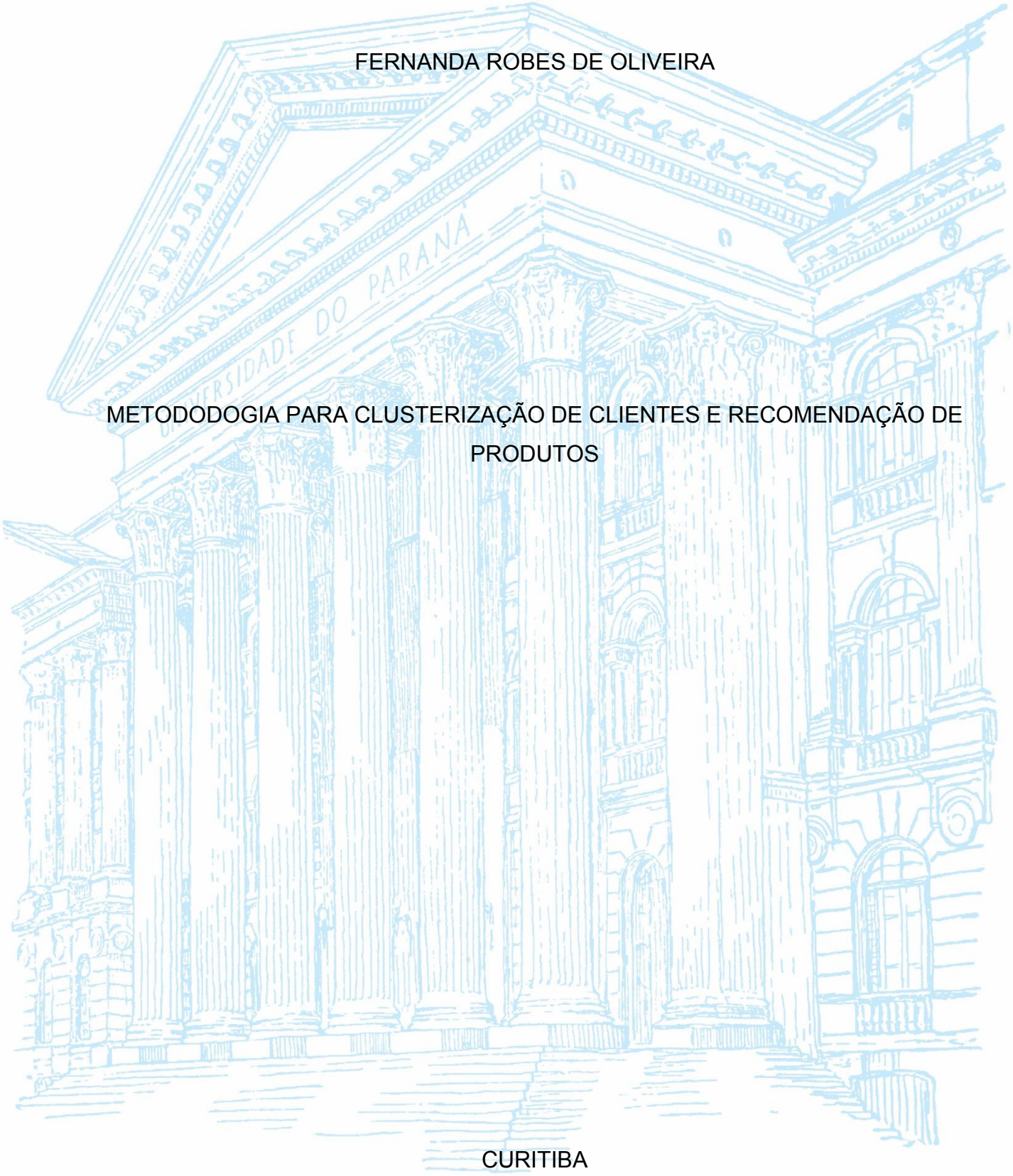
UNIVERSIDADE FEDERAL DO PARANÁ

FERNANDA ROBES DE OLIVEIRA



METODODOGIA PARA CLUSTERIZAÇÃO DE CLIENTES E RECOMENDAÇÃO DE PRODUTOS

CURITIBA

2021

FERNANDA ROBES DE OLIVEIRA

METODODOGIA PARA CLUSTERIZAÇÃO DE CLIENTES E RECOMENDAÇÃO DE PRODUTOS

Dissertação apresentada ao curso de Pós-Graduação em Engenharia de Produção, Setor de Tecnologia, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Engenharia de Produção.

Orientadora: Prof.ª Dra. Mariana Kleina.
Coorientador: Prof. Dr. Marcos Augusto Mendes Marques.

CURITIBA
2021

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

O48m Oliveira, Fernanda Robes de
Metododogia para clusterização de clientes e recomendação de produtos [recurso eletrônico] / Fernanda Robes de Oliveira – Curitiba, 2021.

Dissertação - Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-graduação em Engenharia de Produção .
Orientadora: Prof.ª Dra. Mariana Kleina
Coorientador: Prof. Dr. Marcos Augusto Mendes Marques

1. Clusterização (Análise por conglomerados). 2. K-Medoid (Algorítmo) I. Universidade Federal do Paraná. II. Kleina, Mariana. III. Marques, Marcos Augusto Mendes. IV. Título.

CDD: 004.3

Bibliotecária: Roseny Rivelini Morciani CRB-9/1585

MINISTÉRIO DA EDUCAÇÃO
SETOR DE TECNOLOGIA
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO ENGENHARIA DE PRODUÇÃO - 40001016070P1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ENGENHARIA DE PRODUÇÃO da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **FERNANDA ROBES DE OLIVEIRA** intitulada: **METODOLOGIA PARA CLUSTERIZAÇÃO DE CLIENTES E RECOMENDAÇÃO DE PRODUTOS**, sob orientação da Profa. Dra. MARIANA KLEINA, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre 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, 25 de Maio de 2021.

Assinatura Eletrônica
25/05/2021 22:13:32.0
MARIANA KLEINA
Presidente da Banca Examinadora

Assinatura Eletrônica
26/05/2021 12:26:09.0
WAGNER HUGO BONAT
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
26/05/2021 08:16:49.0
AGNELO DENIS VIEIRA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
31/05/2021 14:41:06.0
WALMES MARQUES ZEVIANI
Avaliador Externo (DEPARTAMENTO DE ESTATÍSTICA DA UNIVERSIDADE FEDERAL DO PARANá)

Documento assinado eletronicamente de acordo com o disposto na legislação federal Decreto 8539 de 08 de outubro de 2015.
Gerado e autenticado pelo SIGA-UFPR, com a seguinte identificação única: 93629
Para autenticar este documento/assinatura, acesse https://www.prppg.ufpr.br/siga/visitante/autenticacaoassinaturas.jsp e insira o codigo 93629

Dedicado à minha mãe que sempre me incentivou a estudar,
e ao meu marido que teve muita paciência comigo durante este trabalho.

RESUMO

O agrupamento de clientes auxilia o *marketing* estratégico, permitindo traçar estratégias diferenciadas a grupos específicos de clientes, com objetivo de criar relacionamento. A identificação do perfil de cada grupo associada a algoritmos de recomendação de produtos, auxilia os clientes a encontrarem os itens mais indicados às necessidades específicas. Esta facilidade pode auxiliar muito empresas que possuem uma extensa gama de produtos, pois, a tarefa exaustiva da busca de produtos por parte dos clientes pode ocasionar que ele compre da concorrência que pode conseguir fazer uma recomendação assertiva. Este trabalho se baseou nesta necessidade para demonstrar a aplicação de um método que combinou o algoritmo de clusterização e o de regras de associação, mostrando cada etapa da aplicação em uma base mista, que possuem tanto variáveis quantitativas quanto qualitativa. Os resultados mostraram que a medida de distância de Gower, utilizada para verificar a semelhança entre os clientes, gerou clusters com estrutura mais forte, de acordo com Coeficiente de Silhueta e o Índice de Davies Bouldin, se comparada a Jaccard. Para possibilitar o agrupamento empregou-se o K-Medoid, por ser mais flexível a utilização de diferentes medidas, o que propiciou a comparação e gerou onze clusters com perfis diferentes de clientes em um estudo de caso no setor de serviços. Para a recomendação de produtos foi avaliado o desempenho dos algoritmos Apriori, Filtragem Colaborativa Baseada em Clientes e Filtragem Colaborativa Baseada no Item, este último apresentou êxito nos dez primeiros clusters, analisando-se as Taxas de Recall e Precisão, e Curva ROC. Porém no cluster onze o Apriori apresentou melhores resultados. Após a identificação dos algoritmos de recomendação, visando otimizar as métricas de eficiência, foi ajustado o número de vizinhos mais próximos do algoritmo de Filtragem colaborativa e os parâmetros de suporte e confiança do Apriori, o que garantiu melhor desempenho de ambos.

Palavras-chave: Clusterização. Recomendação de Produtos. K-Medoid. Filtragem Colaborativa. Apriori.

ABSTRACT

The grouping of clients assists strategic marketing, allowing the design of differentiated strategies to specific groups of clients, with the objective of creating relationships. The identification of the profile of each group associated with product recommendation algorithms, helps customers to find the most suitable items for their specific needs. This facility can help a lot of companies that have an extensive range of products, because the exhaustive task of searching for products on the part of customers can cause them to buy from the competition that may be able to make an assertive recommendation. This work was based on this need to demonstrate the application of a method that combined the clustering algorithm and that of association rules, showing each step of the application on a mixed basis, which have both quantitative and qualitative variables. The results showed that the Gower distance measure, used to verify the similarity between the clients, generated clusters with a stronger structure, according to the Silhouette Coefficient and the Davies Bouldin Index, when compared to Jaccard. To make the grouping possible, K-Medoid was used, as it is more flexible to use different measures, which enabled the comparison and generated eleven clusters with different customer profiles in a case study in the service sector. For the recommendation of products, the performance of the Apriori algorithms, Collaborative Client-Based Filtering and Item-Based Collaborative Filtering was evaluated, the latter was successful in the first ten clusters, analyzing the Recall and Precision Rates, and ROC Curve. However, in cluster eleven, Apriori presented better results. After identifying the recommendation algorithms, in order to optimize the efficiency metrics, the number of neighbors closest to the collaborative filtering algorithm and the support and trust parameters of Apriori were adjusted, which guaranteed better performances by both.

Keywords: Clustering. Product Recommendation. K-Medoid. Collaborative Filtering. Apriori.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em um mercado cada vez mais competitivo, entender o comportamento de compras dos clientes é um fator relevante para que uma empresa se mantenha no mercado, bem como para melhorar os seus indicadores financeiros. A nova economia está amplamente focada em uma melhor prestação de serviços, sendo que a era atual é chamada de economia orientada para o cliente (ZOERAM; MAZIDI, 2018).

Assim, a segmentação pode melhorar o relacionamento entre a empresa e o cliente (BAFGHI, 2017), pois com o agrupamento dos consumidores, por meio das necessidades e comportamentos semelhantes, torna-se mais fácil a criação e a sugestão de um mix de produtos específicos, o que facilita o contato e personaliza o atendimento.

Há possibilidade, portanto, de estabelecer estratégias de *marketing* baseadas nas características particulares dos segmentos e/ou focar em um agrupamento específico, a fim de criar linhas de atendimento personalizadas, já que as diferentes características exigem tratamentos distintos (BEHESHTIAN-ARDAKANI; FATHIAN; GHOLAMIAN, 2018). Também é interessante destacar os ganhos em *marketing* de recomendação, tendo em vista a possibilidade de entregar produtos orientados pelo interesse dos clientes.

A partir da compreensão do cenário mencionado, a empresa pode analisar criteriosamente cada grupo, com o intuito de examinar o potencial de crescimento da sua marca, além de estabelecer uma abordagem diferenciada, promover os produtos específicos, projetar a demanda, verificar satisfação.

Desta forma pode se afastar de uma estratégia de *marketing* de massa em direção a um planejamento mais focado em grupos específicos de clientes (KARA; KAYNAK, 1997). Com maior foco, há a garantia de mais agilidade e rentabilidade para a empresa, bem como uma melhora na prestação dos serviços.

Para compreender e aplicar estes benefícios, se faz necessário a utilização de técnicas que envolvem a Mineração de Dados (RAJAGOPAL, 2011) e Inteligência Artificial, bem como métodos estatísticos multivariados, com a finalidade de processar o alto volume de dados e gerar informações confiáveis. Há muitas ferramentas que podem ser utilizadas para auxiliar neste processo de agrupar clientes tais como: Fuzzy, K-means, K-medoids, Árvore de decisão, bem como de associar a estes grupos

produtos conforme o seu perfil de consumo, com os algoritmos Apriori, Filtragem Colaborativa entre outros.

1.1 TEMA

Estudo e aplicação de técnicas de agrupamento de clientes e regras de associação para recomendação de produtos aplicadas a bases de dados de uma empresa do setor de serviços do Paraná.

1.2 PROBLEMA DE PESQUISA

Como técnicas analíticas de clusterização de clientes e associação de produtos podem melhorar estratégias de *marketing* de uma empresa?

1.3 OBJETIVOS

Nesta sessão serão apresentados o objetivo geral e os específicos deste trabalho.

1.3.1 Objetivo geral

Propor uma metodologia que combine técnicas de agrupamento de clientes em conjunto com regras de associação para recomendação de produtos, visando melhorar o desempenho do *marketing* de serviços.

1.3.2 Objetivos específicos

1. Identificar na literatura técnicas para a clusterização de clientes e técnicas de associação para a recomendação de produtos;
2. Analisar as técnicas existentes que aparecem com maior frequência na literatura para que possam ser estudadas e aplicadas na presente pesquisa
3. Propor uma metodologia para ser utilizada em uma base de dados de uma empresa de serviços;

4. Analisar os resultados obtidos por meio de análise descritiva dos clusters gerados e produtos recomendados;
5. Melhorar os parâmetros dos modelos de recomendação utilizados.

1.4 JUSTIFICATIVA

Nota-se que a competitividade das empresas está cada vez mais baseada no valor agregado que elas geram para o cliente. Desta maneira, é difícil uma empresa atender a todos os clientes com uma única estratégia de *marketing* (KARA; KAYNAK, 1997), assim uma empresa que é capaz de identificar os produtos certos e encaminhá-los aos clientes, conforme a sua necessidade, destaca-se da concorrência e, consequentemente, aumenta sua eficiência em vendas.

O trabalho traz, conceitos e técnicas utilizadas na clusterização de clientes, que podem ser utilizadas por profissionais para melhorar as vantagens competitivas de suas empresas. Destaca-se ainda que esta pesquisa é relevante para a academia, pois aborda conceitos e técnicas na sua fundamentação teórica, que é embasada em uma revisão sistemática sobre o tema e, assim, pode ser vista como uma importante ferramenta, a fim de aumentar a competitividade das empresas no mercado em que estão inseridas, além de auxiliar na elaboração de estratégias de *marketing*, no gerenciamento da base de dados dos clientes e na criação de mix de produtos.

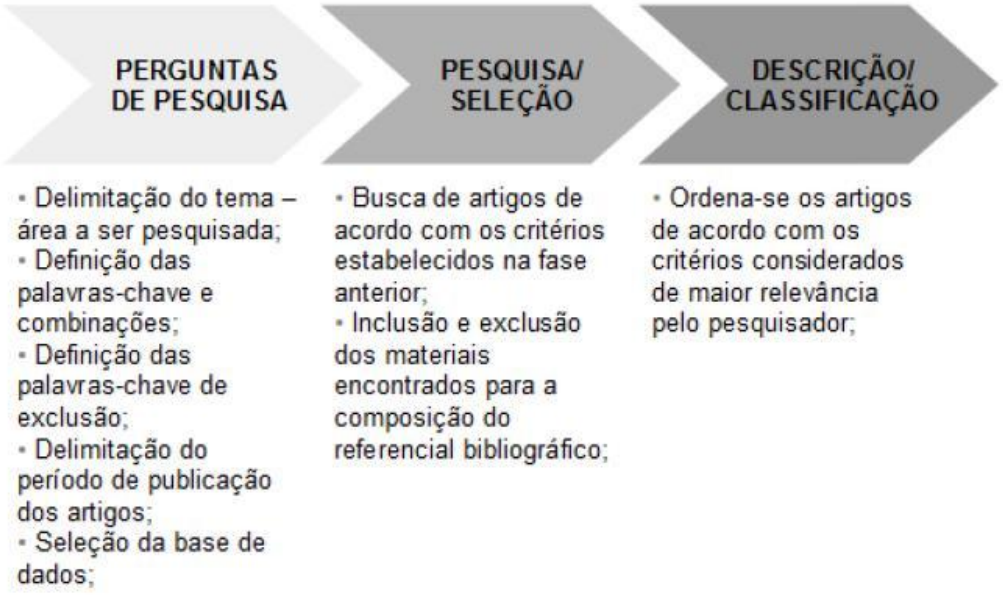
1.5 DELIMITAÇÕES DO TRABALHO

Esta pesquisa está focada em ferramentas de clusterização de clientes e regras de associação para a recomendação de produtos, sendo que o recorte das ferramentas estudadas nos artigos foi elencado na revisão sistemática utilizada para dar o suporte metodológico ao estudo. As técnicas serão aplicadas em uma base de dados de atendimentos realizados no Paraná por uma empresa do setor de serviços.

2 REVISÃO DE LITERATURA E REFERENCIAL TEÓRICO

Antes de definir uma metodologia para o tema ora descrito, foi proposta uma revisão sistemática com o objetivo de mapear e analisar estudos relevantes para a pesquisa, bem como identificar as possíveis lacunas (DRESCH; LACERDA, 2015) que podem ser preenchidas pela pesquisa em questão. Com a finalidade de obter respostas satisfatória e garantir reprodutibilidade do método, bem como evitar o viés do pesquisador, utilizou-se a metodologia proposta por Gohr et al. (2013) , que divide a revisão sistemática em três etapas, conforme demonstrado na FIGURA 1.

FIGURA 1 - ETAPAS DA REVISÃO SISTEMÁTICA



FONTE: Gohr et al. (2013)

2.1 PERGUNTAS DE PESQUISA

A primeira etapa do método, conforme observado na FIGURA 1, é composta pela delimitação do tema a ser pesquisado, escolha das palavras-chave e inclusão e exclusão, período de análise e escolha das bases de dados.

Desta maneira, o tema que esta revisão sistemática buscou elucidar foi “Quais as técnicas que podem ser utilizadas para clusterizar clientes e recomendar produtos?". Assim para responder o tema apresentado, estratificou-se o tema em duas perguntas de pesquisa, uma relacionada a clusterização, “Quais técnicas podem ser

utilizadas para clusterizar clientes?" e outra referente a recomendação, "Quais técnicas podem ser utilizadas para recomendar produtos?", como pode ser observado no QUADRO 1.

Para garantir que as buscas nas bases de dados fossem mais objetivas, foram utilizados caracteres curingas, tais como o asterisco (*) e o cifrão ($). O asterisco (*) foi utilizado com a intenção de obter nas bases de dados palavras que comecem com o termo buscado, porém possam ter diferentes terminações, por exemplo, com o termo cluster* é possível obter na busca artigos que contenham as palavras cluster, clustering, clustered. Já o cifrão ($) foi utilizado para que os termos de busca possam variar para o plural, assim "Association Rule$" considera "Association Rules" e "Association Rule".

Com relação aos operadores booleanos, foi utilizado o OR, que retorna documentos que tenham um dos dois termos e AND utilizado para retornar documentos que contenham ambos termos pesquisados.

Além dos aspectos citados para garantir uma busca de artigos que remetesse melhor ao objetivo, foi considerado na estratégia palavras contidas no título e em tópicos, em que tópicos abrange palavras que apareçam no título, resumo e palavras-chave.

Isto posto, para primeira pergunta de pesquisa "Quais técnicas podem ser utilizadas para clusterizar clientes?", foram utilizadas as seguintes palavras de busca: cluster*, customer* e *marketing*. O termo *marketing* aparece nas estratégias de busca, pois está intrínseco ao objetivo geral do trabalho.

Já para a segunda pergunta de pesquisa "Quais técnicas podem ser utilizadas para recomendar produtos?" foram utilizadas como palavras de busca: *Association Rule*, *recommendation*, *Market Basket*, *products* e *marketing*. Estes termos foram segmentados em duas estratégias de busca diferentes, Busca 1 e Busca 2, conforme QUADRO 1, de forma a coletar um maior número de documentos. Destaca-se que a segunda busca foi realizada pois a primeira retornou poucos artigos que contribuíram com o propósito da pesquisa, desta forma com a leitura dos artigos determinou-se uma segunda estratégia de busca visando complementar e abranger melhor o tema.

Também compuseram o protocolo delimitações do tipo de documento, idioma e tempo, em que foram considerados somente: artigo, inglês, de 2005 a 2020 respectivamente.

Scopus e Web of Science foram selecionadas como as bases utilizadas como fonte de dados para esta revisão. Observa-se no QUADRO 1, que devido as diferenças de plataformas houve variações nas buscas. Também é importante salientar que a utilização das bases foi condicionada ao acesso pela plataforma CAPES por meio do sistema CAFE (Comunidade Acadêmica Federada), o que garante o acesso remoto ao conteúdo assinado do Portal de Periódicos disponível a UFPR.

QUADRO 1 - ESTRATÉGIA DE BUSCA

Estratégia de busca	Termos Utilizados		
Segmentação do tema	Clusterização	Recomendação	
Questão de Pesquisa	Quais técnicas podem ser aplicadas para clusterizar clientes?	Quais técnicas para recomendar produtos aos clientes?	
Bases	Web of Science e Scopus	Web of Science e Scopus	
Strings de busca	Web of science: TI= (cluster* AND customer$) AND TS=(*marketing*) Scopus: TITLE (cluster* AND customer$) AND TITLE-ABS-KEY (*marketing*)	BUSCA 1: Web of science: TI= (((“Association Rule$”) OR recommendation) AND product$) AND TS= (*marketing*) Scopus: TITLE (((“Association Rule$”) OR recommendation) AND product$) AND TITLE-ABS-KEY (*marketing*)	BUSCA 2: Web of science: TI= (((“Association Rule$”) OR "Market Basket")) AND TS= (*marketing* AND product$)) Scopus: TITLE (((“Association Rule$”) OR "Market Basket") AND TITLE-ABS-KEY (*marketing* AND product$))
Tipo de documento	Artigos		
Idioma	Inglês		
Tempo	2005-2020		

FONTE: A autora (2021).

LEGENDA: TI – Pesquisa o campo Título; TS – Pesquisa termos nos campos Título, Resumo e Palavras-chave.

2.2 SELEÇÃO DE ARTIGOS

Nesta fase da metodologia ocorrem as buscas nas bases conforme critérios determinados na fase anterior, bem como são elencados os critérios de exclusão dos artigos de forma a compor o referencial bibliográfico.

