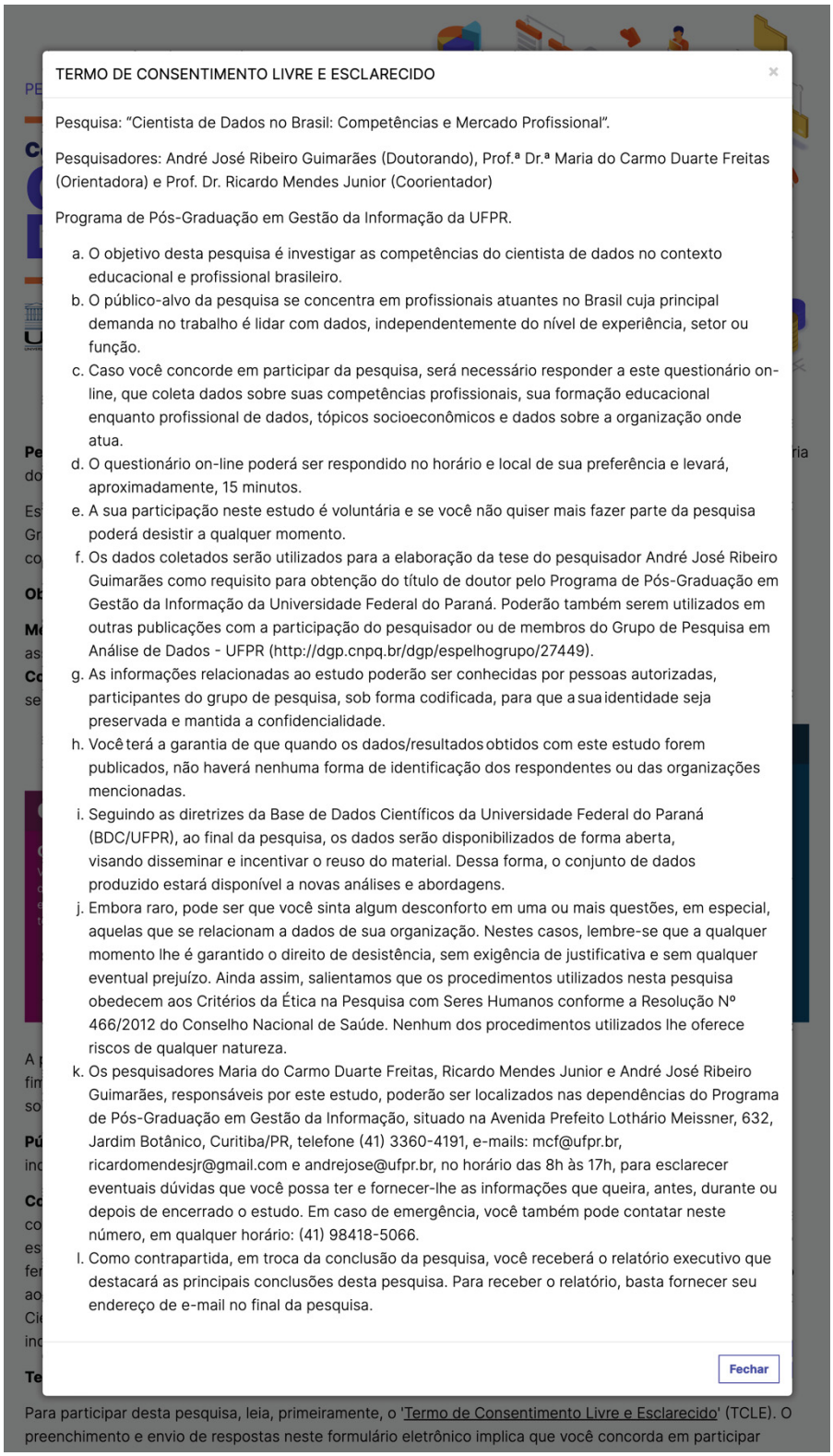
Tempo estimado para conclusão: 20 minutos

Para participar desta pesquisa, leia, primeiramente, o 'Termo de Consentimento Livre e Esclarecido' (TCLE). O preenchimento e envio de respostas neste formulário eletrônico implica que você concorda em participar voluntariamente da pesquisa e manifesta pleno entendimento das condições descritas no TCLE. Caso se sinta desconfortável ou inseguro para participar, fique à vontade em deixar questões não-obrigatórias em branco ou mesmo interromper o preenchimento do formulário.

Desde já, agradecemos sua colaboração em completar todo o processo, contribuindo para o melhor resultado da pesquisa.

⊙ Carregar questionário não finalizado

Próximo >



TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Pesquisa: "Cientista de Dados no Brasil: Competências e Mercado Profissional".

Pesquisadores: André José Ribeiro Guimarães (Doutorando), Prof.ª Dr.ª Maria do Carmo Duarte Freitas (Orientadora) e Prof. Dr. Ricardo Mendes Junior (Coorientador)

Programa de Pós-Graduação em Gestão da Informação da UFPR.

a. O objetivo desta pesquisa é investigar as competências do cientista de dados no contexto educacional e profissional brasileiro.
b. O público-alvo da pesquisa se concentra em profissionais atuantes no Brasil cuja principal demanda no trabalho é lidar com dados, independentemente do nível de experiência, setor ou função.
c. Caso você concorde em participar da pesquisa, será necessário responder a este questionário on-line, que coleta dados sobre suas competências profissionais, sua formação educacional enquanto profissional de dados, tópicos socioeconômicos e dados sobre a organização onde atua.
d. O questionário on-line poderá ser respondido no horário e local de sua preferência e levará, aproximadamente, 15 minutos.
e. A sua participação neste estudo é voluntária e se você não quiser mais fazer parte da pesquisa poderá desistir a qualquer momento.
f. Os dados coletados serão utilizados para a elaboração da tese do pesquisador André José Ribeiro Guimarães como requisito para obtenção do título de doutor pelo Programa de Pós-Graduação em Gestão da Informação da Universidade Federal do Paraná. Poderão também serem utilizados em outras publicações com a participação do pesquisador ou de membros do Grupo de Pesquisa em Análise de Dados - UFPR (http://dgp.cnpq.br/dgp/espelhogrupo/27449).
g. As informações relacionadas ao estudo poderão ser conhecidas por pessoas autorizadas, participantes do grupo de pesquisa, sob forma codificada, para que a sua identidade seja preservada e mantida a confidencialidade.
h. Você terá a garantia de que quando os dados/resultados obtidos com este estudo forem publicados, não haverá nenhuma forma de identificação dos respondentes ou das organizações mencionadas.
i. Seguindo as diretrizes da Base de Dados Científicos da Universidade Federal do Paraná (BDC/UFPR), ao final da pesquisa, os dados serão disponibilizados de forma aberta, visando disseminar e incentivar o reuso do material. Dessa forma, o conjunto de dados produzido estará disponível a novas análises e abordagens.
j. Embora raro, pode ser que você sinta algum desconforto em uma ou mais questões, em especial, aquelas que se relacionam a dados de sua organização. Nestes casos, lembre-se que a qualquer momento lhe é garantido o direito de desistência, sem exigência de justificativa e sem qualquer eventual prejuízo. Ainda assim, salientamos que os procedimentos utilizados nesta pesquisa obedecem aos Critérios da Ética na Pesquisa com Seres Humanos conforme a Resolução Nº 466/2012 do Conselho Nacional de Saúde. Nenhum dos procedimentos utilizados lhe oferece riscos de qualquer natureza.
k. Os pesquisadores Maria do Carmo Duarte Freitas, Ricardo Mendes Junior e André José Ribeiro Guimarães, responsáveis por este estudo, poderão ser localizados nas dependências do Programa de Pós-Graduação em Gestão da Informação, situado na Avenida Prefeito Lothário Meissner, 632, Jardim Botânico, Curitiba/PR, telefone (41) 3360-4191, e-mails: mcf@ufpr.br, ricardomendesjr@gmail.com e andrejose@ufpr.br, no horário das 8h às 17h, para esclarecer eventuais dúvidas que você possa ter e fornecer-lhe as informações que queira, antes, durante ou depois de encerrado o estudo. Em caso de emergência, você também pode contatar neste número, em qualquer horário: (41) 98418-5066.
l. Como contrapartida, em troca da conclusão da pesquisa, você receberá o relatório executivo que destacará as principais conclusões desta pesquisa. Para receber o relatório, basta fornecer seu endereço de e-mail no final da pesquisa.