Após a seleção dos artigos é importante considerar apenas os que possuem conteúdo vinculado ao objetivo da pesquisa, para isto foram estabelecidos alguns critérios de exclusão conforme detalhado a seguir:

1. Exclusão por duplicidade: Após a seleção dos artigos nas bases definidas, os documentos são reunidos, excluindo as duplicidades;
2. Exclusão por título: Efetuada a leitura dos títulos dos artigos, de forma que os artigos em que os títulos que não estejam condizentes aos objetivos da pesquisa são excluídos;
3. Exclusão por resumo: Executada a leitura dos resumos dos documentos restantes, sendo excluídos os artigos cujo resumo não esteja relacionado aos objetivos da pesquisa;

Após a realização das exclusões é realizada a leitura completa dos artigos, registrando-se resumidamente as suas contribuições.

2.3 APLICAÇÃO DAS ESTRATÉGIAS DE BUSCA E SELEÇÃO DE ARTIGOS

A aplicação dos critérios de busca foi segmentada em duas partes, na primeira buscou-se artigos referentes a clusterizaçã*o*, e na segunda parte artigos referentes a recomendação.

O emprego dos critérios relativos a clusterização resultou em 38 artigos da base Web of Science e 44 da Scopus. Aplicando-se a primeira regra de exclusão, ou seja, excluindo os artigos em duplicidade restaram 64 artigos conforme demonstrado na TABELA 1, em que é possível notar como os critérios, tipo de documento, idioma e tempo, reduziram o quantitativo inicial de documentos.

TABELA 1 - QUANTIDADE DE ARTIGOS DE CLUSTERIZAÇÃO

Busca	Web Of Science	Scopus
Strings de busca	106	106
Strings de busca + Tipo de documento	41	48
Strings de busca + Tipo de documento + Idioma	41	44
Strings de busca + Tipo de documento + Idioma + Tempo	38	44
Total após retirar duplicidade	64	

FONTE: A autora (2021).

No tocante a segunda parte, ou seja, o emprego das estratégias de busca pertencentes ao tema recomendação, que responde a segunda pergunta de pesquisa, decorreu em 47 artigos referentes a estratégia de busca 1 e 59 da busca 2, removendo-se os artigos duplicados findou-se em 97 documentos, conforme evidenciado na TABELA 2, que tiveram conteúdo analisado a fim de responder à questão de pesquisa ora levantada.

TABELA 2 - QUANTIDADE DE ARTIGOS DE RECOMENDAÇÃO

	Busca 1		Busca 2	
Bases	Web Of Science	Scopus	Web Of Science	Scopus
Strings de busca	85	73	101	55
Strings de busca + Tipo de documento	45	36	60	27
Strings de busca + Tipo de documento + Idioma	44	36	60	27
Strings de busca + Tipo de documento + Idioma + Tempo	38	30	52	22
Total	47			59
Total após retirar duplicidade	97			

FONTE: A autora (2021).

Nos quantitativos de artigos encontrados foram aplicados os critérios de exclusão, removendo os artigos que possuíam título ou resumo que não eram aderentes ao objeto de pesquisa.

A TABELA 3 apresenta a aplicação dos critérios de exclusão em ambas as temáticas de pesquisa. De maneira que é possível verificar que dos 64 artigos resultantes da estratégia de busca sobre clusterização, após a leitura dos títulos restaram 51 artigos, os quais tiveram os resumos lidos, destes apenas 26 mostraram-se relevantes a pesquisa.

Já na temática de recomendação, dos 97 artigos, 53 tinham títulos favoráveis a pesquisa e tiveram os resumos lidos, restando em 16 artigos que demonstram mais aplicáveis ao objeto da pesquisa.

TABELA 3 - RESULTADOS DOS CRITÉRIOS DE EXCLUSÃO

	Clusterização		Recomendação	
Critério	Título	Resumo	Título	Resumo
Total	64	51	97	53
Excluído	13	25	44	37
Restantes	51	**26**	53	**16**

FONTE: A autora (2021).

Os artigos que foram classificados como relevantes foram lidos na íntegra, e tiveram registrados as suas contribuições para o objeto do estudo.

2.4 CLUSTERIZAÇÃO DE CLIENTES

Como as necessidades dos clientes tornam-se cada vez mais complexas e dinâmicas em virtude de novos produtos e serviços que são lançados todos os dias no mercado (WANG; GAO, 2019), a diversificação dos produtos focada na segmentação baseada nas necessidades dos clientes é uma ferramenta essencial para as empresas se manterem em um ambiente de negócios competitivo (LIN et al., 2019), pois contribui para manter e atrair clientes (KEVREKIDIS et al., 2018).

A clusterização de clientes é um processo que pode ser definido como a ação de separar os clientes com base nas características em grupos distintos e significativos (DESGRAUPES, 2013; TSIPTSIS E CHORIANOPOULOS, 2010; NGAI et al., 2009) de modo que os clientes do mesmo agrupamento tenham necessidades e preferências semelhantes.

Isto é fundamental quando a empresa está buscando entender o mercado, pois, por meio da sua implementação há possibilidade de estruturar uma estratégia de *marketing* mais eficiente para cada segmento (BARMAN; CHOWDHURY, 2019), ação que auxilia as empresas a criar relacionamento com o público-alvo (LIN et al., 2019), oferecendo serviços personalizados, bem como ofertas, promoções e recomendações às preferências de cada segmento (JAGABATHULA; SUBRAMANIAN; VENKATARAMAN, 2018) o que faz os clientes responderem as estratégias de forma mais eficiente (WANG; CHIN, 2017).

Da mesma maneira, identificar os clientes mais valiosos auxilia na alocação de recursos, bem como tem um papel relevante nas decisões de gestão (KHALILI-DAMGHANI; ABDI; ABOLMAKAREM, 2018).

O agrupamento de clientes é um tópico importante para o CRM (*Customer Relationship Management*) (CHEN, 2012), que se concentra em melhorar os processos de negócios associados ao gerenciamento de relacionamento com os clientes, nas áreas de vendas e atendimento (RENUKA DEVI; BHARATHI; PRASAD, 2019). Também incluem adquirir novos clientes e aumentar a lucratividade dos clientes existentes, mantendo na base clientes rentáveis (TSAI; HU; LU, 2015).

Porém conhecer o que os clientes desejam e assim segmentá-los pode não ser tão fácil em empresas que tem milhares de clientes, desta forma é fundamental utilizar técnicas que auxiliem neste contexto (CAMERO et al., 2018). De sorte ferramentas e técnicas de análise de negócios e a tomada de decisões orientada a dados estão cada vez mais presentes para suprir as necessidades empresariais (GRIVA et al., 2018).

Uma técnica de mineração de dados utilizada para este fim é o agrupamento ou clusterização, que é um método não supervisionado para agrupar os dados em grupos semelhantes (BARMAN; CHOWDHURY, 2019), com o objetivo de alocar as observações em agrupamentos homogêneos internamente e heterogêneos entre si com a função de representar o comportamento das variáveis.

2.4.1 Descrição de artigos de clusterização selecionados

A clusterização de clientes tem grande importância para a manutenção das empresas, de tal modo que existem muitas técnicas utilizadas com esta proposta. Na revisão sistemática utilizada foram observados vários métodos propostos para diferentes tipos de empresas e dados. A seguir segue o registro das técnicas utilizadas em cada um dos artigos selecionados:

Zoeram e Mazidi (2018) utilizaram-se do Fuzzy K-means aplicado a variáveis de recência, frequência e valores monetários, para clientes de uma empresa atacadista de pratos. Para avaliar a significância dos grupos obtidos utilizou-se a ANOVA.

Chang e Ho (2017) apresentaram um método de duas camadas, para segmentar clientes de uma empresa de telefonia. Primeiramente os clientes foram segmentados pelo percentual de receita e posteriormente foi utilizado o algoritmo K-means, utilizando o índice de Silhueta para avaliar a quantidade de clusters e ANOVA (Análise de Variância) para significância dos grupos.

Bafghi (2017) propõe algoritmos para maximizar a similaridade em cada cluster e para analisar a qualidade do cluster. Estes algoritmos são utilizados após um algoritmo genético clusterizar os clientes, e são utilizados para mudar os clientes de clusters, considerando as similaridades existentes.

Abbasimehr e Shabani (2019) utilizaram o algoritmo de Ward alternando com diferentes medidas de similaridade, como: *Correlation-based mensure* (COR), *Temporal correlation coefficient* (CORT), *Dinamic time warping* (DTW) e Distância Euclidiana, e analisando a quantidade de grupos pelo índice de Silhueta.

Sheikh, Ghanbarpour e Gholamiangonabadi (2019) aplicaram o algoritmo K-means e o índice Davies-Bouldin em dados de clientes de uma *fintech*.

Yanik e Elmorsy (2019) para gerar grupos de clientes utilizaram primeiramente a análise de componentes principais, considerando variáveis demográficas, transação de cartão de crédito e as categorias de produtos consumidos, com os resultados foram empregados uma rede neural não supervisionada denominada *Self Organizing Map* (SOM) e o K-means. O número de cluster foi verificado por meio do índice de Davies-Bouldin e a significância utilizando-se a ANOVA. A pesquisa demostrou que utilizando somente os dados demográficos e o algoritmo SOM, conseguiram-se melhores resultados.

Khalili-Damghani, Abdi e Abolmakarem (2018) primeiramente tratam os dados utilizando o método *Local outlier factor* (LOF) para remover os dados discrepantes. Os clusters são gerados pelo algoritmo K-means em que é analisada quantidade ideal de clusters por Davies-Bouldin. Após é utilizado uma árvore de decisão.

Munusamy e Murugesan (2020) propõem um algoritmo de Segmentação Dinâmica de Clientes que atualiza os segmentos de clientes com novas informações de forma dinâmica. Assim à medida que novas informações entrem no conjunto de dados, as segmentações de clientes são alteradas, podendo-se criar cluster e eliminar outros. Os autores propõem o algoritmo *Modified dynamic fuzzy c-means* (MDFCM) e comparam com o desempenho do *Crespo's dynamic fuzzy c-means* (CDFCM), demonstrando que o algoritmo proposto tem melhor desempenho, além de gerar clusters significativos.

Renuka Devi, Bharathi e Prasad (2019) aplicam o algoritmo K-means em um conjunto de dados de telecomunicações demonstrando vários índices para melhor analisar os dados e escolher um número de clusters, tais como os de validação interna (Dunn, Silhueta) e validação externa (Rand e Jaccard). Mostram que a padronização

dos dados permite a geração de clusters mais bem separados e densos, bem como a seleção de atributos utilizando os métodos de Filter e Wrapper, reduz o tempo de processamento, bem como remove atributos que podem ser irrelevantes para construir um grupo.

Zheng (2013) demonstra a aplicação do algoritmo C-means em dados de uma empresa de valores imobiliários. No processo para determinação do melhor cluster, realiza a padronização dos dados e utiliza uma função de validade para determinar o melhor agrupamento em uma série de 2 a 11 grupos.

Luo et al. (2013) aplicam K-means em dados de clientes de uma empresa de telecomunicações e evidenciam como o algoritmo pode trazer resultados práticos e uma segmentação eficaz.

Chen (2012) demonstra a aplicação do algoritmo C-means em dados normalizados de recência, frequência e valor monetário, o que evidencia que um cliente pode ser simultaneamente classificado em vários grupos em diferentes graus. Utiliza o teste S (Distância separada) para verificar o melhor número de clusters.

LI, Dai e Tseng (2011) aplicam o algoritmo Ward e nos dados gerados é aplicado o K-means, os clusters são analisados utilizando a ANOVA e o teste Scheffé.

Hosseini, Maleki e Gholamian (2010) aplicam K-means em dados de uma empresa de acessórios de automóveis e utilizam Davies-Bouldin para verificar a quantidade de clusters.

Bose e Chen (2010) utilizam dados de uma empresa de telecomunicações para realizar o agrupamento de clientes. Para isto, normalizam os atributos, pois os dados tinham um alto desvio, utilizando a transformação log, e converte os dados para o intervalo 0 a 1. Devido a uma grande quantidade de atributos, foi realizada uma análise de correlação, sendo que os atributos com correlações superiores a 0,7 foram removidos. Para determinar as quantidades de clusters a serem analisadas, o algoritmo *Self Organizing Map* (SOM) foi utilizado devido a capacidade da técnica de representar visualmente dados quantitativos multidimensionais. Após a preparação dos dados foram aplicados os algoritmos K-means e *Kohonen vector quantization* (KVQ). Foi utilizado Dunn para verificar o número ideal de clusters. Conclui-se que o KVQ é tendencioso para criação de clusters com menor dispersão média intra-cluster, enquanto o K-means tende a criar clusters com uma variação menor.

Sohn e Kim (2008) para descobrir padrões na utilização de serviços baseados nas características dos clientes, analisam 115 variáveis. Primeiro, devido a quantidade

de variáveis utilizam a técnica de Análise de Componentes Principais, em seguida segmentam os clientes com base nos fatores derivados da primeira análise. A clusterização dos clientes foi determinada pelo algoritmo K-means.

Lingras et al. (2005) com o objetivo de encontrar as migrações temporais dentre clusters, utilizam o algoritmo Kohonen SOM modificado. Cada período analisado durou quatro semanas e incluíram os mesmos clientes. As alterações no cluster ao longo do tempo sugerem modificações no comportamento de compra dos clientes.

Bi et al. (2016) propuseram o algoritmo *Semantic Driven Subtractive Clustering Method* (SDSCM), este possui uma velocidade de execução rápida quando comparado com outros métodos.

Sajjadi et al. (2015) utilizam outras variáveis associadas ao método RFM (Recência, Frequência e Valor Monetário) e segmentam os clientes de agências bancárias utilizando o algoritmo K-means.

Seret et al. (2014) utilizam metodologia de mapas auto organizáveis utilizando o algoritmo P-SOM que é uma modificação do tradicional SOM, pois fornece um mecanismo que permite a priorização de variáveis. Na saída do algoritmo é utilizado o K-means para clusterizar os dados e o índice Davies Bouldin para verificar o número ideal de clusters.

Kuo et al. (2016) primeiramente aplicam a técnica de Análise de Componentes Principais para reduzir as dimensões e o algoritmo K-means em dados obtidos em um aplicativo de controle de peso. Para determinar o número de grupos utilizam a Soma do quadrado do erro. Após este passo, algoritmos genéticos são utilizados visando aprimorar os resultados obtidos, estes são *GA-based K-means clustering ensembles* (GKCE), *PSO-based K-means clustering ensembles* (PSOKCE), *ABC-based K-means clustering ensembles* (ABCKCE), eles são comparados utilizando a Média do quadrado do erro. O resultado revela que o algoritmo ABCKE tem a melhor solução.

Hiziroglu e Senbas (2016) comparam os algoritmos K-means e C-means, demonstrando que o segundo produz clusters mais homogêneos quanto a quantidade de clientes.

Brentari, Dancelli e Manisera (2016) aplicam um algoritmo hierárquico utilizando uma matriz de dissimilaridade baseada na medida *Weighted rank correlation* (WRC), utilizada por ser mais adequada a dados categóricos. Além disso, a Média Bipolar (BM) foi utilizada para analisar a medida de variabilidade das variáveis.

Tsai, Hu e Lu (2015) comparam os grupos gerados pelos algoritmos K-means e Expectativa maximização (EM).

Gucdemir e Selim (2015) comparam métodos de agrupamentos Hierárquicos com o K-means e concluem que este produz menores Soma dos Quadrados no Cluster. A Soma de Quadrados foi utilizada por ser mais simples para avaliar a validade dos resultados de agrupamentos e números de clusters.

Gupta, Aggarwal e Rani (2016) agruparam clientes de um shopping, utilizando o algoritmo Two Step e o índice de Silhueta para verificar a quantidade de clusters.

Zheng (2013) utiliza o algoritmo C-means para clusterizar os clientes de uma empresa de valores imobiliários.

O QUADRO 2, resume as técnicas e variáveis utilizadas para a clusterização dos clientes nos artigos selecionados. Nota-se que as técnicas comportam vários tipos de dados e abrangem empresas de vários segmentos, demonstrando que a preocupação na segmentação dos clientes não recai em um único tipo de empresa.

QUADRO 2 - RESUMO DE ARTIGOS DE CLUSTERIZAÇÃO

(Continua)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
A new methodology for customer behavior analysis using time series clustering: A case study on a bank's customers.	Abbasimehr e Shabani (2019).	Clientes que usam dispositivos de banco.	Medidas de silhueta; Medidas de similaridade: COR; CORT, DTW e Distância Euclidiana; Algoritmo de agrupamento de Ward.	Recência (R): tempo de dias desde a última compra; Frequência (F): número de compras em um determinado período; Monetário (M): Quantidade total de dinheiro gasto durante um período específico.
Clustering of Customers Based on Shopping Behavior and Employing Genetic Algorithms.	Bafghi (2017).	Dados de compras de consumidores.	Algoritmos desenvolvidos no artigo; Algoritmo genético; ANOVA.	Transações: Quantidade de vezes que o cliente realiza uma compra; Quantidade de itens comprados; Data e hora da transação; Código do cliente; Valor monetário.
A Two-Layer Clustering Model for Mobile Customer Analysis.	Chang e Ho (2017).	Empresa de telefonia.	K-means; Silhueta; ANOVA.	Receita; Produtos; Consumo.
Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries.	Khalili-Damghani, Abdi e Abolmakarem (2018).	Clientes de uma empresa de seguros e de telefonia.	Método LOF; K-means; Árvore de decisão.	Variáveis de pesquisa, dados quantitativos e qualitativos.

(Continuação)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
A Preliminary Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting.	Sheikh, Ghanbarpour e Gholamiangonabadi (2019).	Clientes de uma *fintech.*	K-means; Davies-Bouldin.	Intervalo entre a primeira e a última transação; Intervalo entre a última transação e no final o período; Número de transações que ocorreu no período; O agregado de valor das transações durante o período; O intervalo médio entre dias que transação.
SOM approach for clustering customers using credit card transactions.	Yanik e Elmorsy (2019).	Clientes de cartão de crédito.	SOM; Análise de Componentes Principais; Davies Boudin; ANOVA.	Demográficas (idade, sexo, estado civil, escolaridade e renda mensal); Dados de transação do cartão de crédito (soma das transações, número de transações, média de transações); Categorias de consumo (Consumo mensal de educação, entretenimento, eletrônicos, vestuário, álcool, artigos de papelaria, acomodações, joalheria, mercado, automóveis e combustíveis, saúde, viagem e transporte, seguro, decoração, comida e outros).

(Continuação)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
A New Approach for Customer Clustering by Integrating the LRFM Model and Fuzzy Inference System.	Zoeram e Mazidi (2018).	Empresa atacadista de pratos.	Fuzzy K-means; ANOVA.	Recência (R): tempo de dias desde a última compra; Frequência (F): número de compras em um determinado período; Monetário (M): Quantidade total de dinheiro gasto durante um período específico. Comprimento (L): Duração do relacionamento com o cliente, número de dias entre a primeira e última visita;
Modified dynamic Fuzzy c-means clustering algorithm – Application in dynamic customer segmentation.	Munusamy e Murugesan (2020).	Dados transacionais de um supermercado.	CDFCM; MDFCM.	Recência (R): tempo de dias desde a última compra; Frequência (F):número de compras em um determinado período; Monetário (M): Quantidade total de dinheiro gasto durante um período específico.
Prediction of customer churn in telecom sector using clustering technique.	Renuka Devi, Bharathi e Prasad (2019).	Dados do setor de telecomunicações.	K-means; Validação Interna: Conectividade; Índice Dunn; Silhueta; Validação externa: Índice Rand; Índice Jaccard; Seleção de atributos: Filter; Wrapper.	Valor da conta; Total de chamadas; Custo do dia; Plano de correio de voz; Minutos internacionais; Rotatividade.

(Continuação)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
Customer segmentation for telecom with the K-means clustering method.	Luo et al. (2013).	Empresa de telecomunicações.	K-means.	Tempo de acesso à rede; Abertura de serviços; Informações sobre pacotes de benefícios; Informações sobre serviços; Taxas mensais comerciais; Taxas de usuários; Taxas de concessão; Serviços de valor agregado; Informações de pagamentos em atraso; Duração da chamada; Horário das chamadas; Duração das chamadas.
The RFM-FCM approach for customer clustering.	Chen (2012).	Dados fictícios.	Fuzzy C-means; Distância separada (teste S).	Recência (R): tempo de dias desde a última compra; Frequência (F): número de compras em um determinado período; Monetário (M): Quantidade total de dinheiro gasto durante um período específico.
A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business.	Li, Dai e Tseng (2011).	Fábrica de tecidos.	K-means; ANOVA; Scheffé.	Comprimento (L): Duração do relacionamento com o cliente, número de dias entre a primeira e última visita; Recência (R): tempo de dias desde a última compra; Frequência (F):número de compras em um determinado período; Monetário (M): Quantidade total de dinheiro gasto durante um período específico.

(Continuação)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty.	Hosseini, Maleki, e Gholamian (2010).	Fábrica de acessórios para automóveis.	K-means; Davies Bouldin.	Recência (R): tempo de dias desde a última compra; Frequência (F): número de compras em um determinado período; Monetário (M): Quantidade total de dinheiro gasto durante um período específico.
Exploring business opportunities from mobile services data of customers: An inter-cluster analysis approach.	Bose e Chen (2010).	Dados de serviços moveis de uma operadora de telecomunicações móveis.	KVQ; K-means; Davies Bouldin.	Qtd. Discagem internacional; Quantidade de chamadas efetuadas; Tempo total das chamadas; Total de serviço de mensagens; Receita total obtida de serviços; Gênero; Idade; Tipo de cliente; Serviços assinados.
Searching customer patterns of mobile service using clustering and quantitative association rule.	Sohn e Kim (2008).	Dados de serviço de telecomunicações móveis.	Análise de Componentes Principais; K-means.	115 variáveis que contêm informações sobre os clientes e a utilização dos serviços.

(Continuação)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
Temporal analysis of clusters of supermarket customers: conventional versus interval set approach.	Lingras et al. (2005).	Dados de um supermercado.	Kohonen SOM modificado.	Inclui informações do cliente sobre gastos; visitas; categorias de produtos; lojas compradas e outros dados transacionais.
A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn.	Bi et al. (2016).	Dados de uma empresa de telecom.	SDSCM.	Registros de chamadas; Registros de cobranças; Informações demográficas; Pacote de serviços; Histórico de compras;
A developing model for clustering and ranking bank customers.	Sajjadi et al. (2015).	Dados de agências bancárias.	K-means.	Recência (R): é a distância do tempo da última transação do cliente; Frequência (F): é o número de interações financeira do cliente com o banco durante um período; Monetário (M): O total de depósitos dos clientes; Total de instalações recebidas (L): total de compromissos que o cliente recebeu; Diferido total (D): significa o total dívidas incobráveis.