Fechar

Para participar desta pesquisa, leia, primeiramente, o 'Termo de Consentimento Livre e Esclarecido' (TCLE). O preenchimento e envio de respostas neste formulário eletrônico implica que você concorda em participar

Competências em
Ciência de Dados

Retomar mais tarde | Sair e apagar o questionário

TECNOLOGIA | ANÁLISE DE DADOS | ENTENDIMENTO DE NEGÓCIOS | COMPETÊNCIAS SOCIAIS | ORGANIZAÇÃO | INFORMAÇÕES ADICIONAIS

Tecnologia para Ciência de Dados

* Nesta seção, você será questionado(a) sobre sua proficiência em tópicos relacionados à **Tecnologia** para a Ciência de Dados.

Por favor, utilize a seguinte escala ao indicar seu nível de proficiência:

PONTOS IMPORTANTES:

- caso não tenha conhecimento acerca de um item específico, selecione o valor 0 (zero);
- seu nível de proficiência é relativo ao item específico perguntado e não ao conjunto "Tecnologia".

Pensando em suas competências atuais relativas a **Tecnologia**, indique quão proficiente você é para cada item listado abaixo:

Algoritmos (por exemplo, complexidade computacional, teoria da ciência da computação).

Administração de banco de dados.

Manipulação de dados semiestruturados (por exemplo, JSON, XML, RDF).

Manipulação de dados não estruturados (por exemplo, imagens, áudios, textos).

Desenvolvimento de softwares (planejar, executar, testar, implantar e manter sistemas de dados).

Engenharia de dados (projetar, desenvolver e gerenciar a infraestrutura para projetos em dados).

Gestão de dados (coleta, preparação, limpeza, armazenamento, proteção, integração de fontes de dados diferentes).

Hacking (habilidades técnicas e lógicas para construir soluções rápidas sem necessariamente possuir os fundamentos da Ciência da Computação).

Inteligência artificial (por exemplo, machine learning, deep learning, árvores de decisão, redes neurais, SVM).

Programação voltada a projetos de dados, seja back-end, front-end ou full-stack.

Administração de sistemas de informação.

Sistemas de dados distribuídos de alto desempenho (por exemplo, Hadoop, MapReduce, Spark).

Anterior

Próximo

Competências em
Ciência de Dados

Retomar mais tarde | Sair e apagar o questionário

TECNOLOGIA | ANÁLISE DE DADOS | ENTENDIMENTO DE NEGÓCIOS | COMPETÊNCIAS SOCIAIS | ORGANIZAÇÃO | INFORMAÇÕES ADICIONAIS

Análise para Ciência de Dados

* Agora, serão apresentados alguns itens relacionados à disciplina de **Análise de Dados**.

Novamente, pedimos que indique sua proficiência, utilizando a mesma escala:

Pensando em suas competências atuais relativas a **Análise de Dados**, indique quão proficiente você é para cada item listado abaixo:

Análise de dados, de forma geral.
ARRASTE 5
0 10

Análises preditivas (tais como otimização de campanhas de marketing, detecção de fraude, redução de risco).
ARRASTE 5
0 10

Estatísticas e modelagem estatística (por exemplo, modelo linear geral, ANOVA, MANOVA, modelo espaço-temporal, geoestatística).
ARRASTE 5
0 10

Matemática (tópicos como álgebra linear, análise real, cálculo, dentre outros).
ARRASTE 5
0 10

Técnicas de mineração de dados (por exemplo, associação, classificação).
ARRASTE 5
0 10

Modelagem em grafos (por exemplo, relacionamentos em redes sociais, definição de rotas, dentre outros).
ARRASTE 5
0 10

Otimização (por exemplo, linear, inteira, convexa, global).

Processamento de linguagem natural (NLP) e mineração de texto.

Formulação de questões de negócio que possam ser respondidas por meio de técnicas analíticas.

Método científico (por exemplo, desenho experimental, projeto de pesquisa).

Visualização de dados (por exemplo, gráficos, mapas, visualização baseada na web, dentre outros).

Técnicas de regressão (linear, múltipla, logística).

Anterior

Próximo

Competências em
Ciência de Dados

Retomar mais tarde | Sair e apagar o questionário

TECNOLOGIA | ANÁLISE DE DADOS | ENTENDIMENTO DE NEGÓCIOS | COMPETÊNCIAS SOCIAIS | ORGANIZAÇÃO | INFORMAÇÕES ADICIONAIS

Entendimento de Negócios para Ciência de Dados

* A seguir, são questionados tópicos de **conhecimentos específicos relativos a negócios em sua área de atuação**.

A escala para indicar sua proficiência continua a mesma:

Pensando em suas competências atuais relativas a **Entendimento de Negócios**, indique quão proficiente você é para cada item listado abaixo:

Conhecimento em domínios específicos relativos à sua área de atuação (por exemplo, assistência médica, finanças, educação, imobiliário, automotivo) relevante para o seu trabalho.

ARRASTE 5

0 10

Projeto e desenvolvimento de novos produtos.

ARRASTE 5

0 10

Gestão de Projetos em Ciência de Dados.

ARRASTE 5

0 10

Desenvolvimento de novos negócios (capacidade de empreender).

ARRASTE 5

0 10

Planejamento financeiro para projetos em Ciência de Dados.

ARRASTE 5

0 10

Governança de dados (definição de estruturas organizacionais, políticas e processos relacionados a dados).

ARRASTE 5

0 10

Garantia de conformidade (compliance) com leis, normativas e demais regulamentações vigentes.

ARRASTE 5

0 10

Atendimento integral à LGPD.

ARRASTE 5

0 10

< Anterior | Próximo >

Competências em
Ciência de Dados

Retomar mais tarde | Sair e apagar o questionário

TECNOLOGIA | ANÁLISE DE DADOS | ENTENDIMENTO DE NEGÓCIOS | COMPETÊNCIAS SOCIAIS | ORGANIZAÇÃO | INFORMAÇÕES ADICIONAIS

Competências Sociais

* As perguntas a seguir se referem às **competências sociais** relacionadas à Ciência de Dados.

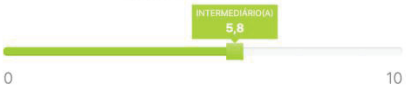
Pensando em suas competências atuais relativas a **Competências Sociais**, indique quão proficiente você é para cada item listado abaixo:

Conhecimento interdisciplinar que permite lidar com todos os aspectos de um problema referente a dados, da coleta às conclusões, interagindo com profissionais de diversas áreas.

Comunicação (por exemplo, compartilhamento de resultados a público não-técnico, comunicação oral e escrita, apresentações, redação de artigos).

Liderança técnica.

Liderança estratégica.



Competência para solucionar qualquer tipo de problema de dados.

< Anterior | Próximo >

Competências em
Ciência de Dados

Retomar mais tarde | Sair e apagar o questionário

TECNOLOGIA | ANÁLISE DE DADOS | ENTENDIMENTO DE NEGÓCIOS | COMPETÊNCIAS SOCIAIS | ORGANIZAÇÃO | INFORMAÇÕES ADICIONAIS

Competências Sociais na **Organização**

* Agora, as perguntas se referem às competências sociais relacionadas à Ciência de Dados na sua organização. Caso trabalhe de forma autônoma, considere sua rede profissional de relacionamento.

Para cada afirmação abaixo, indique o seu grau de concordância em uma escala de 0 a 10, onde "0" significa "Discordo totalmente" e "10" significa "Concordo totalmente", deslizando (arrastando) o marcador pela régua.

0 = DISCORDO TOTALMENTE
10 = CONCORDO TOTALMENTE

A colaboração é uma característica visível entre os membros das equipes de dados.

Cientistas de dados são pessoas criativas que buscam abordagens inovadoras para extrair significado dos dados.

A curiosidade e inquietação perante as possibilidades são atributos necessários para se trabalhar com Ciência de Dados.

A ética é princípio fundamental em todos os projetos que envolvem dados.

Há constante preocupação com o tratamento de dados sensíveis.

O risco de análises tendenciosas (bias) é foco contínuo de atenção.

< Anterior | Próximo >

Competências em
Ciência de Dados

Retomar mais tarde | Sair e apagar o questionário

TECNOLOGIA | ANÁLISE DE DADOS | ENTENDIMENTO DE NEGÓCIOS | COMPETÊNCIAS SOCIAIS | ORGANIZAÇÃO | INFORMAÇÕES ADICIONAIS

Informações **adicionais**

* Na sua opinião, quais os **três termos** que melhor descrevem a Ciência de Dados?

Considere termo como uma palavra ou um conjunto de palavras como, por exemplo, "conhecimento" e "gestão do conhecimento".

Termo 1

Termo 2

Termo 3

Informações pessoais

* Qual sua **idade** em anos?

* Qual seu **gênero**?