(Continuação)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
A dynamic understanding of customer behavior processes based on clustering and sequence mining.	Seret et al. (2014).	Dados de venda de ingressos de uma empresa de eventos.	P-SOM; K-means; Davies Bouldin.	Dias entre a compra do ingresso e o evento; Tempo de relacionamento com a empresa; Quantidade de compras on-line; Número médio de ingressos compradas para cada evento; Gênero do cliente; Valor do cliente; Distância entre o cliente e o local do evento;
An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation.	Kuo et al.(2016).	Dados de usuários de um aplicativo móvel de controle de peso.	SSE; K-means; Análise de Componentes Principais; GKCE; PSOKCE; ABCKCE; MSE.	Dados de pesquisa de restaurantes; Informações de exercícios; Registro de peso.
An application of fuzzy clustering to customer portfolio analysis in automotive industry.	Hiziroglu e Senbas (2016).	Dados de clientes de uma empresa de fabricação de acessórios automotivos.	K-means; Fuzzy C-means; ANOVA.	Rotatividade; Frequência de pedidos; Custo do relacionamento com o cliente.

(Continuação)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
Clustering ranking data in market segmentation: a case study on the Italian McDonald's customers' preferences.	Brentari, Dancelli e Manisera (2016).	Dados de uma pesquisa realizada com clientes de um restaurante.	WRC; Média Bipolar.	Preço; Serviço (tempo de espera, serviço de mesa, cortesia da equipe); Sabor e variedade de alimentos; Qualidade do ingrediente; Qualidade nutricional; Atmosfera (decoração, conforto dos assentos, limpeza); Conveniência (estacionamento, Proximidade a pontos de interesse, Horário de funcionamento).
Customer segmentation issues and strategies for an automobile dealership with two clustering techniques.	Tsai, Hu e Lu (2015).	Dados de uma concessionária de automóveis.	Análise de Componentes Principais; K-means.	60 variáveis em seis categorias: Informações demográficas; Informações sobre transações; Pesquisa de veículo; Pesquisa de serviço de manutenção; Pesquisa de terceirização e manutenção.
Integrating multi-criteria decision making and clustering for business customer segmentation.	Gucdemir e Selim (2015).	Fabricante de televisores.	Ward; K-means; Soma dos Quadrados no Cluster.	Lealdade; Demanda média anual; Potencial de relacionamento de longo prazo; Recência; Frequência; Valo monetário.

(Conclusão)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
Segmentation of retail customers based on cluster analysis in building successful CRM.	Gupta, Aggarwal e Rani (2016).	Dados de clientes de um shopping center.	Two Step; Silhueta.	Sexo; Idade; Estado civil; Educação; Profissão; Renda; Localidade.
Application of silence customer segmentation in securities industry based on fuzzy cluster algorithm.	Zheng (2013).	Empresa de valores imobiliários.	Fuzzy C-means; Função de validade.	Ativos do cliente; Grau de contribuição do cliente; Retorno de ativos do cliente; Preferência por um estoque; Grua de ações bloqueadas; Grau de risco do cliente; Capacidade de resposta do cliente.

FONTE: A autora (2021).

2.4.1.1 Frequência observada de técnicas referentes a clusterização nos artigos selecionados

Com relação a frequências das técnicas, nota-se que dos artigos analisados o algoritmo K-means apareceu em dezesseis (61%), embora haja críticas sobre este método, dado que há necessidade de fornecer a quantidade de clusters (MONALISA; KURNIA, 2019). O algoritmo Fuzzy C-means apareceu em quatro (15%), o SOM e o Ward foram utilizados em três (11%) dos artigos analisados. Houve outras técnicas utilizadas, porém não se repetiram nos artigos.

Outra contribuição da revisão foi mostrar a utilização de técnicas que apoiam a clusterização, dentre elas, a Análise de Componentes Principais ocorreu em quatro artigos (15%), esta técnica foi principalmente citada para reduzir a quantidade de variáveis utilizadas.

Evidencia-se também que técnicas de avaliações da quantidade e qualidade dos clusters também foram empregadas, sendo Davies-Bouldin (19%) e o Coeficiente de Silhueta (15%) os que apresentaram maior frequência. Já para análise de qualidade, a ANOVA foi empregada em seis (23%) artigos.

Em relação as variáveis utilizadas para clusteriza*r* clientes, nove artigos (35%) utilizaram, somente ou incluindo com outras, as variáveis Recência, Frequência e Valor monetário.

É importante destacar que com a execução do projeto de pesquisa, outras técnicas que não apareceram na revisão da literatura foram consideradas importantes. Pois, com o processamento da base de dados observou-se variáveis qualitativas e quantitativas (base mista), o que impediu a distância euclidiana ser considerada, dado ao fato de poder ser utilizada somente em variáveis quantitativas.

Assim o algoritmo K-medoids e distâncias compatíveis com bases mista (Gower e Jaccard), foram utilizadas para suprir os impasses encontrados.

2.4.2 Fundamentos teóricos relacionados a clusterização

2.4.2.1 Análise de Agrupamentos

São análises exploratórias utilizadas para verificar e agrupar comportamentos semelhantes entre observações. É um componente essencial da mineração de dados e fundamental para reconhecimento de padrões. É utilizado para agrupar as

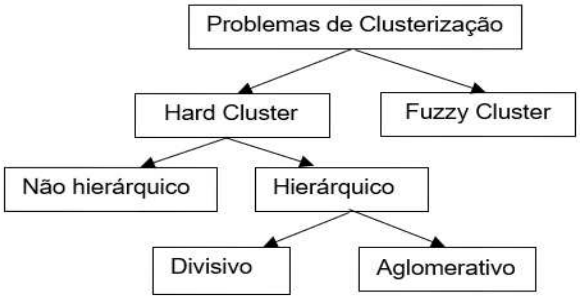
observações de maneira que os grupos sejam homogêneos internamente e heterogêneos entre si, ou seja, as observações das variáveis de um grupo devem ser semelhantes, mas diferentes de um outro grupo (FÁVERO; BELFIORE, 2017). É um processamento de dados não supervisionado, assim não há rótulos, desta forma o algoritmo verifica as relações entre os dados e classifica as observações em alguns grupos, com base na similaridade (TSAI; HU; LU, 2015).

Com relação a classificação, conforme pode-se observar na FIGURA 2 os algoritmos de clusterização podem ser segmentados em Hard Cluster e Fuzzy Cluster. O primeiro permite que uma observação pertença a apenas um cluster , já no segundo pode pertencer a dois ou mais grupos com probabilidades diferentes (GANMAWU; WELLS, 2007).

A tipologia Hard Cluster, pode também ser dividida em dois tipos de algoritmos: Hierárquico e Não hierárquico (KUO et al., 2016), enquanto o primeiro privilegia uma estrutura hierárquica, ou seja, uma vez que uma divisão ou fusão de uma observação é feita, é irrevogável. Ao passo que no processo Não Hierárquico, maximiza-se a homogeneidade dentro dos agrupamentos.

Destaca-se que a tipologia Hierárquica se segmenta em outros dois: Divisivo, que separa as observações sucessivamente em grupos mais refinados e o Aglomerativo que faz sucessivas fusões das observações em grupos.

FIGURA 2 - DIAGRAMA DE TIPOS DE ALGORITOMOS DE CLUSTER



FONTE: Adaptado de Ganmawu e Wells (2007).

Há uma série de técnicas para cada uma das classificações vistas, porém segundo Hennig (2015) não existe a melhor técnica de clusterização, mas elas devem ser selecionadas no contexto dos dados, de forma a escolher a técnica mais apropriada.

Também é importante destacar que um dos principais requisitos dos algoritmos de agrupamento são as medidas de distância (dissimilaridade) para variáveis numéricas ou semelhança (similaridade) para variáveis binárias (FÁVERO; BELFIORE, 2017). Medidas de distância são fundamentais para solucionar vários problema e reconhecimento de padrões (CHA, 2008). Desta forma, o primeiro passo no processo de clusterização é verificar as variáveis a serem utilizadas, pois a escolha da medida de distância dependerá disto.

Segundo Hunt e Jorgensen (2011), caso os dados sejam mistos, ou seja, possuam variáveis qualitativas e quantitativas, poderá ser utilizada uma medida de distância híbrida ou utilizar técnicas para dicotomizar as variáveis e deste modo dispor de uma medida de semelhança para dados binários gerados. Porém, Harikumar e Surya (2015) apontam que utilizar a técnica de dicotomização para transformar um conjunto de dados mistos em homogêneos, pode levar a perda de informação.

Existe na literatura uma série de medidas para calcular a distância entre as observações, a mais popular quando as variáveis são quantitativas é a euclidiana. Porém para variáveis categóricas, em que os atributos assumem valores nominais, é impossível calcular a diferença entre duas observações, logo é comumente utilizada a análise de proximidade entre duas observações, que é calculada pela proporcionalidade de dados equivalentes (AHMAD; DEY, 2007a).

2.4.2.2 Algoritmos de clusterização

2.4.2.2.1 K-means

É um dos algoritmos de agrupamento mais utilizados e foi descrito pela primeira vez por Macqueen (1967). É classificado como não hierárquico e empregado em dados numéricos (BUDIAJI; LEISCH, 2019), pois a distância euclidiana é falha em capturar a similaridade de atributos categóricos (AHMAD; DEY, 2007b). Além disso, depende da definição do número de clusters (k) pelo analista.

O algoritmo baseia-se na atribuição de cada dado ao centroide mais próximo. Gucdemir e Selim (2015) definem as etapas como:

1. Determine o número de clusters (k);
2. Atribua aleatoriamente k pontos para serem os centroides iniciais dos clusters;

3. Atribua o ponto ao cluster cujo centroide estiver mais próximo;
4. Recalcule os centroides dos clusters;
5. Repita as etapas 3 a 5 até a finalização.

Detalhando as etapas anteriores, tem-se que na primeira etapa o analista deve definir o número de clusters, é importante destacar que há diferentes técnicas para auxiliar a escolha desta quantidade.

Na segunda são atribuídos aleatoriamente k pontos centroides, para inicializar o algoritmo.

Na terceira etapa é calculada a distância de cada ponto a cada centroide, geralmente utilizando a distância euclidiana, e então o ponto é associado ao cluster cujo centroide estiver mais próximo.

Na quarta etapa, para cada novo ponto associado, os centroides devem ser atualizados, ou seja, deve ter a respectiva média calculada novamente.

Assim as etapas de três a cinco devem ser repetidas até que não haja mais alterações entre os clusters ou por um número predefinido de iterações (SERET et al., 2014).

2.4.2.2.2 K-medoid

O algoritmo é baseado no objeto medoides, ou seja, em um ponto observado, que é localizado mais centralmente em um cluster. É menos sensível a *outliers*, se comparado ao K-means, além de ser mais flexível, permitindo diferentes tipos de medidas de distância (DE ASSIS; DE SOUZA, 2011).

Usa pontos de referência ao invés do valor médio dos elementos em cada cluster, e precede da entrada do número de clusters (UMAMAHESWARI; DEVI, 2018).

Há vários algoritmos para agrupamento de K-medoids, porém segundo Park e Jun (2009), o Particionamento em torno de medoides (PAM), proposto por Kaufman e Rousseeuw (1990), é conhecido por ser mais poderoso, desta forma foi utilizado neste trabalho.

Reynolds, Richards e Rayward-smith (2004) resumiram o algoritmo da seguinte forma:

1. Selecione k objetos aleatoriamente para se tornarem medoides dos clusters iniciais;
2. Calcule a matriz de dissimilaridade se ela não foi fornecida;

3. Atribua cada objeto ao seu medoide mais próximo;
4. Para cada cluster, pesquise o objeto do cluster que diminui o cociente de dissimilaridade média, selecione o objeto que mais diminui este quociente como o medoide para este cluster;
5. Repita as etapas 3 e 4 até que os medoides se tornem fixos.

2.4.2.3 Análise do número de clusters

A escolha do número de grupos é de importância central, mas ainda é um dos problemas mais difíceis na análise de cluster, porque geralmente é questionável e não há uma única solução existente (BRENTARI; DANCELLI; MANISERA, 2016). Assim, serão detalhados os índices Davies-Boudin e Silhueta, por terem apresentado maior frequência na revisão sistemática da literatura.

2.4.2.3.1 Davies-Bouldin

É um índice que avalia a validade do cluster (SHEIKH; GHANBARPOUR; GHOLAMIANGONABADI, 2019), em função da proporção da dispersão dentro do cluster, para a separação entre clusters. Desta maneira, um valor mais baixo significa que a clusterização é melhor.

Segundo Ganmawu e Wells (2007), para definir o índice é necessário primeiro definir a medida de dispersão e similaridade do cluster.

A medida de dispersão do cluster é apresentada na equação (1).

$$S_i = \left(\frac{1}{|C_i|} \sum_{x \in C_i} d^p(x, c_i) \right)^{\frac{1}{p}}, \qquad p > 0 \tag{1}$$

Onde:

S_i : é a medida de dispersão do cluster C_i;

$|C_i|$: é o número de pontos no cluster C_i;

c_i : é o centro do cluster C_i;

d : é a distância entre x e c_i.

Normalmente o valor de *p* é 2, o que torna esta uma distância euclidiana;

Já medida de similaridade entre os clusters pode ser definida como R_{ij} , que mensura a similaridade entre os clusters C_i e C_j .

Assim R_{ij} é demonstrada na equação (2).

$$R_{ij} = \frac{S_i + S_j}{D_{ij}} \quad (2)$$

Onde:

D_{ij}: é a distância entre os centroides dos *clusters* i e j.

Como S_i e S_j são as distâncias intra-cluster dos clusters i e j, ou seja, a distância média de cada instância ao centroide do cluster e D_{ij} é a distância entre os centroides dos clusters i e j (inter-cluster), assim os clusters i e j devem ser diferentes, caso contrário a distância entre os clusters será zero. Percebe-se que quanto menor R_{ij} , mais distantes e menor a dispersão entre os clusters.

Desta maneira o índice Davies-Bouldin (V_{DB}) é representado na equação (3).

$$V_{DB} = \frac{1}{k} \sum_{i=1}^{k} R_i \quad (3)$$

Onde:

k: Número de clusters;

R_i: É o valor máximo R_{ij}.

A equação (3) demonstra que para cada cluster i será selecionado o cluster j menos semelhante, e este valor será dividido pela quantidade de clusters. Desta forma quanto melhor o agrupamento, mais próximo a zero será o índice.

2.4.2.3.2 Coeficiente de Silhueta

A Silhueta é uma medida da distância entre os grupos utilizada para determinar a qualidade dos clusters. Para cada observação é definido um índice $\epsilon[-1,1]$, que compara a distância da observação ao cluster, com a heterogeneidade do cluster. A análise se dá pela média do índice no cluster, assim quanto mais perto

de 1, melhor a qualidade do cluster, ou seja, sua heterogeneidade é menor. Segundo Chang e Ho (2017), se o coeficiente for maior que 0,5, o cluster é mais eficaz em distinguir entre heterogêneo ou homogêneo. No QUADRO 3, Kaufman e Rousseeuw (1990) demonstram como pode ser interpretado o coeficiente.

QUADRO 3 - INTERPRETAÇÃO DO COEFICIENTE DE SILHUETA

Coeficiente de Silhueta	**Interpretação**
0,71 a 1,00	Estrutura forte
0,51 a 0,70	Estrutura razoável
0,26 a 0,50	Estrutura fraca
Menor que 0,25	Nenhuma estrutura

FONTE: Kaufman e Rousseeuw (1990).

O coeficiente de silhueta é calculado usando a distância intra-cluster média e a distância média do cluster mais próximo.

A equação (4) define uma silhueta para a observação m.

$$s_m = \frac{b_m - a_m}{max\{a_m, b_m\}} \tag{4}$$

Onde:

a_m: é a distância média entre a observação m as demais observações do cluster;

b_m: é a distância média entre a observação m e todas as observações do cluster mais próximo.

Sendo o Coeficiente de Silhueta definido na equação (5).

$$SC = \max_m \overline{s}\,(k) \tag{5}$$

Onde:

$\overline{s}\,(k)$: representa a média de s_m sobre todos os dados para um número específico de clusters k.

2.4.2.4 Medida de distância ou semelhança

Para que um algoritmo de clusterização aloque cada observação em um grupo é necessário que conheça o quanto estas observações são parecidas ou não. Existem várias medidas de distância, e devem ser escolhidas de acordo com o tipo de dados. A seguir será vista a Distância de Gower e Jaccard.

2.4.2.4.1 Distância de Gower

Proposta por Gower em 1971, a distância é capaz de lidar com conjunto de dados mistos e segundo Bektas e Schumann (2019), faz isto de forma eficaz. A distância é calculada pela equação (6) .

$$S_{mn} = \frac{\sum_{v=1}^{p} W_v \, S_v}{\sum_{v=1}^{p} W_v} \tag{6}$$

Em que S_{mn} é a distância entre as observações x_m e x_n , com $m \neq n$. Se a k-ésima variável é qualitativa tem-se S_v pela equação (7).

$$S_v = \begin{cases} 0 \, , se \; x_{vm} = x_{vn} \\ 1 \, , se \; x_{vm} \neq x_{vn} \end{cases} \tag{7}$$

Caso a k-ésima variável for quantitativa, S_v é dado pela equação (8).

$$S_v = \frac{|\, x_{vm} - x_{vn}|}{max(x_v) - min\,(x_v)} \tag{8}$$

Onde:

v : variável elegida;

p : número total de variáveis;

x_{vm} : é o valor da v-ésima variável para a observação m;

W_v : é igual a 1 (um) quando se tem os valores da v-ésima variável para ambos os elementos e 0 (zero), quando não se tem os valores da v-ésima variável para quaisquer dos dois elementos.

2.4.2.4.2 Distância de Jaccard

Medida utilizada para variáveis binárias, leva em conta a presença do atributo nas observações, excluindo o número de ausências conjuntas (HENNIG, 2015). A distância de Jaccard (dJ) é apresentada na equação (9).

$$dJ(M,N) = \frac{C_{01} + C_{10}}{C_{01} + C_{10} + C_{11}} \quad (9)$$

Onde:

M e N: São os conjuntos de observações com n atributos binários.

C_{11}: Representa o número de atributos em que M e N têm valor 1;

C_{01}: Representa o número de atributos em que M tem valor 0 e N é igual a 1;

C_{10}: Representa o número de atributos em que M é igual a 1 e N é 0.

2.5 RECOMENDAÇÃO DE PRODUTOS

A mudança das necessidades dos clientes juntamente com um rápido avanço da tecnologia aumentou a variedade das ofertas de produtos (WANG; CHIN, 2017), assim frequentemente os clientes precisam realizar uma extensa pesquisa de alternativas disponíveis para fazer uma compra (DELLAERT; HÄUBL, 2012).

A concorrência força as empresa a desenvolver atividades de *marketing* inovadoras para capturar as necessidade do cliente, melhorar a satisfação e retenção (LIU; SHIH, 2005b), ou seja, orientar-se para o cliente é uma prioridade no mercado (VIDELA-CAVIERES; RIOS, 2014).

Deste modo para criar melhores estratégias de *marketing*, as empresas devem entender as associações de vendas entre os produtos e serviços (WENG, 2016), pois assim podem fornecer recomendações mais atraentes (DELLAERT; HÄUBL, 2012). De modo a manter-se competitivas, várias empresas costumam utilizar alguns sistemas de recomendação para indicar produtos relevantes, que podem interessar aos clientes, ou auxiliá-los a encontrar produtos mais adequados. (LIN; GOH; HENG, 2017).

É importante entender que os sistemas de recomendação, além de auxiliar os clientes a encontrar produtos, tem inúmeros outros benefícios, como:

a) Auxiliar a gerenciar o sortimento de produtos e assim impulsionar a demanda (LIN; GOH; HENG, 2017). Isto ocorre devido ao fato que há uma maior exposição dos produtos recomendados para clientes com o perfil de compra selecionado (LIN; GOH; HENG, 2017);
b) Auxilia a implantação de um *marketing* de estratégias (MOSTAFA, 2015), pois ao detectar produtos comprados juntos, a empresa inicia uma ação de *marketing* fornecendo descontos para aumentar as vendas (MITRA et al., 2016);
c) Táticas podem ser utilizadas para projetar a estratégia de preços em toda a gama de produtos associados (MOSTAFA, 2015);
d) Permite que as empresas desenvolvam *marketing* individual para cada cliente (LIU; SHIH, 2005b);
e) Os sistemas de recomendação aumentam a probabilidade de venda cruzada, lealdade do cliente e descobrem produtos que os clientes podem estar interessados (LIU; SHIH, 2005b).

O crescimento do volume de dados e os avanços tecnológicos beneficiam as empresas, que podem utilizar a mineração de dados para analisar os perfis de clientes e assim melhorar a decisão de *marketing* (LIU; SHIH, 2005b).

2.5.1 Descrição de artigos de recomendação selecionados

Como já exposto, a recomendação de produtos tem muitas vantagens para as empresas em geral, ainda mais em um ambiente competitivo.

Desta forma a revisão sistemática visou aprofundar o conhecimento das técnicas utilizadas para recomendação. A seguir, estão resumidos os artigos selecionados.

Hung (2005) demonstra um algoritmo que pondera a probabilidade de compra de um determinado produto pela taxonomia do produto, frequência e valor gasto. O valor é analisado por classificação do produto, tipo e marca, para então recomendar os três produtos com maiores chances. O autor detalha que novos produtos podem ser indicados pela marca mais propensa de compra de um determinado cliente.

Liu e Shih (2005a) utilizam como primeiro passo da recomendação de produtos a clusterização de clientes, para isto utilizam o algoritmo K-means nas

variáveis frequência, recência e valor monetário, para só então utilizar o algoritmo de Filtragem Colaborativa.

Liu e Shih (2005b) utilizam Filtragem Colaborativa para recomendar produtos, tendo como medida de dissimilaridade a correlação de um índice formado pela ponderação os valores das variáveis de frequência, recência e valor monetário. Segundo os autores, a técnica auxilia na qualidade das recomendações, porém a metodologia é mais eficaz para clientes fiéis.

Shih e Liu (2008) demonstram que o método híbrido que combina métodos baseados em ponderação das variáveis recência, frequência e valor monetário e dados de demandas e compras anteriores dos clientes, é mais eficaz que outros métodos. Utilizam o agrupamento dos clientes pelo K-means e para recomendações o algoritmo de Filtragem Colaborativa. Além disso, comparam a qualidade das recomendações utilizando as métricas Recall, Precisão e F1.