- ○ Feminino
- ○ Masculino
- ○ Prefiro não dizer
- ○ Outros:

* Em qual **estado** você mora atualmente?

Por favor, selecione...

* Em qual **cidade** você mora atualmente?

Informações educacionais

* Qual é o nível mais alto de **educação formal** que você atingiu?

Por favor, selecione...

* Em que curso fez ou está fazendo **graduação**? Indique, por favor, a instituição educacional correspondente.

Caso nunca tenha se matriculado em um curso de graduação, por favor, preencha com "NA" referente a "não se aplica".

Se já fez ou está fazendo alguma **pós-graduação** (especialização, mestrado, doutorado ou pós-doutorado), indique os cursos e as instituições educacionais correspondentes.

Se já fez alguma **formação** em Ciência de Dados (*Data Science*), por favor, indique o nome do curso e a instituição educacional:

Informações profissionais

* Selecione o título que mais se assemelha à sua **função atual**.
Caso sua função atual não se encaixe em nenhuma das opções pré-estabelecidas, utilize a opção "Outros", descrevendo sucintamente sua situação profissional em relação à Ciência de Dados.

- Analista de BI
- Analista de Dados
- Analista de Negócios
- Cientista de Dados
- DBA/Engenheiro de Banco de Dados
- Engenheiro de Dados
- Engenheiro de Software
- Estatístico
- Matemático
- Gerente de Projeto/Produto
- Outros:

* Qual a principal **metodologia** que você utiliza em seus projetos de análise, mineração de dados ou ciência de dados?

- CRISP-DM
- SEMMA
- KDD
- Metodologia desenvolvida por mim
- Metodologia da minha organização
- Metodologia voltada ao segmento que atuo
- Nenhuma
- Outros:

* **Quantos anos** de experiência você possui em analisar dados, seja trabalhando ou estudando?

* Qual é a sua **remuneração** mensal atual em Reais?

- ○ Até 2 salários mínimos (até R$ 2.200,00).
- ○ De 2 a 4 salários mínimos (de R$ 2.201,00 até R$ 4.400,00).
- ○ De 4 a 6 salários mínimos (de R$ 4.401,00 até R$ 6.600,00).
- ○ De 6 a 8 salários mínimos (de R$ 6.601,00 até R$ 8.800,00).
- ○ De 8 a 10 salários mínimos (de R$ 8.801,00 até R$ 11.000,00).
- ○ Acima de 10 salários mínimos (mais de R$ 11.001,00).
- ○ Nenhuma remuneração
- ○ Prefiro não responder

Caracterização da empresa

* Em qual **segmento** (área de atividade) melhor se encaixa a organização na qual trabalha?
Se atua de forma autônoma, por favor, apresente os principais segmentos atendidos. E caso atualmente não realize projetos profissionais na área, por favor, preencha com "NA" referente a "não se aplica".

* Qual o **número total de pessoas** que colaboram (celetistas, estagiários, terceirizados etc.) com a organização em 2021?

- ○ Trabalho como autônomo(a)
- ○ Até 4 pessoas
- ○ 5 a 9 pessoas
- ○ 10 a 19 pessoas
- ○ 20 a 29 pessoas
- ○ 30 a 49 pessoas
- ○ 50 a 99 pessoas
- ○ 100 a 249 pessoas
- ○ Acima de 249 pessoas
- ○ Não se aplica

* Quantas pessoas atuam diretamente como **cientista de dados** ou cargos análogos em seu local de trabalho?
Se trabalha de forma autônoma ou é o único responsável por projetos de dados, por favor, insira o número 1. Caso atualmente não realize projetos profissionais na área ou esta pergunta não se aplica à sua situação atual, por favor, preencha com 0.

Identificação

Informe seu e-mail caso queira receber o **relatório executivo gratuito** com as principais conclusões da pesquisa. Neste relatório, você poderá também avaliar suas competências em relação aos demais respondentes.

Assim, embora seja opcional, o fornecimento de seu e-mail é muito importante para a pesquisa.

E-mail

Por favor, verifique o formato de sua resposta

Comentários

Caso queira fazer algum **comentário adicional** sobre a pesquisa ou sobre a área de Ciência de Dados, fique à vontade:

Anterior

Enviar

Muito obrigado
por completar o questionário!

Sua participação será fundamental para o sucesso da pesquisa.