Weng (2016) propõe um algoritmo, para descobrir as informações específicas de regras de associação para itens comercializados em momentos diferentes. Compara o algoritmo proposto com o Apriori e demonstra que para conjunto de produtos específicos os valores de suporte e confiança apresentados pelo algoritmo proposto são superiores se comparados com o Apriori.

Huang et al. (2008), ao contrário de modelos convencionais que consideram apenas os produtos que aparecem com frequência, propõem um modelo que considera as regras de associação pelo preço e categoria de produtos, o que permite uma saída mais personalizada, já que há clientes bem diferentes, alguns que procuram preços mais baixos e outros itens exclusivos. Para atingir o objetivo são utilizados os seguintes algoritmos: Apriori TiD para extrair as regras de associação, ART2 (teoria da ressonância adaptativa) para agrupar os preços de venda, Fuzzy para alterar os agrupamentos de preços para graus de associação, Backpropagation (BP) que é um dos métodos de aprendizagem mais utilizados de rede neural artificial.

Chiang et al. (2011) discorrem que em geral os produtos com mais tempo de mercado tendem a ter um suporte mais alto, assim produtos novos tendem a ter um suporte, em geral, inferior ao mínimo e dificilmente serão recomendados. Assim propõem um algoritmo que verifica a associação entre produtos tais como A-> BvC, para conseguir aumentar a métrica de confiança e recomendar ambos os produtos novos.

Babu e Bhuvaneswari (2014) sugerem uma modificação no algoritmo Apriori, de maneira que o algoritmo ganhe mais velocidade, uma vez que precisa varrer o banco uma única vez. Compara o tempo de execução do algoritmo proposto com os algoritmos FP-Growth, Apriori e DynFP, demonstrando que o algoritmo proposto é mais veloz.

Videla-Cavieres e Rios (2014) apresentam o grafo como uma maneira de especificar o relacionamento entre conjunto de itens. Desta forma geram uma rede de produtos, de forma que cada produto está vinculado a outro por uma ponderação de transações no tempo.

Zekic-Susac e Has (2015) demonstram um método de recomendação de segundo nível, em que o algoritmo utiliza não o produto, mas o segmento dele, por exemplo, produtos lácteos ao invés de leite. Assim nota-se que no nível de produtos, devido a grande quantidade de itens e quantidade de regras gerada pode ser grande, podendo haver prejuízo na qualidade destas regras. O agrupamento de segundo nível permite extrair regras mais claras e precisas para o uso dos tomadores de decisão.

Turčínek e Turčínkova (2015) aplicaram o algoritmo FP-Growth e Apriori em uma base de dados referente a pesquisa sobre hábitos de compra de produtos à base de carne, com o objetivo de verificar a associação dos atributos avaliados de escolha da loja aos grupos de clientes, assim os algoritmos foram utilizados na exploração do comportamento do consumidor. O Algoritmo Priori forneceu dados mais refinados se comparado com o outro algoritmo, acredita-se que isto foi decorrente a alteração de escala para binária exigida pelo FP-Growth.

Mostafa (2015) utiliza *Multidimensional Scaling* (MDS), que é uma técnica de pesquisa de *marketing* usada para detectar visualmente padrões complexos em conjuntos de dados de alta dimensão, para representar as relações entre produtos comprados em uma cesta. As relações analisadas são geradas por meio do algoritmo Apriori.

Valle, Ruz e Morras (2018) apresentam uma análise de rede baseada em *Mini-Spanning Tree* (MST) como uma abordagem MBA (*Marketing Basket Analysis* ou em português análise da cesta de compras). Em que com o algoritmo gera interpretações fáceis e possibilita reconhecer inter-relações entre um conjunto de produtos. As vantagens do MST destacadas são: capacidade de superar dados em excesso, maior controle sobre relações espúrias e ruídos.

Unvan (2020) compara os algoritmos Apriori e FP-Growth, em uma aplicação de dados de compras em um supermercado. Demonstra que FP-Growth aplicado em dados categóricos tem uma maior capacidade de processamento e ocupa menos memória.

Dos dezesseis artigos selecionados, dois ao invés de demonstrar técnicas aplicadas, traziam conceitos gerais sobre as técnicas o que permitiu ampliar a visão sobre os termos utilizados.

O QUADRO 4 resume as técnicas e variáveis utilizadas para a recomendação de produtos que foram vistas nos artigos selecionados. Nota-se que as técnicas comportam vários tipos de dados e abrangem principalmente empresas do segmento de varejo.

QUADRO 4 - RESUMO DOS ARTIGOS SOBRE RECOMENDAÇÃO

(Continua)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
A personalized recommendation system based on product taxonomy for one-to-one marketing on-line.	Hung (2005).	Empresa de varejo on-line.	Utiliza uma estratégia segmentação dos produtos por tipo e marca, em que estabelece um algoritmo denominado RRT (Recompra recente do cliente) utilizado para fornecer estimativa de recompra do produto ou produtos similares.	Valor total das compras em um determinado período; Valor do produto; Frequência de compras.
Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences.	Liu e Shih (2005a).	Empresa de varejo de equipamento como rodas, rodízios, plataformas e carrinhos de mão.	Segmenta os clientes em grupos utilizando K-means e aplica a Filtragem colaborativa para recomendação de produtos em cada grupo.	Frequência; Recência; Valor monetário;
Integrating AHP and data mining for product recommendation based on customer lifetime value.	(LIU; SHIH, 2005b).	Empresa de varejo de equipamento como rodas, rodízios, plataformas e carrinhos de mão.	*Clusteriza* os clientes utilizando o K-means com a metodologia RFM e utiliza a mineração de regras de associação em cada grupo por meio da Filtragem Colaborativa.	Frequência; Recência; Valor monetário;
Product recommendation approaches: Collaborative filtering via customer lifetime value and customer demands.	Shih e Liu (2008).	Empresa de varejo de equipamento como rodas, rodízios, plataformas e carrinhos de mão.	Filtragem Colaborativa.	Quantidade comprada por produto; Frequência; Recência; Valor monetário.

(Continuação)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
Identifying association rules of specific later-marketed products.	Weng (2016).	Aplicado em três conjuntos de dados: supermercado, loja de móveis e de uma mercearia.	Proposto um algoritmo que utiliza a variável tempo para calcular o suporte e confiança.	Data de compra do produto; Dados de transação do produto;
Using backpropagation to learn association rules for service personalization.	Huang et al. (2008).	Banco de dados transacional.	Utiliza o Apriori TiD para criar as regras de associação, após os produtos são agrupados pela variável preço utilizando o algoritmo ART-2, desta forma somente são oferecidos aos clientes os produtos que estão no mesmo cluster.	Dados transacionais; Preço do produto; Categoria do produto;
Mining disjunctive consequent association rules.	Chiang et al.(2011).	Aplicado em uma base de empresa de seguro de automóveis.	Oferece uma nova proposta de regas de associação disjuntivas entre produtos, para aumentar a métrica de confiança.	Dados de vendas por produto; Quantidade total de produtos vendidos;

(Continuação)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
Association rule mining for identifying optimal customers using MAA algorithm.	Babu e Bhuvaneswari (2014).	Dados gerados.	Compara os algoritmos Apriori, MAA (*Algoritmo Apriori Modificado*), algoritmo proposto, FP-Growth and DynFP-growth, demonstrando que o tempo de execução do algoritmo proposto é menor.	Dados de vendas por produto; Quantidade total de produtos vendidos.
Extending Market Basket Analysis with graph mining techniques: A real case.	Videla-Cavieres e Rios (2014).	Cadeia de supermercados atacadista e uma cadeia de supermercados de varejo.	Desenvolve uma metodologia de rede, e demonstra que um grafo é uma maneira de evidenciar o relacionamento entre um conjunto de itens.	Dados de vendas por produto; Quantidade total de produtos vendidos; Dia da venda.
Discovering market basket patterns using hierarchical association rules.	Zekic-Susac e Has (2015).	Dados de uma cadeia de lojas de varejo.	Utiliza o algoritmo FP-tree (Árvore de padrão frequente) analisando as regras de associações geradas em uma estrutura hierárquica de produtos.	Dados de vendas por produto; Quantidade total de produtos vendidos; Segmento do produto.
Exploring consumer behavior: Use of association rules.	Turčínek e Turčínkova (2015).	Dados proveniente de questionário aplicado para obter percepções de clientes sobre forma de compra de produtos à base de carne.	Utiliza algoritmos de associação para analisar o comportamento dos clientes identificado por uma pesquisa quantitativa. Compara os algoritmos: Apriori e FP-Growth, sendo que os resultados do algoritmo Apriori foram melhores.	Questionário que elenca 20 atributos sobre forma de escolha de uma loja que vende produtos à base de carne. As respostas das perguntas eram configuradas em uma escala de 1 a 10.

(Conclusão)

Artigo	Autor(es) e ano	Aplicação	Técnicas Utilizadas (Algoritmos)	Variáveis Utilizadas
Knowledge discovery of hidden consumer purchase behaviour: A Market Basket Analysis.	Mostafa (2015).	Dados de compra em supermercado.	Extraiu-se regras de associação por meio do algoritmo Apriori e demonstra visualmente padrões complexos utilizando a metodologia MDS.	Dados de vendas por produto; Quantidade total de produtos vendidos.
Market basket analysis: Complementing association rules with minimum spanning trees.	Valle, Ruz e Morras (2018).	Dados de compras de supermercado.	Demonstra os algoritmos de Árvore de abrangência Mínima (*MST – minimum spanning tree*) como uma alternativa a Análise de Regra de Associação utilizada na cesta de mercado.	Data da transação; ID do cliente; Código interno do item; Quantidade comprada de cada item; Preço pago.
Market basket analysis with association rules.	Unvan (2020).	Dados de compras de supermercado.	Compara os algoritmos Apriori e FP-Growth e conclui que o último utiliza menos memória e tem um menor tempo de execução sendo melhor para trabalhar com dados binários.	Dados de vendas por produto; Quantidade total de produtos vendidos.

FONTE: A autora (2021).

2.5.1.1 Frequência observada de técnicas referentes a recomendação nos artigos selecionados

Dos artigos selecionados, observa-se que os algoritmos de associação que tiveram maior destaque foram o Apriori que apareceu em quatro (29%) dos artigos analisados, Filtragem Colaborativa (FC) e FP-Growth, ambos com três artigos (21%). Todos os artigos trouxeram índices que auxiliam a mensurar a qualidade da associação tais como Suporte e Confiança.

As técnicas de associação que apareceram com maior frequência na revisão da literatura, para melhor compreensão e análise, serão detalhadas adiante.

2.5.2 Fundamentos teóricos relacionados a recomendação

Antes de iniciar com os algoritmos de recomendação é fundamental conhecer alguns conceitos para entendimento mais amplo sobre a temática, como Mineração de Dados e Mineração de Regras de associação.

2.5.2.1 Mineração de dados

A mineração de dados pode ser definida como um processo para extrair informações importantes de dados existente para permitir uma melhor tomada de decisão em toda a organização. Ela tem se tornado cada vez mais importante devido a crescente disponibilidade de grande volume de dados (LIU; SHIH, 2005b).

Além disso, tem-se tornado amplamente aceita pelas empresas por aprimorar o desempenho organizacional e obter vantagem competitiva, uma vez que prevê tendências e comportamentos futuros. Uma das ferramentas mais amplamente utilizadas é a Análise de cesta de Mercado (MBA), também chamada de Mineração de Regras de Associação, que é um método muito utilizado pelo setor de varejo para promover produtos.

2.5.2.2 Mineração de Regras de Associação

A técnica originou-se no *marketing* e foi primeiramente utilizada para analisar quais itens em um supermercado são frequentemente comprados juntos (MUSALEM; ABURTO; BOSCH, 2018). É uma técnica quantitativa, que visa descobrir associações

entre as escolhas dos consumidores em relação a diferentes produtos (MOSTAFA, 2015) (MUSALEM; ABURTO; BOSCH, 2018), para verificar quais produtos são vinculados, ou seja, se o produto A for comprado qual é a probabilidade de comprar o produto B.

Com os resultados da análise é possível desenvolver campanhas de *marketing* para um nicho específico de clientes, visando influenciar o comportamento de compra e estimulando a demanda (SOLNET; BORTUG; DOLNICAR, 2016).

Assim as minerações de regras de associação podem ser definidas como um meio de encontrar regras de associação em um banco de dados de transações, sendo que estas associações devem atender a algumas restrições pré-estabelecidas (WENG, 2016). Com esta técnica é possível definir os produtos que os clientes provavelmente comprem posteriormente, ou seja, determinar o comportamento de compra, pois permite a descoberta de padrões não óbvios e geralmente ocultos (MUSALEM; ABURTO; BOSCH, 2018). Além do mais, permite extrair quais produtos e categorias de produtos tendem a ser comprados juntos, como também determinar quais produtos são determinantes para compra de produtos específicos (MUSALEM; ABURTO; BOSCH, 2018).

A análise auxilia na aquisição e retenção de clientes, pois trabalha como um facilitador para que se entenda as necessidades dos clientes, permitindo criar melhores promoções e incentivos e reduzindo a perda de clientes para a concorrência (HEMALATHA, 2012). O que corrobora para o aumento da lucratividade do negócio, já que, conforme relatado por Faed e Forbes (2011), um aumento de 5% na retenção de clientes reflete em um incremento de 25 a 85% na lucratividade.

Também é possível identificar os principais produtos, aqueles que podem prejudicar os negócios se estiverem indisponível ou mais caros (UNVAN, 2020).

Além de todos os benefícios ora expostos, considerando os algoritmos, é importante destacar que o MBA não está vinculado a premissas de linearidade ou normalidade, dentre outras como exigidas por outros métodos. Outra vantagem é que as regras de associação são menos influentes a dados discrepantes se comparadas a outras abordagens analíticas (MUSALEM; ABURTO; BOSCH, 2018).

Contudo, apesar de todas as vantagens, algoritmos de recomendação de produtos necessitam de dados históricos de compras para ser possível prever as intensões futura (WANG, 2015), pois extraem o conhecimento a partir de informações

de transação de produtos (LIU; SHIH, 2005b). Além disso, o MBA não é imune ao problema de ausência de valores.

2.5.3 Algoritmos de associação

Algoritmos baseados em regras de associação, procuram nas bases de dados de transações (vendas de produtos), regras que permitam reconhecer padrões para realizar a recomendação de produtos. Existem vários algoritmos com esta finalidade, tais como o Apriori e os de Filtragem Colaborativa.

2.5.3.1 Apriori

Publicado em 1994 por Agrava e Srkikant, é um dos mais conhecidos segundo Babu e Bhuvaneswari (2014). Este algoritmo não se aplica a dados numéricos.

Em geral os passo do algoritmo são:

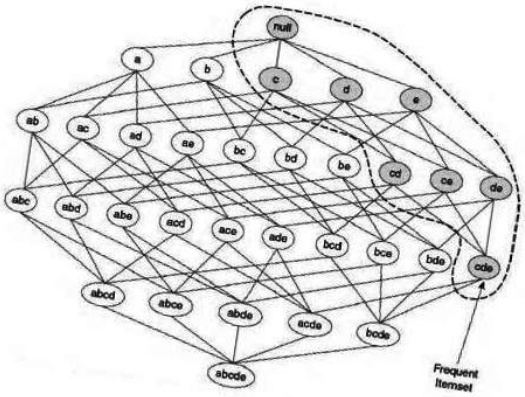
1) Gerar um conjunto de itens frequentes que tem um suporte acima do mínimo;
2) Gerar todas as regras de associação confiáveis, ou seja, acima da confiança mínima.

Detalhando melhor o processo, primeiramente gera-se o conjunto de candidatos com somente um item ($k = 1$), produto e a frequência deles. O conjunto de dois produtos ($k = 2$) é gerado analisando os indicadores de suporte e confiança mínimos dos candidatos com somente um produto.

Assim, os candidatos frequentes de dois produtos será um subconjunto do conjunto de somente um produto que continham os indicadores de suporte e confiança iguais ou maiores do que os apresentados. Repete-se este processo até a finalização das verificações, sempre armazenando os candidatos frequentes, ou seja, as relações que apresentaram indicadores de suporte e confiança maiores ou iguais do que os indicados pelo pesquisador.

Esta forma de criação dos conjuntos frequentes faz parte do princípio do algoritmo, em que se um conjunto de itens é frequente, então todos os seus subconjuntos também devem ser frequentes (TAN; MICHAEL; VIPIN, 2016), conforme ilustra a FIGURA 3.

FIGURA 3 - PRINCÍPIO DO ALGORITMO APRIORI



FONTE: Tan, Michael e Vipin (2016).

Desta forma o algoritmo faz inúmeras passagens pela base de dados (ZHANG; ZHANG, 2002), e devido à sua natureza combinatória, se houver um conjunto de itens muito grande, pode ser difícil processar os dados (WITTEN et al., 2017).

Embora muito conhecido, este algoritmo apresenta certas dificuldades:

a) Se ele for inicializado com uma regra muito alta, não será possível encontrar regras que envolvam itens pouco frequentes ou raros;
b) Para encontrar regras que envolvam ambos os produtos (frequentes e pouco frequentes), o algoritmo deve inicializar com um suporte baixo;
c) O número de varreduras no banco aumenta à medida que o banco de dados aumenta.

2.5.3.1.1 Qualidade de associações

O Apriori é baseado em regras de associação entre produtos, e para encontrar associações interessantes e eliminar as sem relevância são utilizados alguns indicadores para medir o quão forte as associações são estabelecidas, são eles: Suporte e Confiança.

Para detalhar os indicadores é necessário determinar algumas variáveis, sejam:

- $X = \{x_1, x_2, ..., x_m\}$ um conjunto de produtos de uma empresa;
- D: um conjunto de transações de vendas, sendo que cada transação registra os produtos comprados anteriormente pelo cliente.

Uma regra de associação é uma implicação do tipo $x_1 \rightarrow x_2$, dado que $\{x_1$ e $x_2\} \subset X$ e $x_1 \cap x_2 = \emptyset$. Weng (2016) definiu estes indicadores como:

a) Suporte: Utilizado para mediar quão significativa é a regra, é calculado como sendo a porcentagem de transações em D que contém $x_1 \cup x_2$, conforme a equação (10). Uma desvantagem deste índice, segundo Musalem, Aburto e Bosch (2018), é que sua utilidade diminui muito na presença de um conjunto muito grande de transações e com milhares de produtos. Desta maneira, é provável que os valores do indicador sejam muito baixos, pois a presença de outras transações serve como ruído no conjunto de dados, o que dificulta distinguir a força de uma associação.

$$Sup(x1 \cup x2) = \frac{|(x_1 \cup x_2)|}{|D|} \tag{10}$$

b) Confiança: Mede a probabilidade das relações e calcula-se como a porcentagem de transações em D que contenham $x_1 \cup x_2$ pela porcentagem que contém x_1, de acordo com a equação (11). Ao contrário do suporte, a confiança é útil mesmo em conjunto de dados muito grandes.

$$Conf(x1 \rightarrow x2) = \frac{Sup(x_1 \cup x_2)}{Sup(x_1)} \tag{11}$$

Desta maneira, para realizar a análise de um conjunto de produtos considerado frequente, deverá ser definida uma medida de Suporte e Confiança (LIU; SHIH, 2005a).

Em geral os indicadores Suporte e Confiança tem alguns problemas, conforme descrito por Chiang et al. (2011):

a) Todos os produtos podem não ser comercializados ao mesmo tempo (ex: produtos novos), assim os comercializados posteriormente têm suporte e confiança menores. Desta maneira, apesar de poder haver um interesse maior nos produtos recém comercializados, as regras de associação podem não ser descobertas, devido não alcançarem os níveis de suporte e confiança mínimo;
b) A quantidade excessiva de produtos, que também reduz o indicador de suporte. Para solucionar este quesito, as regras de associação podem ser formuladas considerando as categorias de produtos, o que se denomina regra multinível;

c) Quando as transações ocorrem em diferentes lojas, pois há possibilidade de alguns produtos não serem vendidos em todas (CHEN et al., 2005).

2.5.3.2 Filtragem Colaborativa

A Filtragem Colaborativa (FC) prediz se um cliente gostaria ou não de um novo item, prevendo como eles classificariam este item (JALILI et al., 2018). Esta técnica recomenda produtos com base na semelhança entre os clientes ou produtos (chamados vizinhos), assim o termo “colaborativa” vem da necessidade de filtrar clientes/produtos semelhantes, ou seja, estes dados colaboram na tarefa de recomendação.

Neste trabalho será visto dois tipos de Filtragem Colaborativa: a Baseada no Item e a Baseada no Usuário.

Segundo Gorakala e Usuelli (2015), existem algumas limitações nesta técnica:

a) Um cliente novo, que ainda não comprou nenhum produto, não terá previsões. Ambos os algoritmos precisam calcular a semelhança, e ela é calculada com base em produtos já comprados;
b) Produtos novos, ainda não adquiridos por nenhum cliente, não serão recomendados. É importante notar que a técnica utiliza dados históricos de produtos já consumidos.

Outra limitação destacada por Han, Kamber e Pei (2012), é em relação a um conjunto de produtos muito grande, pois reduz a chance de usuários possuírem itens comuns.

2.5.3.2.1 Filtragem Colaborativa Baseada no Item

Este tipo de filtragem colaborativa considera que os usuários preferem itens similares aos demais itens que gostaram (HAHSLER, 2011). Desta maneira o método considera o conjunto de itens que o usuário, ao que se quer recomendar, avaliou e verifica o quanto cada um deles é similar a outros itens não comprados (GORAKALA; USUELLI, 2015).

Assim a primeira fase é calcular uma matriz de similaridade contendo todos os itens avaliados, para isto deverá ser utilizado uma medida de similaridade como por exemplo Cosseno, Jaccard ou Pearson.

Para reduzir a complexidade do modelo, para cada item apenas uma lista dos k itens mais semelhantes e seus valores de similaridades são armazenados. O conjunto dos k itens mais semelhantes ao item i são denominados como a vizinhança do item i.

Após a definição da vizinhança, é necessário calcular estimativas das notas para os itens ainda não adquiridos, para que se possa gerar as previsões. Uma das técnicas utilizadas é a soma ponderada, que consiste em calcular as estimativas, pela ponderação da nota dada ao item adquirido pela similaridade do item ainda não comprado.

Na FIGURA 4 é ilustrado um exemplo de Filtragem Colaborativa Baseada no item. Neste caso, há oito itens que tem a sua similaridade calculada, sendo identificado, na linha, os três mais similares (em negrito). Após isto é verificado os itens avaliados, ou seja, a nota dada pelo usuário alvo u_a (em laranja), o qual quer se sugerir um novo produto, ainda não comprado. Para cada item avaliado e não comprado é realizada uma média ponderada relacionando à similaridade do item versus a nota dada.