Se puder compartilhar a pesquisa com outros profissionais de dados, agradecemos ainda mais!

APÊNDICE 3– CURSOS SUPERIORES ANALISADOS

Instituição (IES)	Sigla	Categoria Administrativa	Nome do Curso	Grau	Modalidade	URL
UNIVERSIDADE DO VALE DO ITAJAÍ	UNIVALI	Privada sem fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	A Distância	https://ead.univali.br/cursos-graduacao/ciencia-de-dados-ead
UNIVERSIDADE ESTÁCIO DE SÁ	UNESA	Privada com fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	A Distância	https://estacio.br/cursos/graduacao/ciencia-de-dados
UNIVERSIDADE CRUZEIRO DO SUL	UNICSUL	Privada com fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	A Distância	https://www.cruzeirodosulvirtual.com.br/graduacao/ciencia-de-dados/
CENTRO UNIVERSITRIO ANHANGUERA PITÁGORAS AMPLI	AMPLI	Privada com fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	A Distância	https://www.ampli.com.br/graduacao/ciencia-de-dados
UNIVERSIDADE NOVE DE JULHO	UNINOVE	Privada sem fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	A Distância	https://www.uninove.br/cursos/graduacao/ead/tecnologia-em-ciencia-de-dados
CENTRO UNIVERSITRIO DE JOÃO PESSOA	UNIPÊ	Privada com fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	Presencial	https://www.unipe.edu.br/graduacao/ciencia-de-dados/
CENTRO UNIVERSITÁRIO RITTER DOS REIS	UNIRITTER	Privada com fins lucrativos	CIÊNCIA DE DADOS	Bacharelado	A Distância	https://www.uniritter.edu.br/graduacao/ciencia-de-dados
UNIVERSIDADE ANHEMBI MORUMBI	UAM	Privada com fins lucrativos	CIÊNCIA DE DADOS	Bacharelado	Presencial	https://portal.anhembi.br/graduacao/ciencia-de-dados/
Centro Universitário Anhanguera Pitágoras Unopar de Niterói	UNIAN-RJ	Privada com fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	A Distância	https://www.anhanguera.com/curso/ciencia-de-dados-tecnologo/
UNIVERSIDADE FEDERAL DO CEARÁ	UFC	Pública Federal	CIÊNCIA DE DADOS	Tecnológico	Presencial	https://itapaje.ufc.br/pt/ciencia-de-dados/
UNIVERSIDADE VILA VELHA	UVV	Privada com fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	A Distância	https://uvv.br/ead/graduacao/ciencia-de-dados/
Centro Universitário Anhanguera Pitágoras	UNOPAR	Privada com fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	A Distância	https://www.unopar.com.br/curso/ciencia-de-dados-tecnologo/

Instituição (IES)	Sigla	Categoria Administrativa	Nome do Curso	Grau	Modalidade	URL
Unopar de Campo Grande						
UNIVERSIDADE POSITIVO	UP	Privada com fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	A Distância	https://www.up.edu.br/graduacao/ciencia-dados-inteligencia-artificial/
Centro Universitário das Américas	CAM	Privada com fins lucrativos	CIÊNCIA DE DADOS	Bacharelado	Presencial	https://vemprafam.com.br/cursos/ciencia-de-dados-data-science/
CENTRO UNIVERSITÁRIO INTERNACIONAL	UNINTER	Privada com fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	A Distância	https://www.uninter.com/graduacao-ead/ciencia-de-dados-2/
Centro Universitário Joaquim Nabuco de Recife	UNINABUCO	Privada com fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	A Distância	https://graduacao.uninabuco.digital/nossos-cursos/ciencia-de-dados/618/5
FACULDADE CAPITAL FEDERAL	FECAF	Privada com fins lucrativos	CIÊNCIA DE DADOS	Tecnológico	Presencial	https://www.fecaf.com.br/cursos/ciencia-de-dados
FACULDADE SENAC PORTO ALEGRE - FSPOA	SENACRS	Privada sem fins lucrativos	CIÊNCIA DE DADOS E INTELIGÊNCIA ANALÍTICA	Tecnológico	Presencial	https://senacrs.com.br/cursos/curso-superior-de-tecnologia-em-ciencia-de-dados-e-inteligencia-analitica_WylxNTg2lixudWxsLG51bGwsbnVsbF0
PONTIFÍCIA UNIVERSIDADE CATÓLICA DE CAMPINAS	PUC-CAMPINAS	Privada sem fins lucrativos	CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL	Bacharelado	Presencial	https://www.puc-campinas.edu.br/graduacao/ciencia-de-dados-e-inteligencia-artificial/
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL	PUCRS	Privada sem fins lucrativos	CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL	Bacharelado	Presencial	https://www.pucrs.br/politecnica/curso/ciencia-de-dados/
UNIVERSIDADE DE SOROCABA	UNISO	Privada sem fins lucrativos	CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL	Bacharelado	Presencial	https://uniso.br/graduacao/curso/ciencia-de-dados-e-inteligencia-artificial
PONTIFÍCIA UNIVERSIDADE	PUCSP	Privada sem fins lucrativos	CIÊNCIA DE DADOS E	Bacharelado	Presencial	https://www.pucsp.br/graduacao/ciencia-de-dados-e-inteligencia-artificial