Desta forma, nota-se que o primeiro item a ser recomendado é o i_3, pois possui a maior avaliação estimada ($\hat{r}_\alpha = \frac{(0{,}4*4+0{,}5*5)}{9} = 4{,}6$).

FIGURA 4 - EXEMPLO DE FILTRAGEM COLABORATIVA BASEADA NO ITEM

S	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	$\hat{r}_a$	$k=3$
i_1	-	0.1	0	**0.3**	**0.2**	**0.4**	0	0.1	-	
i_2	0.1	-	**0.8**	**0.9**	0	**0.2**	0.1	0	0.0	
i_3	0	**0.8**	-	0	**0.4**	0.1	0.3	**0.5**	4.6	
i_4	**0.3**	**0.9**	0	-	0	0.1	0	**0.2**	3.2	
i_5	**0.2**	0	**0.4**	0	-	0.1	**0.2**	0.1	-	
i_6	**0.4**	**0.2**	0.1	**0.3**	0.1	-	0	0.1	2.0	
i_7	0	**0.1**	**0.3**	0	**0.2**	0	-	0	4.0	
i_8	**0.1**	0	**0.5**	**0.2**	0.1	0.1	0	-	-	
u_a	2	?	?	?	4	?	?	5		

Fonte: Hahsler (2011).

A Filtragem Colaborativa Baseada em Item é mais eficiente computacionalmente do que a baseada no usuário, pois o modelo é relativamente pequeno, assim ele consome menos recursos uma vez que não há necessidade de analisar todos os índices de similaridade entre os usuários mas somente similaridade entre os itens (JALILI et al., 2018).

2.5.3.2.2 Filtragem Colaborativa Baseada no Usuário

Neste modelo a suposição é que os usuários com preferências afins classificarão itens de maneira semelhante (HAHSLER, 2011), ou seja, identifica-se usuários com o mesmo perfil de comportamento. Deste modo, as avaliações feitas por vários indivíduos são utilizadas para prever as classificações ausentes, isto é feito por meio de uma vizinhança de usuários semelhantes.

Para definir a vizinhança, primeiramente é calculada à similaridade entre usuários, em seguida é determinado o número de usuários mais semelhantes (k vizinhos mais próximos).

Na FIGURA 5 é ilustrado um exemplo do funcionamento da Filtragem Colaborativa Baseada no Usuário, em que é demonstrado as notas (avaliações dadas aos produtos) de seis usuários (u_n) para até oito dos itens (i_n).

Para um determinado usuário alvo u_a (em laranja), o qual objetiva-se recomendar produtos, é calculado à similaridade com os demais pelas avaliações dadas aos produtos. Neste exemplo foi determinado três vizinhos mais próximos, que são o usuário u_1, u_2 e u_4, que possuem maior similaridade (S_α) com o usuário alvo. A nota média destes vizinhos é utilizada como avaliação estimada ($\hat{r}_\alpha$) do usuário alvo. Assim as estimativas de notas são utilizadas para recomendar produtos.

FIGURA 5 - EXEMPLO DE FILTRAGEM COLABORATIVA BASEADA NO USUÁRIO

R

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	?	4.0	4.0	2.0	1.0	2.0	?	?
u_2	3.0	?	?	?	5.0	1.0	?	?
u_3	3.0	?	?	3.0	2.0	2.0	?	3.0
u_4	4.0	?	?	2.0	1.0	1.0	2.0	4.0
u_5	1.0	1.0	?	?	?	?	?	1.0
u_6	?	1.0	?	?	1.0	1.0	?	1.0
u_a	?	?	4.0	3.0	?	1.0	?	5.0
$\hat{r}_a$	3.5	4.0			2.3		2.0	

S_a

	u_a
u_1	0.3
u_2	1.0
u_3	0.2
u_4	0.3
u_5	0.1
u_6	0.1

FONTE: Hahsler (2011).

2.5.3.2.3 Filtragem Colaborativa Utilizando Dados Binários

Como visto nos métodos supracitados (baseado no item e no usuário), eles necessitam avaliações de clientes sobre os produtos adquiridos. Porém, se não houver avaliações, ou houver poucas, as recomendações devem ocorrer somente avaliando o comportamento do cliente (HAHSLER, 2011). O fato de não haver avaliação é bastante comum, exemplos disto são a visualização de páginas de internet, a compras em supermercados, etc.

Neste caso, se configura 1 (um) como produto comprado e 0 (zero) como não comprado, porém este não comprado pode significar que ele não conhece, ou não gosta do produto. Assumindo o zero como ausência de informação, uma medida de similaridade como a de Jaccard pode ser utilizada para recomendar os produtos.

2.5.3.3 Métricas de Avaliação de Algoritmos de Recomendação

Dado um conjunto de modelos de recomendação, é necessário decidir qual é o melhor, que deve ser implantado. Para avaliar o desempenho dos modelos, há um conjunto de métricas a serem utilizadas (GUNAWARDANA; SHANI, 2009). Para serem calculada, elas utilizam uma matriz de confusão (FIGURA 6), que é uma espécie de tabela de contingência, que apresenta os dados em previstos (recomendados pelo algoritmo) e real.

FIGURA 6 - MATRIZ DE CONFUSÃO

		ACTUAL	
		POSITIVE	**NEGATIVE**
PREDICTED	**POSITIVE**	*TRUE POSITIVE*	*FALSE POSITIVE*
	NEGATIVE	*FALSE NEGATIVE*	*TRUE NEGATIVE*

Fonte: Gorakala e Usuelli (2015)

A matriz apresenta quatro indicadores que são descritos por Jalili et al. (2018) como:

a) Verdadeiro Positivo (TP): São considerados os itens relevantes recomendados pelo sistema;
b) Verdadeiro Negativo (TN): São os itens irrelevantes que são corretamente não recomendados;

c) Falso Positivo (FP): São os itens que são irrelevantes que são incorretamente recomendados;
d) Falso Negativo (FN): Itens relevantes que são incorretamente não recomendados pelo sistema (sistema falhou em não recomendar).

Os indicadores apresentados na Matriz de Confusão dão origem a uma série de métricas utilizadas na avaliação dos modelos de recomendação, como:

a) Precisão: É calculada como sendo a proporção de recomendações corretas (TP) sobre o total de itens recomendados (TP e FP), demonstrada na equação (12). Esta métrica evidência a proporção de recomendações que foram realmente corretas, ou seja, quão bem o modelo recomendou.

$$Precisão = \frac{TP}{(TP + FP)} \tag{12}$$

b) Recall: Também chamado de Taxa de Verdadeiros Positivos (TPR), é calculado conforme equação (13), que avalia a proporção de itens relevantes que forma recomendados (TP), sobre todos os itens relevantes, recomendados ou não (TP+FN). Assim a métrica verifica a proporção de itens relevantes que foram correntemente identificados.

$$Recall\ ou\ TPR = \frac{TP}{(TP + FN)} \tag{13}$$

Conforme Gunawardana e Shani (2009), permitir uma grande quantidade de itens recomendados melhora o Recall, porém provavelmente reduza a Precisão. Para estabelecer o melhor modelo, estes dois indicadores podem ser comparados graficamente para uma determinada quantidade de itens recomendados.

Também há a Taxa de Falso Positivo (FPR), utilizada juntamente com a TPR na análise da Curva ROC:

c) Taxa de Falso Positivo (FPR): avalia a proporção de itens irrelevantes que são recomendados (FP) sobre o total de itens irrelevantes (FP+TN).

$$FPR = \frac{FP}{(TN + FP)}$$

2.5.3.3.1 CURVA ROC

ROC (*Receiver Operating Characteristic*) é uma representação gráfica da probabilidade de detecção do sistema, também chamada de sensibilidade ou Taxa de Verdadeiros Positivos (TPR), pela probabilidade de alarme falso, também chamada de Taxa de Falsos Positivos (FPR) (HAHSLER, 2011).

Uma forma possível de analisar a eficiência de dois sistemas é comparando o tamanho da área sob a curva, em que uma área maior indica melhor desempenho.

3 METODOLOGIA

A metodologia da pesquisa discorre sobre o percurso metodológico adotado, e evidencia o método inerente a pesquisa.

3.1 MÉTODO DE PESQUISA

A respeito da classificação da pesquisa, esta refere-se a uma pesquisa aplicada dado que aborda sobre a problemática de segmentação de clientes, que é recorrente na maioria das empresas, desta forma o resultado desta pesquisa tem como objetivo descrever uma metodologia prática que pode ser utilizada pelas empresas que tenham este dilema.

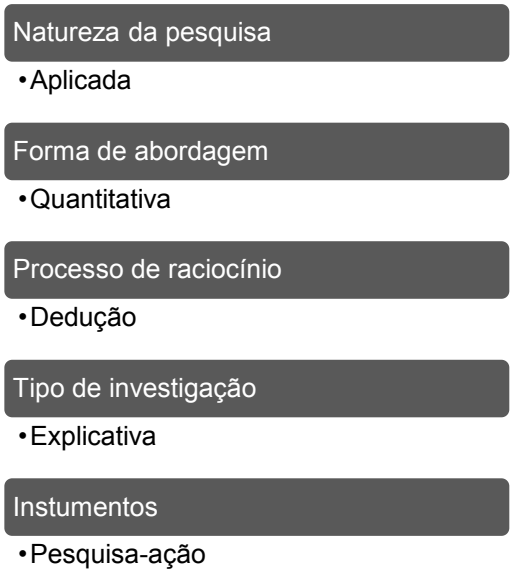
Já em relação a sua forma de abordagem, a pesquisa em questão é quantitativa, pois utiliza variáveis quantificáveis em seus métodos.

Quanto ao processo de raciocínio, classifica-se como dedutiva, já que se baseia na hipótese que há diferenças entre clientes, de forma que existe a possibilidade de segmentá-los em grupos, garantindo maior efetividade nas campanhas de *marketing*, bem como recomendar produtos com base em regras de associação.

Também é possível denominar esta pesquisa como explicativa, quanto ao seu tipo de investigação, visto que, um de seus objetivos é identificar os fatores fundamentais para a segmentação.

Da mesma forma, no que concerne aos procedimentos técnicos, pode-se identificá-la quanto aos seus procedimentos técnicos como pesquisa-ação, dado que há envolvimento do pesquisador no problema que está sendo estudado. Isto posto, a FIGURA 7 resume a classificação da pesquisa quanto aos aspectos metodológicos.

FIGURA 7 - CLASSIFICAÇÃO DA PESQUISA



FONTE: A autora (2021).

3.2 ETAPAS METODOLÓGICAS

Nesta seção serão abordadas as etapas metodológicas inerentes a pesquisa, discorrendo-se sobre a unidade de análise, seleção do público-alvo, detalhamento do método e etapas da pesquisa.

3.3 UNIDADE DE ANÁLISE

A pesquisa tem como unidade de análise os clientes de uma empresa. Neste contexto, de forma a desenvolver o projeto de estudo serão utilizados dados de uma empresa B2B (*Business to business*), ou seja, os clientes da empresa estudada são pessoas jurídicas.

3.4 SELEÇÃO DO PÚBLICO-ALVO

Este estudo utilizou dados dos clientes de uma empresa de serviços de apoio a gestão empresarial. Os clientes selecionados foram atendidos entre 2017 a início de 2020 e contou somente com clientes presentes na base de *score* da empresa, ou seja, pontuaram em uma metodologia que analisa variáveis de tempo de

relacionamento e consumo. A seleção foi executada diretamente em banco de dados Oracle, de forma a extrair todas as variáveis necessárias para execução da pesquisa.

3.5 DESCRIÇÃO DO MÉTODO DE PESQUISA

O método de pesquisa foi dividido em sete etapas conforme FIGURA 8, detalhadas na sequência.

Etapa 1: Iniciou com a especificação do problema de pesquisa, pois é ele que norteou toda a execução do trabalho;

Etapa 2: Dedicada à revisão de literatura, que teve o objetivo estabelecer quais as técnicas mais utilizadas para clusterizar os clientes e recomendar produtos. Nesta fase foi necessária uma pesquisa aprofundada nas bases de dados (Web of Science e Scopus) com propósito de conhecer os métodos a fim de desenvolver uma metodologia específica;

Na revisão verificou-se que o algoritmo mais utilizado para clusterização foi o K-means que apareceu em 61% dos artigos analisados, já para recomendação de produtos o Apriori, Filtragem Colaborativa e FP-Growth foram encontrados em 29%, 21% e 21% dos artigos respectivamente;

Etapa 3: Os algoritmos foram analisados e escolhidos dada a sua maior frequência na revisão sistemática, bem como na sua possibilidade de aplicação, dada as características das variáveis disponíveis na base de dados.

Etapa 4: Extração da base de dados que foi utilizada;

Etapa 5: Foi realizada uma análise de cada variável extraída, com objetivo de verificar as medidas de dispersão, com a finalidade de encontrar pontos influentes, possíveis *outliers*, quantificar observações faltantes e desenvolver um processo de limpeza dos dados. Além disso, foi desenvolvida uma análise descritiva com o objetivo de conhecer mais os dados antes de utilizar os métodos propostos. Esta etapa é importante para a melhor escolha dos algoritmos uma vez que demonstra para o pesquisador detalhes sobre os dados.

Nesta etapa, foi observado que a base era mista (possuía variáveis quantitativas e qualitativas), o que impediu a utilização da distância euclidiana, por ser aplicada somente a variáveis quantitativas. Desta maneira, o K-means foi substituído pelo K-medoids por ser mais flexível a diferentes tipos de distância.

Etapa 6: Nesta etapa as variáveis foram tratadas para inclusão nos algoritmos;

Para a clusterização esta etapa contou com a aplicação de técnicas como padronização das variáveis, transformação de variáveis quantitativas e qualitativas em *dummies*;

Para recomendação, houve a transformação da base de transações em binária, utilizando-se 1 (um) quando o produto foi consumido e 0 (zero) caso o contrário;

Etapa 7: Aplicação dos algoritmos escolhidos, para realizar a clusterização dos clientes e para recomendar produtos.

Nesta etapa a clusterização contou com a aplicação das medidas de distância Gower e Jaccard, escolhidas por serem aplicáveis em bases mista. Em ambas as matrizes de distância foi aplicado o algoritmo K-medoid, sendo o número de clusters e a qualidade mensurada pelo Coeficiente de Silhueta e Índice de Davies Bouldin.

Para recomendação de produtos, foi aplicado, em cada cluster, e comparado os algoritmos Apriori e Filtragem Colaborativa, porém virtude do tempo, não foi possível executar análises utilizando o FP-Growth.

Os algoritmos de recomendação foram utilizados com o objetivo de auxiliar no processo de fidelização dos clientes, dado que como a metodologia precede uma base histórica de compras, somente clientes que já adquiriram produtos, podem receber alguma recomendação, da mesma forma só produtos que já foram utilizados, são recomendados.

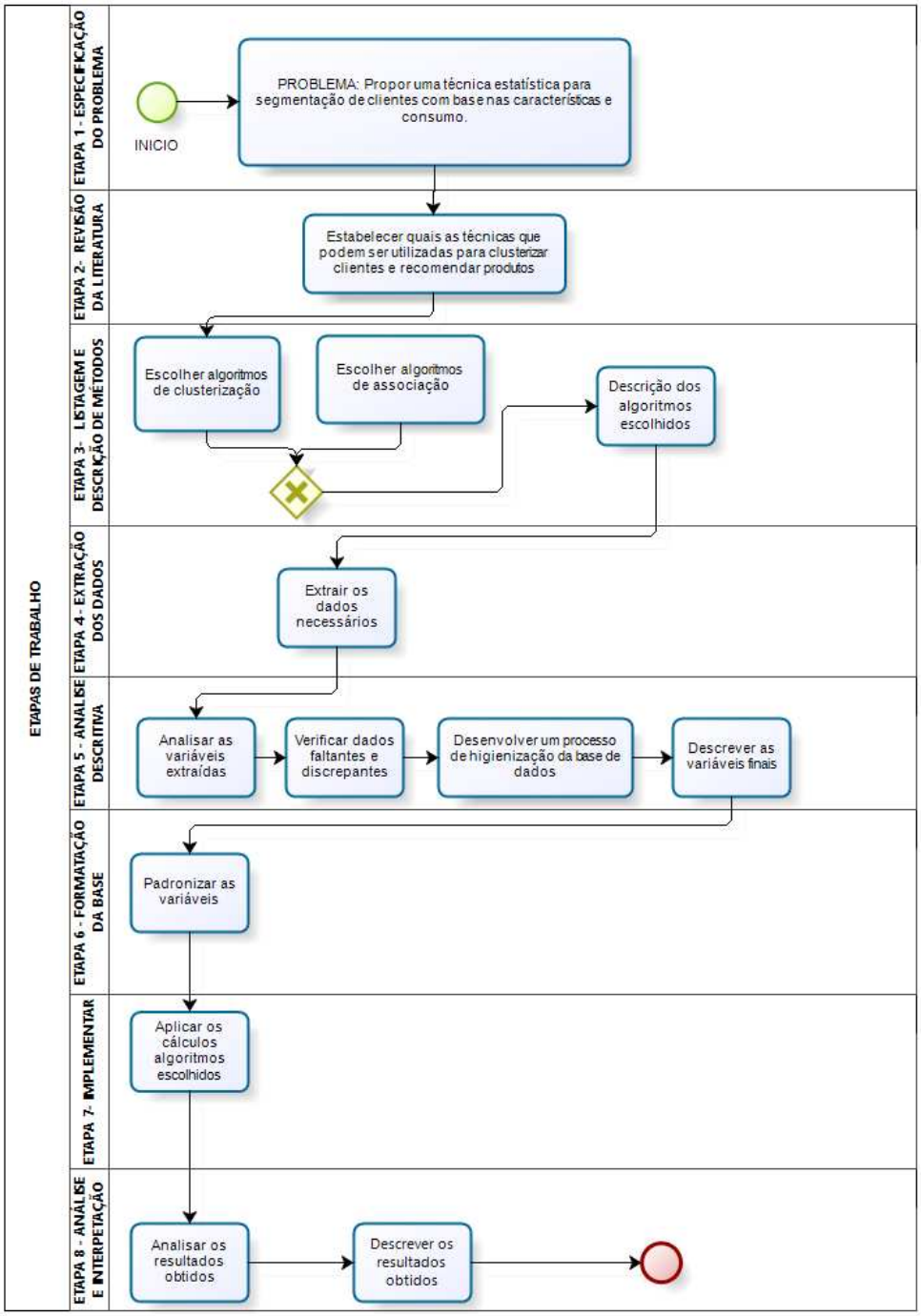
Para mensurar o desempenho dos algoritmos de recomendação, foi utilizada uma avaliação *off-line,* ou seja, foi utilizado dados históricos já levantados nas transações. Esta é uma alternativa para testar o algoritmo mais adequado, dado a onerosidade de testes *on-line*. Nela alguns dados de compras são ocultados, com a finalidade de simular a atividade *on-line*, em que é feita a recomendação e o usuário interage utilizando ou não o produto (GUNAWARDANA; SHANI, 2009).

A escolha do dos algoritmos levou em conta métricas como Precisão e Recall. Também foi buscada melhorar o desempenho nos resultados dos algoritmos de recomendação com ajuste nos parâmetros.

Etapa 8: Análise e interpretação dos resultados.

Após a aplicação de todas as etapas foi possível os resultados para a proposição de uma metodologia que combinasse técnicas de clusterização de clientes e recomendação de produtos.

FIGURA 8 - ETAPAS DA PESQUISA



FONTE: A autora (2021).

3.6 CONJUNTO DE DADOS

A base de dados a utilizada compreende 339.061 clientes Pessoa Jurídica (PJ), 1.739.805 transações de 3.033 produtos. Cada transação possui o código do cliente Pessoa Jurídica, o código da Pessoa Física (PF) que consumiu, a data do consumo, o local em que ocorreu, o código do produto e o tipo de produto (consultoria, curso, palestra etc.).

É importante salientar que os produtos empregados têm o propósito de auxiliar o empresário a melhorar a gestão do negócio, podendo este aperfeiçoamento incidir em várias áreas, como *Marketing*, Recursos Humanos, Vendas, Finanças, Legislação, Processos de Produção, Gestão Ambiental, entre outros. Estas áreas estão contidas nas classificações dos produtos.

Além da base de transações, há outras bases que farão parte da análise, tais como a base de:

a) Perfil, em que estão contidas as informações sobre Data de Fundação da Empresa, Código Nacional de Atividade Empresarial (CNAE), Natureza Jurídica, Porte etc.;
b) Produtos: que contêm algumas características dos produtos, tais como, área de conhecimento, tema etc.;

Os dados serão analisados e os algoritmos processados utilizando o *software R* (R CORE TEAM, 2018), em um *notebook* que utiliza sistema operacional Windows 10, processador Intel Core i3 e 8 GB de memória RAM.

4 RESULTADOS

4.1 METODOLGIA PROPOSTA

A metodologia foi proposta para uma base de dados mista, ou seja, que possui variáveis quantitativas e qualitativas, sendo segmentada em duas fases, sendo a primeira referente a clusterização de clientes e a segunda sobre recomendação de produtos.

4.1.1 Primeira fase - clusterização

A fase de clusterização de clientes têm seis etapas, conforme evidenciado na FIGURA 9, que são descritas a seguir:

- Etapa 1: Análise descritiva – Destinada a análise descritiva de todas as variáveis da base de dados. A análise auxilia o pesquisador a compreender melhor as variáveis, sendo nesta fase observados itens como, medidas de tendência central, quantidade de dados faltantes etc. Após a compreensão, o pesquisador consegue escolher as variáveis que serão utilizadas na clusterização, bem como selecionar os algoritmos que serão utilizados;
- Etapa 2: Preparação dos dados – Com as variáveis escolhidas, é necessário preparar os dados para inclusão nos algoritmos.
- Etapa 3: Criação da matriz de distância – Após preparação dos dados, pode-se calcular a matriz de distância, que é um elemento necessário para a clusterização;
- Etapa 4: Aplicação do algoritmo de Clusterização e Análise da quantidade de agrupamentos – Com a matriz de distância o algoritmo de clusterização pode ser aplicado. Nesta fase são geradas diversas possibilidades de agrupamento, que são analisadas considerando índices de qualidade que avaliam a dispersão interna e externa do cluster, desta forma é possível escolher o melhor agrupamento;
- Etapa 6: Análise dos dados – Escolhida a melhor quantidade de agrupamentos, eles são analisados com a intensão de verificar as

características mais proeminentes, de forma a estabelecer um rótulo que identifique mais facilmente o perfil de cliente determinante em cada cluster.

FIGURA 9 - METODOLOGIA PRIMEIRA FASE - CLUSTERIZAÇÃO

FONTE: A autora (2021).