Instituição (IES)	Sigla	Categoria Administrativa	Nome do Curso	Grau	Modalidade	URL
CATÓLICA DE SÃO PAULO			INTELIGÊNCIA ARTIFICIAL			
UNIVERSIDADE FEDERAL DA PARAÍBA	UFPB	Pública Federal	CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL	Bacharelado	Presencial	https://sigaa.ufpb.br/sigaa/public/curso/portal.jsf?id=19420831&lc=pt_BR ou https://www.ufpb.br/cdn
Centro Universitário IBMEC	IBMEC	Privada com fins lucrativos	CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL	Bacharelado	Presencial	https://matriculas.ibmec.br/curso/ciencia-de-dados-e-inteligencia-artificial-4972#2
CENTRO UNIVERSITÁRIO DO INSTITUTO DE EDUCAÇÃO SUPERIOR DE BRASÍLIA - IESB	IESB	Privada com fins lucrativos	CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL	Bacharelado	Presencial	https://www.iesb.br/cursos/ciencia-de-dados-e-inteligencia-artificial/
CENTRO UNIVERSITÁRIO UNIDOM - BOSCO	UNIDOM - BOSCO	Privada com fins lucrativos	CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL	Bacharelado	A Distância	https://www.unidombosco.edu.br/cursos/ciencia-de-dados-e-inteligencia-artificial-ead/
CENTRO UNIVERSITÁRIO FUNDAÇÃO SANTO ANDRÉ	CUFSA	Privada sem fins lucrativos	CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL	Bacharelado	Presencial	https://www.fsa.br/graduacao/ciencia-de-dados-inteligencia-artificial/
ESCOLA DE MATEMTICA APLICADA	EMAp-FGV	Privada sem fins lucrativos	CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL	Bacharelado	Presencial	https://emap.fgv.br/curso/ciencia-de-dados-e-inteligencia-artificial
CENTRO UNIVERSITÁRIO DE BRASÍLIA	UNICEUB	Privada sem fins lucrativos	CIÊNCIA DE DADOS E MACHINE LEARNING	Bacharelado	Presencial	https://www.uniceub.br/pdp/graduacao/ti/ciencia-de-dados-e-machine-learning-245

Instituição (IES)	Sigla	Categoria Administrativa	Nome do Curso	Grau	Modalidade	URL
UNIVERSIDADE DE SÃO PAULO	USP	Pública Estadual	ESTATÍSTICA E CIÊNCIA DE DADOS	Bacharelado	Presencial	https://icmc.usp.br/graduacao/estatistica-bacharelado
UNIVERSIDADE DA AMAZÔNIA	UNAMA	Privada com fins lucrativos	DATA SCIENCE	Tecnológico	A Distância	https://graduacao.unama.br/nossos-cursos/data-science/427/94
CENTRO UNIVERSITÁRIO DO NORTE	UNINORTE	Privada sem fins lucrativos	DATA SCIENCE	Tecnológico	A Distância	https://graduacao.uninorte.com.br/nossos-cursos/data-science/427/130
CENTRO UNIVERSITÁRIO MAURÍCIO DE NASSAU	UNINASSAU	Privada com fins lucrativos	DATA SCIENCE	Tecnológico	A Distância	https://graduacao.uninassau.digital/nossos-cursos/data-science/427/60
Centro Universitário das Américas	CAM	Privada com fins lucrativos	MARKETING DIGITAL E DATA SCIENCE	Tecnológico	A Distância	https://www.famonline.com.br/cursos/marketing-digital-data-science/