4.1.2 Segunda fase - recomendação

Com os clusters criados, a segunda fase trata-se da recomendação de produtos para os clientes de cada um dos agrupamentos. Esta fase tem seis etapas conforme a FIGURA 10 sendo as etapas descritas a seguir:

- Etapa 1: Análise da base de dados - Destina-se a análise das transações, com o objetivo de verificar a quantidade de produtos, vendas em cada cluster;

- Etapa 2: Preparação da base – Trata-se da preparação da base de dados para inclusão nos algoritmos de recomendação;
- Etapa 3: Definição da forma de avaliação – É de fundamental importância que seja definido um método de avaliação para que se possa comparar os algoritmos testados;
- Etapa 4: Algoritmos e parâmetros testados – São definidos os algoritmos de recomendação que serão testados, e os parâmetros que serão utilizados;
- Etapa 5: Análise dos resultados - Nesta fase são definidas as métricas que serão avaliadas para a escolha algoritmo;
- Etapa 6: Ajustes dos parâmetros - pós a análise dos resultados e escolha do melhor algoritmo, pode-se testar diferentes valores de parâmetros com o objetivo de melhorar os resultados.

FIGURA 10 – METODOLOGIA SEGUNDA FASE - RECOMENDAÇÃO

FONTE: A autora (2021)

4.2 PRIMEIRA FASE: CLUSTERIZAÇÃO DOS CLIENTES

Nesta seção será apresentada cada etapa da primeira fase da metodologia proposta, ou seja, o processo para clusterização da base de clientes.

4.2.1 Etapa 1: Análise da base de dados

Antes da utilização de técnicas de clusterização foi realizada uma série de análises nos dados disponíveis na base de Perfil, que apresenta algumas características dos clientes PF (Pessoa Física) e PJ (Pessoa Jurídica), conforme descrito na etapa 5 do método de pesquisa, de forma a compreender melhor os dados que poderiam ser utilizados para clusterizar os clientes. É importante salientar que esta base representa o perfil dos clientes atendidos nos anos de 2017 a 2020, compreendendo um total de 966.194 clientes, sendo 627.130 PF e 339.064 PJ.

4.2.1.1 Análise das variáveis

O QUADRO 5 apresenta a descrição das variáveis da base Perfil que foram analisadas.

QUADRO 5 - DETALHAMENTO DAS VARIÁVEIS DA BASE DE PERFIL

Nome	Descrição
CODIGO	É a variável de identificação do cliente, representa o código que pode ser de uma PJ ou PF.
ATIVO	Variável utilizada para verificar se a empresa está ativa, ou seja, em operação. Também é utilizada para PF.
TIPO	Tipo do cliente, pode adotar uma das seguintes classificações: 1- Física – Utilizada para segmentar PF; 2- Formal – Utilizada para determinar se um PJ possui atividade formalizada, com CNPJ (Cadastro Nacional de Pessoa Jurídica); 3- Informal - Utilizada quando a PJ não possui atividade formalizada, ou seja, não possui CNPJ; 4- Produtor rural – Há casos que a PJ não possui CNPJ, mas a atividade é formalizada por um CADPRO (Inscrição no Cadastro de Produtor Rural).
PORTE	Categoria das empresas segundo o faturamento: MEI (Microempreendedor Individual) - Fatura até R$ 81 mil reais/ano; ME (Microempresa) - Fatura até R$ 360 mil/ano; EPP (Empresa de pequeno porte) - Fatura de R$ 360 mil a R$ 4,8 milhões/ano; MGE (Média e Grande empresa) - Fatura mais de 4,8 milhões/ano.
NÚMERO_COLABORADORES	Contabiliza o número de funcionários das empresas.
DATA_ANIVERSARIO	Data de abertura da PJ ou Nascimento da PF.
SETOR	Representa o setor de atuação da empresa: Indústria, Comércio, Serviço ou Agropecuária.
GRUPO_NATUREZA_JURIDICA	Denota o código de natureza jurídica da PJ, ou seja, o regime jurídico em que ela se enquadra: Administração pública, Entidades empresariais, Entidades sem fins lucrativos e Pessoas físicas.
STARTUP	Variável que indica se a empresa é *startup*. *Startup* é uma empresa que possui um modelo de negócios repetível e escalável.
CNAE_DIVISAO	Divisão do Código Nacional Atividade Empresarial (CNAE).
CNAE_CLASSE	Classe do Código Nacional de Atividade Empresarial (CNAE).
CNAE_SEGMENTACAO	Agrupamento do CNAE em segmentos empresariais.
TERRITÓRIO	Indica em qual área de abrangência territorial que a empresa está. O território abrange um aglomerado de municípios.
ESCOLARIDADE	Escolaridade da PF vinculada à PJ.
TIPO_SEXO	Gênero da PF vinculada a PJ.

FONTE: A autora (2021).

Primeiramente foi verificada a quantidade de observações duplicadas, desta forma foram eliminadas 56, resultando em 966.138 registros únicos, sendo 627.077 PF e 339.061 PJ.

4.2.1.1.1 Variável ativo

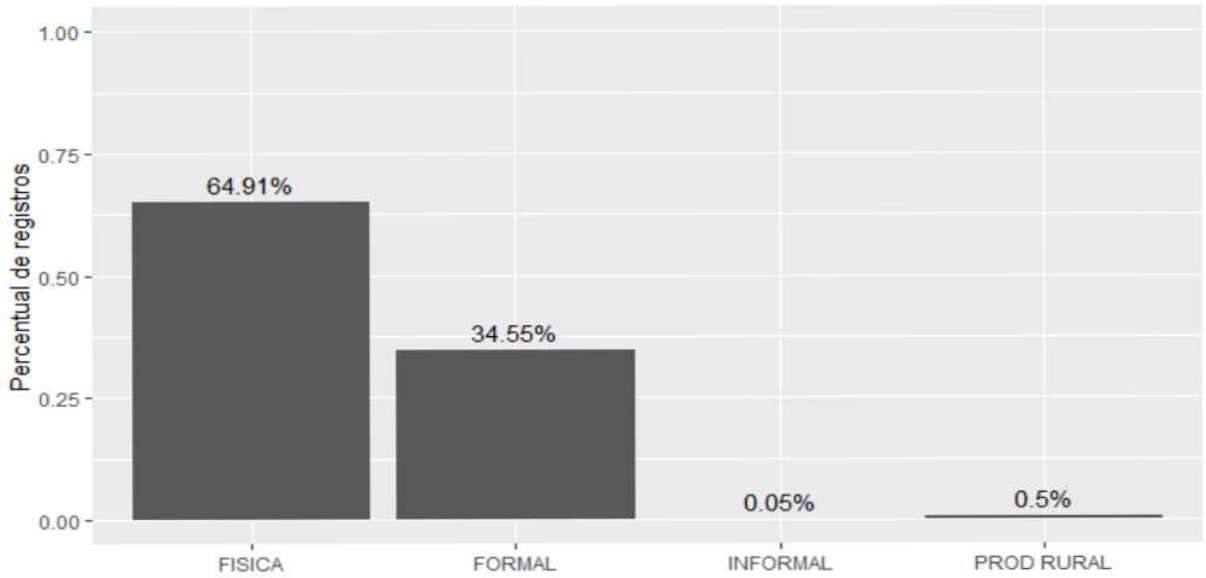
Com relação a variável ativo, como todos os clientes da base possuíam registro ativo, esta variável foi removida da análise.

4.2.1.1.2 Variável tipo

Com relação a variável tipo (FIGURA 11), 64,91% dos registros referem-se ao tipo física (Pessoa Física), sendo o restante PJ, em que 34,55% possuem o tipo formal, ou seja, são empresas com CNPJ, 0,05% são empresas informais e 0,5% produtores rurais.

Com relação a entropia, que é uma medida de dispersão aplicada em variáveis qualitativas, em que quando o resultado é próximo de zero, significa que a distribuição de frequências é mais concentrada e quando é próximo de um elas são mais equilibradas, resultou no valor de 0,2944, evidenciando que há concentração, ou seja, pouca diversidade entre classes.

FIGURA 11 - PERCENTUAL DE REGISTRO POR TIPO



FONTE: A autora (2021).

Também foi feita uma análise cruzada da variável Tipo e Porte (TABELA 4), com objetivo de avaliar possíveis inconsistências na base. Nesta análise notou-se que havia 16 clientes sem porte cadastrado, que foram removidos da base.

TABELA 4 - TIPO x PORTE

	EPP	ME	MGE	MEI	PF	NA[1]
FISICA	0	0	0	0	627.077	0
FORMAL	17.577	111.290	13.210	191.672	0	13
INFORMAL	54	275	38	84	0	3
PROD RURAL	0	4.845	0	0	0	0

FONTE: A autora (2021).

Como as estratégias de atendimento da empresa fornecedora da base de dados estão fortemente pautadas no atendimento a PJ, para este projeto foram consideradas somente as análises relativas a PJ.

Destaca-se que dos 627.077 códigos de PF existentes, 66,86% estão vinculados a uma PJ como empresários ou funcionários. Desta forma a base possui somente 21,51% de PF sem nenhum vínculo empresarial, sendo pessoas que vieram em busca de capacitação com objetivos como de abrir uma empresa futuramente, entre outros.

4.2.1.1.3 Variável porte

Com relação ao percentual de PJ por porte (TABELA 5), tem-se que 56,56% das empresas são do porte MEI (Microempreendedor Individual) e 34,33% são ME (Microempresas), estes dois portes compreendem 90,86% das empresas.

A entropia foi de 0,4211, o que evidencia uma diversidade moderada entre as classes.

TABELA 5 - PERCENTUAL DE PJ POR PORTE

MEI	ME	EPP	MGE
56,56%	34,33%	5,20%	3,91%

FONTE: A autora (2021).

[1] NA: *"Not Available",* tradução: Não disponível.

4.2.1.1.4 Variável número de colaboradores

Nota-se na TABELA 6, que há uma inconsistência, pois há empresas que possuem número de funcionários negativo na base. Além disso 13 empresas do porte ME não possuem nenhum registo de funcionários.

Também é possível notar que há MEI com mais de um funcionário, o que é proibido por lei, porém, pode ser indício de algum tipo de informalidade na contratação de funcionários.

TABELA 6 - NÚMERO DE FUNCIONÁRIOS POR PORTE

PORTE	MÍNIMO	1° QUARTIL	MEDIANA	MÉDIA	3° QUARTIL	MÁXIMO	DESVIO	NA
MEI	-2	0	0	0,07	0	10	0,27	0
ME	-4	0	1	2,88	3	1.164	8,27	13
EPP	-17	1	5	9,79	12	804	18,85	0
MGE	0	0	4	98,83	25	90.000	1.300,29	0

FONTE: A autora (2021).

Foi verificado que somente cinco empresas possuíam valores negativos em relação a quantidade de funcionários, estas e as treze empresas que não tinham nenhuma informação a respeito desta variável foram eliminadas da base.

Com relação aos MEI com mais de um funcionário, foi verificado que isto ocorreu somente em 0,3% das empresas deste porte.

Além disso nota-se que os valores máximos são bem distantes do terceiro quartil em que é possível notar que a maioria das empresas possuem poucos funcionários. Porém quanto maior o porte, maior é o desvio padrão ao redor da média.

Para remover as divergências na variável quantidade de funcionários, a princípio pensou-se em utilizar a técnica de 1,5 vezes o Intervalo Interquartílico. Porém a técnica removeria as empresas do porte MEI com um ou mais funcionários, retirando-se da base 12.557 observações ou 7% dos MEI. Assim, seriam removidos da base os MEI com um funcionário, o que é permitido pela legislação.

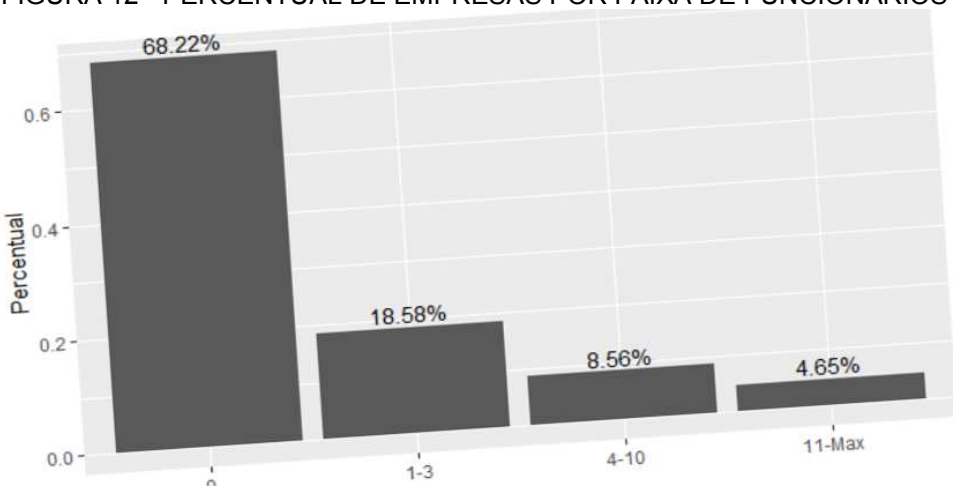
Desta forma, com o intuito de utilizar a variável e remover as discrepâncias, utilizou-se a técnica de quantização, que cria uma variável qualitativa derivada de uma quantitativa. Para esta criação, utilizou-se a média de funcionários de cada porte, criando-se assim quatro faixas, a saber: "0","1-3","4-10","11-Max”.

Na FIGURA 12 é possível verificar que 68,22% das empresas da base não possuem nenhum funcionário. O que demonstra que são empresas na sua maioria

muito pequenas, reforçando que o que foi visto na análise da variável porte, em que 56% das empresas são MEI.

Para esta variável qualitativa a entropia foi de 0,4024 o que evidencia uma moderada diversidade entre as classes. Conforme visto na FIGURA 12, a classe que não apresenta funcionários (zero), engloba quase 70% da base.

FIGURA 12 - PERCENTUAL DE EMPRESAS POR FAIXA DE FUNCIONÁRIOS



FONTE: A autora (2021).

4.2.1.1.5 Variável data de aniversário

Com relação a data de aniversário, pode-se notar na TABELA 7 que as empresas mais novas abriram em 2020, já as empresas mais velhas abriram em 1900, o que representa uma grande diferença com relação a consolidação da empresa no mercado. Também é possível notar que a mediana é de empresas que abriram em 2015, e o terceiro quartil que representa 75% das empresas da base são de 2017.

Foram removidas 14 empresas que não continham valor para a variável.

TABELA 7 - ANÁLISE DO ANO DE FUNDAÇÃO

MÍNIMO	1° QUARTIL	MEDIANA	MÉDIA	3° QUARTIL	MÁXIMO	NA
1900	2009	2015	2011	2017	2020	14

FONTE: A autora (2021).

Utilizando a data de aniversário para calcular a idade da empresa (Ano atual – Ano de aniversário) demonstrada na TABELA 8, é possível notar que 75% (terceiro quartil) tem até 11 anos de idade.

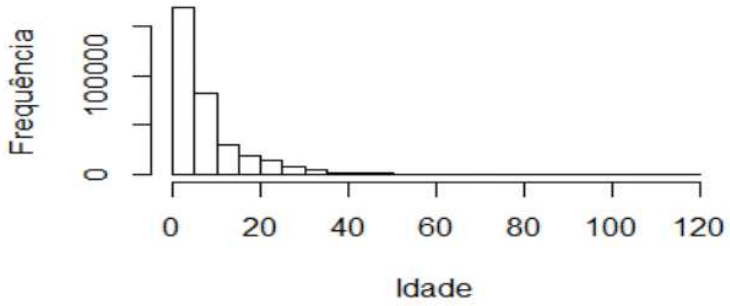
TABELA 8 - IDADE DAS EMPRESAS

MÍNIMO	1° QUARTIL	MEDIANA	MÉDIA	3° QUARTIL	DESVIO	MÁXIMO
0	3	5	9	11	8,68	120

FONTE: A autora (2021).

O histograma da FIGURA 13 também evidencia a alta concentração de empresas mais novas na base.

FIGURA 13 - HISTOGRAMA DA IDADE



FONTE: A autora (2021).

Segmentando a idade por porte (TABELA 9), nota-se que a idade é relacionada ao porte, quanto maior o porte, maior a idade da empresa.

TABELA 9 - IDADE POR PORTE

PORTE	MÍNIMO	1° QUARTIL	MEDIANA	MÉDIA	3° QUARTIL	DESVIO	MÁXIMO
MEI	0	2	3	4,1	6	2,9	62
ME	0	6	11	13,2	18	9,2	64
EPP	1	9	16	17,3	24	11,2	62
MGE	1	9	17	19,9	27	13,4	120

FONTE: A autora (2021).

Como visto no histograma, a distribuição da idade das empresas da base é desbalanceada, ou seja, há muitas empresas novas e há poucas empresas com idade mais avançada. De forma a suavizar este fato e evitar distorções na análise, a variável idade foi transformada em qualitativa, definindo-se faixas de corte de acordo com os quartis. A TABELA 10 mostra a quantidade de empresas por faixa de idade.

TABELA 10 - PERCENTUAL DE EMPRESAS POR FAIXA DE IDADE

0-2	3-5	6-10	11-Max
22,19%	28,21%	24,51%	25,10%

FONTE: A autora (2021).

A variável faixa de idade possui uma entropia de 0,5697.

4.2.1.1.6 Variável setor

Outra variável analisada foi o setor da empresa, na TABELA 11 é demonstrado a quantidade de empresas por setor. Nota-se maior presença das empresas nos setores de Comércio e Serviços. Para esta variável, a entropia calculada foi de 0,6020.

TABELA 11 - QUANTIDADE DE EMPRESAS POR SETOR

AGROPECUÁRIA	COMÉRCIO	CONSTRUÇÃO	INDÚSTRIA	SERVIÇO	NA
7.088	120.591	31.789	43.115	134.895	1.535

FONTE: A autora (2021).

Na TABELA 12 verifica-se que a maioria dos dados faltantes estão vinculados ao tipo produtor rural. Também se percebe que quase todas as empresas desta classe estão classificadas como pertencentes ao setor da agropecuária.

TABELA 12 - QUANTIDADE DE EMPRESAS POR SETOR E TIPO

TIPO	AGROPECUÁRIA	COMÉRCIO	CONSTRUÇÃO	INDÚSTRIA	SERVIÇO	NA
FORMAL	3360	120513	31776	43067	134608	412
INFORMAL	2	78	13	47	287	16
PRODUTOR RURAL	3726	0	0	1	0	1107

FONTE: A autora (2021).

Porém antes de simplesmente imputar as 1.107 empresas o setor da agropecuária, foi analisada a divisão do CNAE, conforme demonstrado na TABELA 13. Desta maneira foram imputadas no setor da agropecuária as empresas da divisão CNAE zero e um, a empresa da divisão 10 foi vinculada ao setor da indústria, enquanto a empresa da divisão 47 ao setor do comércio.

TABELA 13 - EMPRESAS SEM SETOR POR DIVISÃO CNAE

DIVISÃO CNAE	PRODUTOR RURAL
0 - Não especificado	676
1 - Agricultura, Pecuária E Serviços Relacionados	429
10 - Fabricação De Produtos Alimentícios	1
47 - Comércio Varejista	1

FONTE: A autora (2021).

Após a correção, as demais 428 empresas que não possuíam vinculação a setor, e estavam atreladas ao tipo Formal e Informal, foram analisadas em relação a variável divisão CNAE (TABELA 14).

As 21 empresas que possuíam na descrição da divisão CNAE a palavra “fabricação” (Divisão 10, 13, 22, 25, 27,31 e 32) foram vinculadas ao setor da indústria. Além disso houve seis empresas que possuíam vínculo à divisão “CNAE 41 – Construção de Edifícios” e foram incluídas no setor da indústria.

Houve 43 empresas que, embora apresentassem a variável setor faltante, possuíam na descrição das divisões CNAE a palavra “comércio” (Divisão 47,45 e 46), e desta forma foram imputadas ao respectivo setor.

Além disso, 58 empresas que possuíam descrição na divisão CNAE a palavra “serviço” e foram vinculadas ao respectivo setor. Da mesma forma 120 empresas que estavam vinculadas a Divisão CNAE com correspondência ao setor de serviços, foram vinculadas a ele, sendo: 51 empresas estavam vinculadas a divisão “53 - Correio e outras atividades de entrega”, 14 a “85 - Educação”, 13 a “70 - Atividades de sedes de empresas e de consultoria em gestão empresarial” e 13 a “86 - Atividades de atenção à saúde humana”, 6 a "94 - Atividades de organizações associativas", 8 a "69 - Atividades jurídicas, de contabilidade e de auditoria", 2 a "90 - Atividades artísticas, criativas e de espetáculos", 7 a "73 - Publicidade e pesquisa de mercado", 4 a "58 - Edição e edição integrada à impressão", 3 a "74 - Outras atividades profissionais, científicas e técnicas", 1 a "59 - Atividades cinematográficas, produção de vídeos e de programas de televisão; gravação de som e edição de música", 1 a "77 - Aluguéis não imobiliários e gestão de ativos intangíveis não financeiros". Sendo 178 empresas no total classificadas como pertencentes ao setor de serviços por meio da análise da divisão CNAE.

Restaram, após a análise e vinculação, sem classificação de setor 177 empresas que representam 0,05% da base, e devido à baixa representatividade foram excluídas.

TABELA 14 - EMPRESAS SEM SETOR POR DIVISÃO CNAE

DIVISÃO CNAE	EMPRESAS
0 - Não especificado	126
10 - Fabricação De Produtos Alimentícios	6
13 - Fabricação De Produtos Têxteis	2
14 - Confecção De Artigos Do Vestuário E Acessórios	6
18 - Impressão E Reprodução De Gravações	3
22 - Fabricação De Produtos De Borracha E De Material Plástico	1
25 - Fabricação De Produtos De Metal, Exceto Máquinas E Equipamentos	3
27 - Fabricação De Máquinas, Aparelhos E Materiais Elétricos	1
31 - Fabricação De Móveis	4
32 - Fabricação De Produtos Diversos	4
33 - Manutenção, Reparação E Instalação De Máquinas E Equipamentos	3
38 - Coleta, Tratamento E Disposição De Resíduos, Recuperação De Materiais	1
41 - Construção De Edifícios	6
43 - Serviços Especializados Para Construção	12
45 - Comércio E Reparação De Veículos Automotores E Motocicletas	10
46 - Comércio Por Atacado, Exceto Veículos Automotores E Motocicletas	4
47 - Comércio Varejista	29
49 - Transporte Terrestre	8
53 - Correio E Outras Atividades De Entrega	51
56 - Alimentação	20
58 - Edição E Edição Integrada À Impressão	4
59 - Atividades Cinematográficas, Produção De Vídeos E De Programas De Televisão, Gravação De Som E Edição De Música	1
60 - Atividades De Rádio E De Televisão	1
61 - Telecomunicações	4
62 - Atividades Dos Serviços De Tecnologia Da Informação	2
63 - Atividades De Prestação De Serviços De Informação	1
66 - Atividades Auxiliares Dos Serviços Financeiros, Seguros, Previdência Complementar E Planos De Saúde	4
69 - Atividades Jurídicas, De Contabilidade E De Auditoria	8
70 - Atividades De Sedes De Empresas E De Consultoria Em Gestão Empresarial	13
71 - Serviços De Arquitetura E Engenharia, Testes E Análises Técnicas	4
73 - Publicidade E Pesquisa De Mercado	7
74 - Outras Atividades Profissionais, Científicas E Técnicas	3
77 - Aluguéis Não Imobiliários E Gestão De Ativos Intangíveis Não Financeiros	1
79 - Agências De Viagens, Operadores Turísticos E Serviços De Reservas	1
82 - Serviços De Escritório, De Apoio Administrativo E Outros Serviços Prestados Principalmente Às Empresas	18
85 - Educação	14
86 - Atividades De Atenção À Saúde Humana	13
87 - Atividades De Atenção À Saúde Humana Integradas Com Assistência Social, Prestadas Em Residências Coletivas E Particulares	1
90 - Atividades Artísticas, Criativas E De Espetáculos	2
94 - Atividades De Organizações Associativas	6
95 - Reparação E Manutenção De Equipamentos De Informática E Comunicação E De Objetos Pessoais E Domésticos	4
96 - Outras Atividades De Serviços Pessoais	15
97 - Serviços Domésticos	1

FONTE: A autora (2021).

4.2.1.1.7 Variável grupo de natureza jurídica

A maioria da base está classificada como tendo a natureza jurídica de entidades empresariais (TABELA 15). É possível verificar que as três empresas que não tem classificação de natureza jurídica, são do tipo formal (TABELA 16) e porte MEI (TABELA 17). Como a grande maioria das empresas do porte MEI são entidades empresariais, esta foi imputada para as três empresas.

TABELA 15 - NATUREZA JURÍDICA

ADMINISTRAÇÃO PÚBLICA	ENTIDADES EMPRESARIAIS	ENTIDADES SEM FINS LUCRATIVOS	PESSOAS FÍSICAS	NA
510	330.189	3.248	4.882	3

FONTE: A autora (2021).

A entropia calculada foi de 0,06, o que demonstra baixa diversidade nas classes, ou seja, alta concentração em uma classe. O que neste caso, acontece na classe em entidades empresariais que concentra 0,97% das observações.

Além disso constata-se que a natureza jurídica referente a pessoas físicas está atrelada grande parte ao tipo produtor rural. Nas normas da Receita Federal consta que neste tipo de natureza jurídica constam além de produtores rurais, candidato a cargo político eletivo e empresa individual imobiliária.

TABELA 16 - NATUREZA JURÍDICA POR TIPO

	FISICA	FORMAL	INFORMAL	PRODUTOR RURAL
ADMINISTRAÇÃO PÚBLICA	0	509	1	0
ENTIDADES EMPRESARIAIS	0	329.772	415	2
ENTIDADES SEM FINS LUCRATIVOS	0	3.236	12	0
PESSOAS FÍSICAS	0	50	0	4.832
ORGANIZAÇÕES INTERNACIONAIS E OUTRAS INSTITUIÇÕES EXTRATERRITORIAIS	0	4	0	0
NA	0	3	0	0

FONTE: A autora (2021).

TABELA 17 - NATUREZA JURÍDICA POR PORTE

NATUREZA JÚRIDICA	MEI	ME	EPP	MGE
ADMINISTRAÇÃO PÚBLICA	0	0	0	510
ENTIDADES EMPRESARIAIS	191.613	111.498	17.624	9.454
ENTIDADES SEM FINS LUCRATIVOS	0	17	4	3.227
PESSOAS FÍSICAS	1	4.832	0	49
ORGANIZAÇÕES INTERNACIONAIS E OUTRAS INSTITUIÇÕES EXTRATERRITORIAIS	0	0	0	4
NA	3	0	0	0

FONTE: A autora (2021).

4.2.1.1.8 Variável *startup*

Há poucas empresas classificadas como *startups*, somente 456 (0,13%), o que demonstra muita concentração em apenas uma classe (entropia 0,0044). A grande maioria delas 269 (58,99%) são do porte ME (TABELA 18) e 301 (66,00%) são do setor de serviços (TABELA 19).

TABELA 18 - *STARTUP* POR PORTE

PORTE	NÃO	SIM
MEI	191.526	91
ME	116.078	269
EPP	17.574	54
MGE	13.202	42

FONTE: A autora (2021).

TABELA 19 - *STARTUP* POR SETOR

SETOR	NÃO	SIM
Agropecuária	8.191	2
Comércio	120.546	89
Construção	31.776	13
Indústria	43.092	51
Serviço	134.775	301

FONTE: A autora (2021).

4.2.1.1.9 Variável CNAE divisão

Na base há 86 classificações para esta variável. A TABELA 20 demonstra as vinte divisões do CNAE mais representativas, em que é possível notar uma maior representatividade da divisão 47, vinculada ao Comércio Varejista. Não foram encontradas empresas que não tinham a classificação.

TABELA 20 - CNAE DIVISÃO POR EMPRESAS

CNAE DIVISÃO	QTD. EMPRESAS	%
47 - Comércio Varejista	91.520	27,45%
43 - Serviços Especializados Para Construção	28.751	8,62%
96 - Outras Atividades De Serviços Pessoais	23.259	6,98%
56 - Alimentação	19.811	5,94%
45 - Comércio E Reparação De Veículos Automotores E Motocicletas	17.273	5,18%
85 - Educação	11.052	3,31%
82 - Serviços De Escritório, De Apoio Administrativo E Outros Serviços Prestados Principalmente Às Empresas	10.338	3,10%
49 - Transporte Terrestre	9.600	2,88%
73 - Publicidade E Pesquisa De Mercado	9.019	2,70%
14 - Confecção De Artigos Do Vestuário E Acessórios	8.699	2,61%
46 - Comércio Por Atacado, Exceto Veículos Automotores E Motocicletas	8.040	2,41%
95 - Reparação E Manutenção De Equipamentos De Informática E Comunicação E De Objetos Pessoais E Domésticos	6.047	1,81%
10 - Fabricação De Produtos Alimentícios	5.644	1,69%
1 - Agricultura, Pecuária E Serviços Relacionados	4.970	1,49%
25 - Fabricação De Produtos De Metal, Exceto Máquinas E Equipamentos	4.525	1,36%
86 - Atividades De Atenção À Saúde Humana	4.384	1,31%
33 - Manutenção, Reparação E Instalação De Máquinas E Equipamentos	3.830	1,15%
97 - Serviços Domésticos	3.709	1,11%
69 - Atividades Jurídicas, De Contabilidade E De Auditoria	3.670	1,10%
81 - Serviços Para Edifícios E Atividades Paisagísticas	3.256	0,98%

FONTE: A autora (2021).

4.2.1.1.10 Variável CNAE classe

Esta variável possui 632 categorias possíveis. A TABELA 21 mostra as vinte classes com mais representatividade, e nota-se que não há uma classe com uma representatividade muito elevada. Além disso, 5.411 das empresas não possuíam cadastro nesta variável e foram excluídas da análise.

TABELA 21 - CNAE CLASSE POR EMPRESAS

CNAE CLASSE	QTD. EMPRESAS	%
47814 - Comércio varejista de artigos do vestuário e acessórios	22.749	6,71%
96025 - Cabeleireiros e outras atividades de tratamento de beleza	20.784	6,13%
43991 - Serviços especializados para construção não especificados anteriormente	13.731	4,05%
56112 - Restaurantes e outros estabelecimentos de serviços de alimentação e bebidas	12.999	3,84%
45200 - Manutenção e reparação de veículos automotores	9.943	2,93%
47890 - Comércio varejista de outros produtos novos não especificados anteriormente	9.690	2,86%

73190 - Atividades de publicidade não especificadas anteriormente	8.425	2,49%
43304 - Obras de acabamento	7.776	2,29%
85996 - Atividades de ensino não especificadas anteriormente	7.593	2,24%
14126 - Confecção de peças do vestuário, exceto roupas íntimas	7.452	2,20%
49302 - Transporte rodoviário de carga	6.691	1,97%
47121 - Comércio varejista de mercadorias em geral, com predominância de produtos alimentícios - minimercados, mercearias e armazéns	5.667	1,67%
45307 - Comércio de peças e acessórios para veículos automotores	5.455	1,61%
47440 - Comércio varejista de ferragens, madeira e materiais de construção	5.080	1,50%
43215 - Instalações elétricas	5.015	1,48%
47296 - Comércio varejista de produtos alimentícios em geral ou especializado em produtos alimentícios não especificados anteriormente	4.458	1,32%
56201 - Serviços de catering, bufê e outros serviços de comida preparada	4.382	1,29%
47725 - Comércio varejista de cosméticos, produtos de perfumaria e de higiene pessoal	3.779	1,12%
97005 - Serviços domésticos	3.709	1,09%

FONTE: A autora (2021).

4.2.1.1.11 Variável segmentação CNAE

A variável segmentação CNAE foi criada pela própria empresa e trata-se de um agrupamento de CNAE, que foi executado observando metodologias já existentes, como por exemplo os agrupamentos provenientes das pesquisas do IBGE.

Nota-se na TABELA 22 que não há dados faltantes, porém cerca de 19,62% das empresas estão classificadas como outros. Não há uma classificação predominante, as empresas estão bem separadas entre as segmentações existentes, sendo as três maiores classificações voltadas para Casa e Construção com 11,36%, Tecidos, Vestuário e Calçados com 8,21% e Estética e Beleza com 6,23%.

TABELA 22 - CNAE SEGMENTAÇÃO POR EMPRESAS

CNAE SEGMENTAÇÃO	QTD. EMPRESAS	%
Outros	65.427	19,62%
Casa e Construção	37.892	11,36%
Tecidos, Vestuário e Calçados	27.369	8,21%
Estética e Beleza	20.784	6,23%
Alimentação	19.811	5,94%
Supermercados, Hipermercados, Produtos Alimentícios, Bebidas e Fumo	19.217	5,76%
Educação	11.052	3,31%
Reparação Veículos	10.607	3,18%
Transporte	10.590	3,18%
Vestuário e Moda	9.047	2,71%
Material de Construção	7.373	2,21%
Eletroeletrônico	6.956	2,09%
Artigos Farmacêuticos, Médicos, Ortopédicos, Perfumaria e Cosméticos	6.827	2,05%
Gráfica e Editoração	6.699	2,01%
Veículos e Motocicletas, Peças e Partes	6.666	2,00%
Móveis e Eletrodomésticos	6.166	1,85%
Artigos Uso Pessoal e Domésticos	5.836	1,75%
Produção Primária/Agroindústria	5.680	1,70%
Metal Mecânica	5.636	1,69%
Saúde	5.255	1,58%
Equipamentos e Materiais de Escritório, Informática e Comunicação	4.720	1,42%
Indústria de alimentos e bebidas	4.635	1,39%
Madeira e Móveis	3.737	1,12%
T.I.	3.624	1,09%
Audiovisual	3.192	0,96%
Turismo	3.045	0,91%
Representação Comercial	2.458	0,74%
Atividades de Entrega	2.097	0,63%
Livros, Jornais, Revistas e Papelaria	1.990	0,60%
Esporte e Recreação	1.918	0,58%
Combustíveis e Lubrificantes	1.764	0,53%
Têxtil	1.420	0,43%
Mercado Financeiro	1.176	0,35%
Coleta e Tratamento de Resíduos	730	0,22%
Produtos Químicos	566	0,17%
Plástico	495	0,15%
Segurança	433	0,13%
Papel e Celulose	413	0,12%
Vidro	102	0,03%
Extração Mineral	20	0,01%

FONTE: A autora (2021).

4.2.1.1.12 Território

As empresas estão segmentas em vinte e cinto territórios, sendo que eles representam agrupamentos dos 399 municípios do Paraná. A TABELA 23 apresenta o quantitativo de empresas por território. Nota-se que o maior percentual de empresas está situado no território de Curitiba, além disso, há 3,24% de empresas sem território, estas são as que foram atendidas, mas estão lotadas fora do Paraná.

TABELA 23 - EMPRESA POR TERRITÓRIO

TERRITÓRIO	QTD. EMPRESAS	%
Curitiba	72.004	21,60%
Oeste Integrado	41.548	12,46%
Curitiba-RMC	30.491	9,14%
Norte do Paraná	24.617	7,38%
Terra Roxa	23.196	6,96%
Campos Gerais	15.194	4,56%
Sudoeste	13.387	4,01%
NA	10.811	3,24%
Vale do Ivaí	9.917	2,97%
Norte Pioneiro	9.688	2,91%
Litoral	8.528	2,56%
Paraná Centro	8.099	2,43%
Piquirivaí	7.754	2,33%
Arenito-Caiua	7.333	2,20%
Costa Noroeste	7.258	2,18%
Gralha Azul	6.859	2,06%
Paranapanema	6.632	1,99%
Cantuquiriguaçu	5.729	1,72%
Fronteira	4.693	1,41%
Vale do Tibagi	4.288	1,29%
Vale dos Pinheirais	4.108	1,23%
Iguaçu	3.693	1,11%
Da Moda	3.520	1,06%
Procopense	2.580	0,77%
Vale do Ribeira	1.498	0,45%

FONTE: A autora (2021).

4.2.1.2 Público-Alvo

Após a análise descritiva, limpeza dos dados e correções, a base de dados que inicialmente continha 339.061 empresas, resultou em 333.425, uma eliminação de 1,66% dos dados.

Porém, devido a capacidade computacional foi feita uma segmentação dos dados, trabalhando-se inicialmente somente com as empresas situadas no território de Curitiba (72.004), mas esta quantidade ainda foi excessiva para o processamento da análise de cluster, tendo de ser reduzida.

Assim, para superar esta barreira, foram selecionadas aleatoriamente 10 mil empresas do território de Curitiba.

4.2.1.3 Seleção das variáveis

Das quinze variáveis analisadas, foram selecionadas, primeiramente como pertinentes para a etapa de clusterização, somente sete: Tipo, Porte, Natureza Jurídica, Setor, *Startup*, Número de Colaboradores e Idade (proveniente do cálculo da data de abertura da empresa).

O QUADRO 6 apresenta as variáveis que não foram selecionadas e o motivo da exclusão.

QUADRO 6 - VARIÁVEIS NÃO SELECIONADAS

Nome	Motivo
CODIGO	Variável identificadora, não representa uma característica da empresa.
ATIVO	Somente fizeram parte da análise empresas ativas, assim como todas eram, não havia motivo de considerar a variável.
DATA_ANIVERSARIO	Ao invés de utilizar a variável Data de Aniversário, foi criada a variável Idade.
CNAE_DIVISAO	Não foi utilizada devido a possuir 86 categorias possíveis, o que aumentaria muito o número de variáveis *dummies*.
CNAE_CLASSE	Não foi utilizada devido a possuir 632 categorias possíveis, o que aumentaria muito o número de variáveis *dummies*.
CNAE_SEGMENTACAO	Não foi utilizada devido a quantidade de categorias, pois geraria 39 variáveis *dummies,* o que aumentaria muito a base de dados.
TERRITÓRIO	Não foi utilizada com fins de clusterização, porém foi utilizada para segmentar a base, de forma a reduzir o tamanho do conjunto de dados.
ESCOLARIDADE	Não foi utilizada pois a base para a clusterização baseou-se em dados de PJ.
TIPO_SEXO	Não foi utilizada pois a base para a clusterização baseou-se em dados de PJ.

FONTE: A autora (2021).

Embora selecionadas inicialmente para a etapa de clusterização, as variáveis Tipo, Natureza Jurídica e Startup não foram consideradas por serem praticamente constantes, ou seja, apresentam uma das categorias com mais de 99% de presença conforme pode-se observar no QUADRO 7, além disso, conforme visto nas análises descritivas, estas variáveis apresentam baixa entropia, o que demonstra uma alta concentração das observações. Desta maneira elas foram eliminadas do processo, permanecendo somente as variáveis Porte, Setor, Número de Colaboradores e Idade.

QUADRO 7 - PROPORÇÃO DAS VARIÁVEIS QUALITATIVAS SELECIONADAS

VARIÁVEL TIPO			
FÍSICA	**FORMAL**	**INFORMAL**	**PRODUTOR RURAL**
0,000	0,999	0,001	0,000
VARIÁVEL PORTE			
MEI	**ME**	**EPP**	**MGE**
0,516	0,389	0,054	0,041
VARIÁVEL NATUREZA JURÍDICA			
ADM. PÚBLICA	**ENT. EMPRESARIAIS**	**SEM FINS LUCRATIVOS**	**PESSOA FÍSICA**
0,001	0,992	0,006	0,000

VARIÁVEL SETOR				
AGRO	**COMÉRCIO**	**CONSTRUÇÃO**	**INDÚSTRIA**	**SERVIÇOS**
0,008	0,313	0,079	0,106	0,494

VARIÁVEL *STARTUP*	
SIM	**NÃO**
0,001	0,999

FONTE: A autora (2021).

4.2.2 Etapa 2: Preparação dos dados

Antes de inserir os dados no algoritmo é necessário preparar a base, esta fase é crucial, visto que a base é mista, ou seja, possui tanto dados quantitativos, quanto qualitativos. Para driblar este fato duas técnicas foram propostas:

a) Transformar as variáveis qualitativas em *dummies* e normalizar as variáveis quantitativas. Neste caso poderá ser usada uma medida de distância híbrida, que será utilizada pelo algoritmo de clusterização;

b) Transformar as variáveis qualitativas em *dummies* e dicotomizar as variáveis quantitativas, e assim utilizar uma medida de distância para dados binários.

Assim, na preparação dados serão necessárias três transformações nas variáveis para que se possa comparar as técnicas, são elas: transformar as qualitativas em *dummies*; transformar as quantitativas em *dummies;* e normalizar as quantitativas. Estas transformações serão combinadas para que as técnicas possam ser aplicadas e analisadas. Para a geração das variáveis *dummies* foi utilizada a função *createDummyFeatures* do pacote *mlr do software R* (BISCHL et al., 2016).

Na próxima sessão serão abordadas cada uma das transformações necessárias.

4.2.2.1 Transformando as variáveis qualitativas em *dummies*

Primeiramente as variáveis qualitativas (Porte e Setor) foram transformadas em *dummies*, ou seja, transformadas em binárias. Variáveis binárias são as que assumem valor de 0 ou 1, conforme a presença (valor 1) ou ausência (valor 0) da categoria. A quantidade de variáveis *dummies* é determinada pelo número de níveis, ou categorias, da variável original.

É importante destacar que para o processo de clusterização, diferentemente do que ocorre em regressão, todas as classes da variável são transformadas em *dummies*, pois neste processo é importante calcular a distância ou similaridade considerando todas as classes da variável. Este fato ~~que~~ não ocorre na regressão, em que as *dummies* são formadas pela quantidade de classes da variável menos um, caso contrário haverá a situação de colinearidade perfeita ou multicolinearidade perfeita (GUJARATI; PORTER, 2011).

Desta forma, por exemplo, a variável porte, que tem quatro categorias (MEI, ME, EPP e MGE), é transformada em quatro *dummies*, conforme demonstra a TABELA 24, em que é possível notar como cada variável categórica é representada de maneira dicotomizada.

A dicotomização das variáveis é necessária para que se possa analisar a similaridade entre as observações, item necessário na clusterização de variáveis qualitativas.

Com relação a quantidade de variáveis formadas, das duas variáveis qualitativas presentes na base (Porte e Setor), formou-se nove variáveis *dummies,* uma para cada classe da variável original.

TABELA 24 - EXEMPLO DE DICOTOMIZAÇÃO

CATEGÓRICA	MEI	ME	EPP	MGE
MEI	1	0	0	0
ME	0	1	0	0
EPP	0	0	1	0
MGE	0	0	0	1

FONTE: A autora (2021).

4.2.2.2 Transformando as variáveis quantitativas em *dummies*

Variáveis numéricas, antes de serem dicotomizadas, precisam primeiramente serem convertidas em categorias.

Este processo de transformação das variáveis numéricas em categóricas é chamado de discretização de dados. O objetivo é tornar uma base de dados mista em categórica, dado que muitos algoritmos são desenvolvidos para este tipo de base.

Na variável idade a conversão foi realizada considerando como intervalos as medidas dos quartis. Assim criou-se as seguintes faixas de idade: De zero a dois, de três a cinco, de seis a dez e mais que onze.

Desta maneira, cada uma das quatro categorias tornou-se uma nova variável *dummy*.

4.2.2.3 Normalizando as variáveis quantitativas

Na variável quantitativa foi aplicada a técnica de normalização. O objetivo da normalização é alterar os valores para uma escala comum, sem distorcer as diferenças. Nos dados a idade varia de 0 a 55 anos.

Desta forma a normalização reduz o intervalo dos dados, fixando entre 0 e 1. Isto, segundo Umamaheswari e Devi (2018), pode aumentar a precisão e desempenho de algoritmos de mineração envolvendo distâncias. Desta maneira, os atributos numéricos devem ser normalizados para serem considerados na mesma escala (AHMAD; DEY, 2007b). Para isso foi considerada a normalização Min-Max, conforme demonstrado na equação (14).

$$Z = \frac{y - \min(y)}{\max(y) - \min(y)} \quad (14)$$

4.2.3 Etapa 3: Criação da matriz de distâncias

Após a preparação dos dados, é realizada a aplicação dos algoritmos que vão calcular as distâncias entre as observações, o que vai permitir que o algoritmo de clusterização execute o agrupamento das observações mais similares.

Nesta fase foi proposta a criação de duas matrizes, a primeira utilizando a distância de Gower e a segunda a de Jaccard. Cada uma delas utilizará de uma combinação das transformações dos dados vistas anteriormente.

4.2.3.1 Distância de Gower

Para aplicação desta técnica, foi utilizada a transformação das variáveis qualitativas (Porte e Setor) em *dummies* e a normalização da variável numérica Idade. Das três variáveis iniciais, originaram-se dez, conforme exemplo mostrado pela TABELA 25, que apresenta as observações de cinco clientes.

TABELA 25 - EXEMPLO DE BASE PARA APLICAÇÃO DE GOWER

Idade	Porte				Setor				
	MEI	ME	EPP	MGE	Comércio	Construção	Indústria	Serviços	Agro
0,218182	0	1	0	0	0	0	1	0	0
0,072727	1	0	0	0	0	0	0	1	0
0,072727	1	0	0	0	0	0	0	1	0
0,072727	1	0	0	0	0	0	0	1	0
0,036364	1	0	0	0	0	0	1	0	0

FONTE: A autora (2021).

Após os dados processados, é aplicada a distância de Gower, que é uma medida híbrida, assim considera em seu cálculo tanto as variáveis quantitativas, quanto as binárias. No R foi utilizada a função *daisy(x, metric = 'gower')* do pacote *Cluster* (MAECHLER et al., 2018). A TABELA 26 mostra a matriz de distância de Gower para os cinco primeiros clientes. Nela pode-se observar, por exemplo, que o cliente um é mais similar ao cliente cinco (distância menor), que se comparado aos demais, em que a distância é maior.

TABELA 26 - EXEMPLO DE MATRIZ DE DISTÂNCIA DE GOWER

	1	2	3	4	5
1	0,00000000000	0,34953186962	0,34951340956	0,34953186962	0,13135005144
2	0,34953186962	0,00000000000	0,00001846006	0,00000000000	0,22626262626
3	0,34951340956	0,00001846006	0,00000000000	0,00001846006	0,22628108632
4	0,34953186962	0,00000000000	0,00001846006	0,00000000000	0,22626262626
5	0,13135005144	0,22626262626	0,22628108632	0,22626262626	0,00000000000

FONTE: A autora (2021).

4.2.3.2 Distância de Jaccard

Para aplicar a distância de Jaccard, é necessária uma transformação nas variáveis, assim as quantitativas e qualitativas foram transformadas em *dummies*. A transformação aplicada nas três variáveis, originou treze, conforme mostra a TABELA 27.

TABELA 27 - EXEMPLO DE BASE PARA APLICAÇÃO DE JACCARD

Porte				Setor					Idade			
MEI	ME	EPP	MGE	Comércio	Construção	Indústria	Serviços	Agro	De 0 a 2	De 3 a 5	De 6 a 10	Mais de 10
0	1	0	0	0	0	1	0	0	0	0	0	1
1	0	0	0	0	0	0	1	0	0	1	0	0
1	0	0	0	0	0	0	1	0	0	1	0	0
1	0	0	0	0	0	0	1	0	0	1	0	0
1	0	0	0	0	0	1	0	0	1	0	0	0

FONTE: A autora (2021).

Com a transformação dos dados é possível aplicar a distância de Jaccard (No R: função *dist(x, method='binary')* do pacote *Stats*). A TABELA 28 apresenta os dados dos cinco primeiros clientes, nela é possível notar, que o cliente dois é mais parecido com o cinco, pois a distância é menor.

TABELA 28 - EXEMPLO DE MATRIZ DE DISTÂNCIA DE JACCARD

	1	2	3	4	5
1	0,00	0,80	0,80	1,00	0,80
2	0,80	0,00	0,80	1,00	0,50
3	0,80	0,80	0,00	0,80	0,08
4	1,00	1,00	0,80	0,00	1,00
5	0,80	0,50	0,80	1,00	0,00

FONTE: A autora (2021).

4.2.4 Etapa 4: Aplicação do algoritmo de clusterização e análise da quantidade de agrupamentos

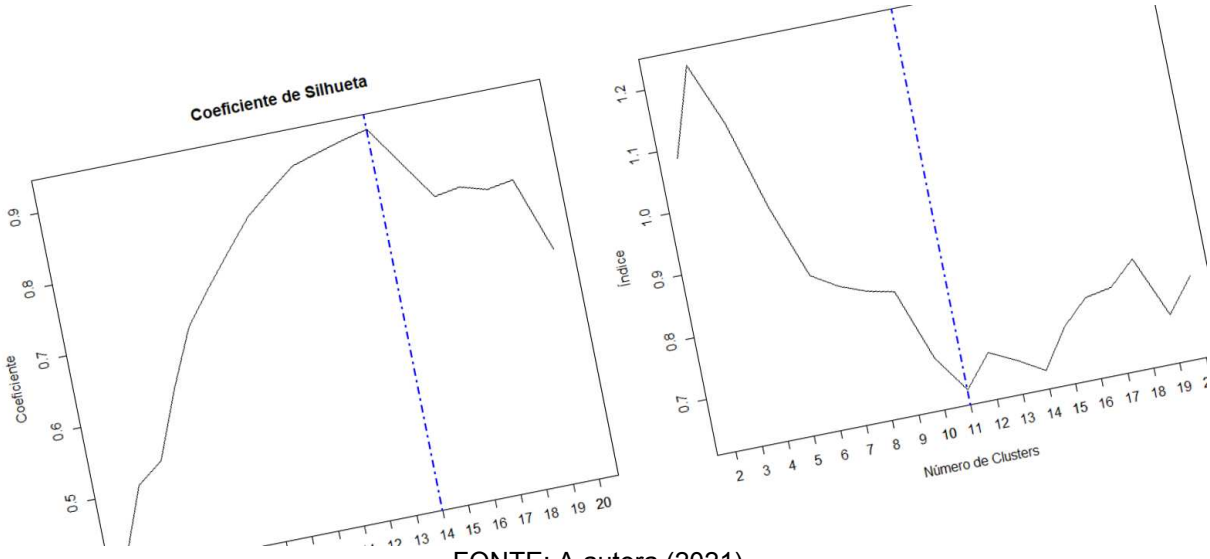
Para verificar o melhor número de clusters foram utilizados os índices de qualidade Davies Bouldin e o Coeficiente de Silhueta, executados no R pelas funções *silhouette()* do pacote *Cluster* e *index.DB()* do pacote *clusterSim*. Ambos foram aplicados nos dados provenientes das matrizes de distância geradas pelos algoritmos de Gower e Jaccard. Para a aplicação foram gerados de 2 a 20 clusters utilizando o algoritmo K-medoid, utilizando a função *pam()* do pacote *Cluster* (MAECHLER et al., 2018).

Destaca-se que a quantidade máxima de clusters igual a 20 foi escolhida pelo fato de que um número maior que este ficaria difícil para realizar o gerenciamento dos clientes na empresa que está aplicando a metodologia proposta.

A FIGURA 14 ilustra os índices de qualidade aplicados na matriz de distância proveniente do algoritmo de Gower.

Para o Coeficiente de Silhueta o valor máximo foi 0,924, o que significa que a estrutura forte ocorre quando o número de clusters é igual a 14. Já o Índice de Davies Bouldin, cuja interpretação é o quanto menor o índice, melhor a estrutura, confirma a quantidade de 11 clusters, como a melhor divisão. Com esta quantidade de agrupamentos o índice de Davies Bouldin foi de 0,633.

FIGURA 14 - ÍNDICES DE QUALIDADE PARA GOWER



FONTE: A autora (2021).

Como visto na FIGURA 14 não houve convergência dos índices de qualidade, enquanto o Coeficiente de Silhueta aponta 14 clusters, Davies-Bouldin sugere 11. Os índices detalhados podem ser observados na TABELA 29. É visto que no que compete ao índice de silhueta, ambas as quantidades de agrupamento possuem estrutura forte. Já para Davies Bouldin a diferença do índice entre os agrupamentos é de 0,0047, o que é relativamente baixo.

TABELA 29 –ÍNDICES DE QUALIDADE POR QUANTIDADE DE CLUSTERS PARA GOWER

Cluster	Silhueta	Davies - Bouldin
2	0,4120782	1,0843627
3	0,5066305	1,2257387
4	0,5329370	1,1233637
5	0,6268189	0,9783254
6	0,7057741	0,8601195
7	0,7538687	0,8349403
8	0,7973850	0,8188431
9	0,8395032	0,8087792
10	0,8686206	0,6929739
11	0,8951100	0,6335578
12	0,9061119	0,6852868
13	0,9164001	0,6649722
14	0,9245436	0,6383049
15	0,8701494	0,7003872
16	0,8155121	0,7405344
17	0,8216758	0,7477657
18	0,8112560	0,7857310
19	0,8168630	0,6869822
20	0,7128005	0,7415119

FONTE: A autora (2021).

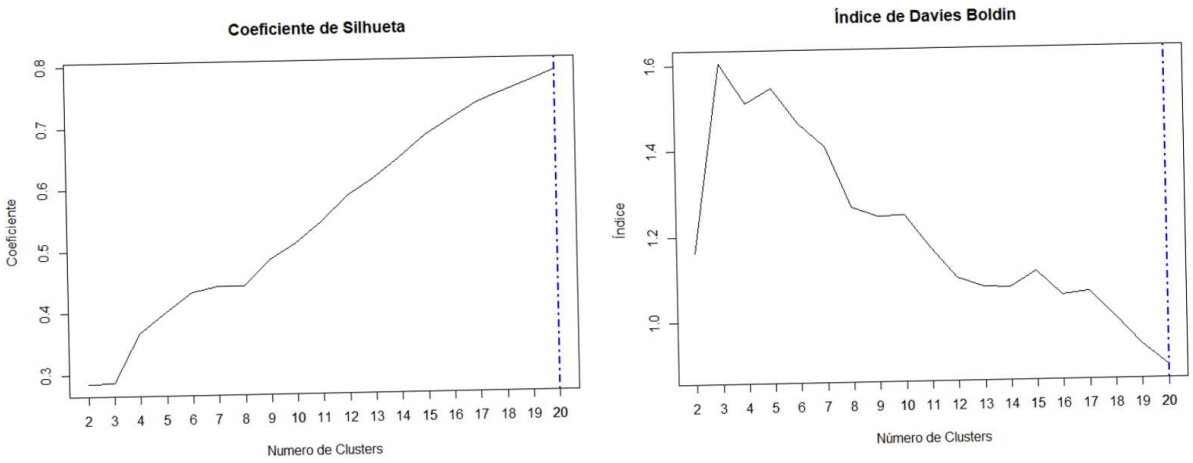
Aplicando os índices de qualidade na matriz de distância proveniente do algoritmo de Jaccard, conforme ilustra a FIGURA 15 e demonstra a TABELA 30, tem-se que para o Coeficiente de Silhueta o índice máximo apresentado foi de 0,785, indicando uma estrutura forte para 20 clusters.

Davies Bouldin também apontou 20 clusters como o melhor agrupamento, para o qual apresentou o menor índice (0,884).

Destaca-se que ambos os índices atingiram seu melhor resultado no número máximo de clusters fixado. Intuitivamente, pela FIGURA 15, espera-se que ambos os índices continuem melhorando à medida que se aumente o número de clusters,

entretanto 20 já é considerado um número expressivo de agrupamentos, e aumentar esse número é inviável para análise de perfil do cliente, que será apresentada adiante.

FIGURA 15 - ÍNDICES DE QUALIDADE PARA JACCARD



FONTE: A autora (2021).

TABELA 30 –ÍNDICES DE QUALIDADE POR QUANTIDADE DE CLUSTERS PARA JACCARD

Cluster	Silhueta	Davies - Bouldin
2	0,285296	1,161948
3	0,287023	1,604729
4	0,366152	1,510014
5	0,400905	1,546718
6	0,432479	1,463988
7	0,441362	1,407385
8	0,442083	1,263173
9	0,484117	1,242749
10	0,508863	1,245278
11	0,543431	1,165329
12	0,585274	1,095987
13	0,612221	1,07386
14	0,6456	1,07278
15	0,681474	1,110093
16	0,707823	1,051861
17	0,733236	1,059495
18	0,751609	0,999377
19	0,768151	0,934523
20	0,785449	0,884346

FONTE: A autora (2021).

Conforme demonstrado nas análises anteriores, a matriz de distância de Gower apresentou melhores estruturas para os clusters, pois ambos os índices de qualidade, tanto para 11 como 14 clusters exibiram melhores resultados.

Diante do exposto, foi considerado os agrupamentos provenientes da matriz de distância de Gower. Pois, além de possuir melhores estruturas, também possuem menor quantidade de clusters, o que auxilia no processo de gerenciamento dos grupos pelas áreas de relacionamento com o cliente.

Devido a pouca diferença entre os índices de qualidade provenientes de Gower, qualquer um dos agrupamentos poderá ser escolhido, pois ambos têm estruturas fortes. Desta forma, devido ser mais facilmente gerenciável menos clusters decidiu-se pela menor quantidade de agrupamentos, ou seja, 11 clusters.

Acredita-se que melhores estruturas de agrupamento podem ser obtidas com a distância de Jaccard, se for aplicada outras metodologias para binarizar as variáveis quantitativas. Porém, como mencionado por Harikumar e Surya (2015), para transformar um conjunto de dados mistos em homogêneos, pode-se haver a perda de informação.

4.2.5 Etapa 5: Análise dos clusters gerados

A análise descritiva é fundamental ter clareza das características embutidas em cada cluster, pois assim consegue-se fazer a melhor utilização, permitindo rotulá-los com as características mais evidentes do agrupamento, o que auxilia no momento de direcionar campanhas de *marketing*. A análise descritiva foi executada no *software R* utilizando o pacote *tidyverse* (WICKHAM et al., 2019).

Com relação a quantidade de clientes, TABELA 31, nota-se que a quantidade de clientes não é homogênea, ou seja, há grupos que concentram 27% dos clientes (cluster 2) e alguns que apenas 2% (cluster 11).

TABELA 31 - TAMANHO DOS CLUSTERS

Cluster	Quantidade	%
1	328	3%
2	2.688	27%
3	684	7%
4	553	6%
5	1.751	18%
6	1.682	17%
7	1.274	13%
8	252	3%
9	273	3%
10	182	2%
11	333	3%
Total	10.000	100%

FONTE: A autora (2021).

Já com relação a variável porte, (TABELA 32), é possível notar que houve uma segregação bem expressiva dos clusters, pois em quatro clusters (36%) houve total representatividade de porte, ou seja, 100% dos clientes dentro destes agrupamentos são de um único porte.

Nota-se também que o porte MEI e o ME predominaram em quatro clusters sendo respectivamente no 2, 3, 4 e 7 o MEI e 1, 5, 6 e 10 o ME. O EPP teve predominância em dois clusters (9 e 10), já o MGE somente no cluster 11.

TABELA 32 - CLUSTERS POR PORTE

Cluster	MEI	ME	EPP	MGE	Porte predominante
1	0	313	11	4	ME
2	2.688	0	0	0	MEI
3	678	0	5	1	MEI
4	539	0	5	9	MEI
5	0	1.751	0	0	ME
6	0	1.665	0	17	ME
7	1.264	0	0	10	MEI
8	0	0	243	9	EPP
9	0	0	273	0	EPP
10	0	156	9	17	ME
11	0	0	0	333	MGE

FONTE: A autora (2021).

Considerando os setores, pode-se observar na TABELA 33, que os clusters estão bem segmentados. O setor Agropecuário não teve predominância em nenhum cluster. Comércio predominou nos clusters 6 (ME), 7 (MEI) e 8 (EPP), sendo nos dois primeiros com exclusividade. Construção predominou os clusters 4(MEI) e 10 (ME). Indústria esteve presente nos clusters 1 (ME) e 3 (MEI). Serviços foi o setor que

predominou em quatro dos onze clusters, tendo representatividade em todos os portes, 2 (MEI), 5 (ME) ,9 (EPP) e 11 (MGE).

TABELA 33 - CLUSTERS POR SETOR

Cluster	Agropecuário	Comércio	Construção	Indústria	Serviços	Porte predominante
1	9	0	0	319	0	ME
2	25	0	0	0	2.663	MEI
3	0	0	0	684	0	MEI
4	30	0	523	0	0	MEI
5	7	0	0	0	1.744	ME
6	0	1.682	0	0	0	ME
7	0	1.274	0	0	0	MEI
8	0	208	14	30	0	EPP
9	0	0	4	5	264	EPP
10	0	0	182	0	0	ME
11	4	32	10	17	270	MGE

FONTE: A autora (2021).

Com relação a idade das empresas por cluster (TABELA 34), nota-se que os clusters nota-se que a idade mediana dos clusters segue uma relação de Porte e Setor. Assim quanto maior o porte, maior a idade das empresas. Notou-se também que para empresas que dentro do Porte, o setor de serviços tende a ter uma idade mais baixa.

TABELA 34 - IDADE POR CLUSTERS

Cluster	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo	Porte predominante
1	1	8	12	14,56	19	54	ME
2	0	2	3	3.914	5	27	MEI
3	0	2	4	4.278	6	23	MEI
4	0	2	4	4.526	6	12	MEI
5	1	5	10	12,09	16	54	ME
6	1	7	12	14,15	20	55	ME
7	0	3	4	4.719	6	34	MEI
8	1	10	18	19,78	26	55	EPP
9	1	7	13	14,99	21	54	EPP
10	2	8	11	13,96	18	48	ME
11	1	10	20	20,7	28	54	MGE

FONTE: A autora (2021).

Resumidamente, pode-se descrever os grupos da seguinte maneira:

— Grupo 1: Empresas do porte ME, do setor da Indústria, com 12 anos medianos de idade;

— Grupo 2: Empresas do porte MEI, alta proporção de Serviço, com 3 anos medianos de idade;

- Grupo 3: Alta proporção de empresas do porte MEI, do setor da Indústria, com 4 anos medianos de idade;
- Grupo 4: Empresas do porte MEI, grande concentração do setor de Construção Civil, com 4 anos medianos de idade;
- Grupo 5: Empresas do porte ME, grande concentração do setor de Serviços, com 10 anos medianos de idade;
- Grupo 6: Empresas do porte ME, do setor de Comércio, com 12 anos medianos de idade;
- Grupo 7: Empresas do porte MEI, do setor de Comércio, com 4 anos medianos de idade;
- Grupo 8: Empresas do porte EPP, do setor do Comércio, com 18 anos medianos de idade;
- Grupo 9: Empresas do porte EPP, do setor de Serviços, com 13 anos medianos de idade;
- Grupo 10: Empresas do porte ME, grande concentração do setor Construção Civil, com 11 anos medianos de idade;
- Grupo 11: Empresas do porte MGE, grande concentração do setor Serviços, com 20 anos medianos de idade.

Com o resumo fica mais fácil entender o perfil das empresas de cada um dos grupos e assim traçar estratégias mais eficientes de *marketing* e relacionamento com os clientes, de modo a conseguir maior engajamento nas ações.

4.3 SEGUNDA FASE: RECOMENDAÇÃO DE PRODUTOS

Com os grupos de clientes definidos, além de uma comunicação mais eficaz, pode-se também recomendar produtos com base no perfil de consumo. Nesta fase aplicam-se os algoritmos de regras de associação, com a finalidade de recomendar o produto mais indicado para o cliente, evitando que ele procure em uma grande variedade de produtos.

4.3.1 Etapa 1: Análise da base de interações

Antes de aplicar o algoritmo é importante analisar as variáveis, com o intuito de compreender melhor a base de dados, para fazer o processo de limpeza e ajustes nos dados para implementação dos modelos.

4.3.1.1 Interações

Interações são os atendimentos recebidos pelos clientes, assim pode haver mais atendimentos que produtos consumidos, pois um cliente pode consumir um mesmo produto mais de uma vez.

A análise de dados levou em conta as interações dos dez mil clientes que foram clusterizados. O período utilizado foi de 24/04/2017 a 04/02/2020 (cerca de 33 meses). Neste período houve 64.381 interações.

Com relação as interações, como mostra a TABELA 35, há clientes que receberam somente um atendimento, porém um cliente chegou a receber 161 atendimentos neste período. A média de atendimentos por cliente foi de 6,44 e a mediana foi de 3 atendimentos.

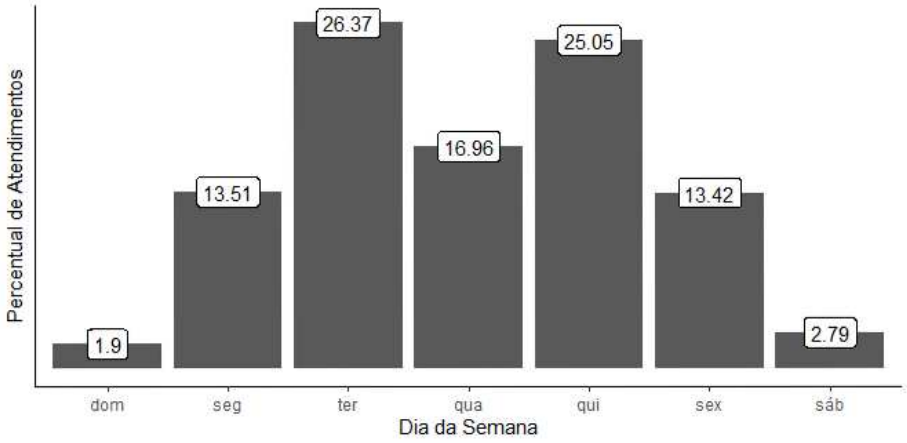
TABELA 35 - ATENDIMENTOS POR CLIENTE

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
1	2	3	6,44	8	161

FONTE: A autora (2021).

A FIGURA 16 ilustra o percentual de atendimentos por dia da semana, nela é possível notar que mais de 50% dos atendimentos ocorrem terça e quinta. Nos finais de semana o atendimento é menor, devido a ocorrência de eventos pontuais.

FIGURA 16 - ATENDIMENTOS POR DIA DA SEMANA



FONTE: A autora (2021).

4.3.1.2 Produtos

Embora no período ocorreram 64.381 atendimentos, houve a disponibilização de somente 736 produtos.

Na TABELA 36 é possível notar que há clientes que consumiram somente um tipo de produto, sendo o maior consumo foi de 59 produtos diferentes. Em média um cliente chegou a utilizar até 4 produtos. Nota-se também que 75% dos clientes (3º quartil) utilizaram até 5 produtos diferentes, o que demonstra uma baixa diversidade de produtos por cliente.

TABELA 36 - PRODUTOS CONSUMIDOS POR CLIENTE

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
1	2	3	4	5	59

FONTE: A autora (2021).

Na TABELA 37 nota-se que os primeiros seis produtos são responsáveis por 53% dos atendimentos do período, sendo que o produto de código 9198 responde sozinho por quase 24% dos atendimentos.

TABELA 37 - QUANTIDADE DE ATENDIMENTOS GERADOS PELOS VINTE PRODUTOS MAIS CONSUMIDOS

CÓD PRODUTO	QTD. ATENDIMENTOS GERADOS	%	% ACUMULADO
9198	15.012	23,74%	23,74%
13177	3.855	6,10%	29,84%
8248	3.852	6,09%	35,93%
7498	3.779	5,98%	41,90%
13179	3.746	5,92%	47,83%
675	3.450	5,46%	53,28%
13190	2.736	4,33%	57,61%
6733	2.717	4,30%	61,91%
10066	2.646	4,18%	66,09%
13187	1.462	2,31%	68,40%
10064	1.391	2,20%	70,60%
10062	1.162	1,84%	72,44%
10063	1.056	1,67%	74,11%
488	1.020	1,61%	75,72%
13186	1.020	1,61%	77,34%
9197	953	1,51%	78,84%
11558	936	1,48%	80,32%
13178	925	1,46%	81,79%
10070	905	1,43%	83,22%
10065	848	1,34%	84,56%

FONTE: A autora (2021).

Na TABELA 38 observa-se que há produtos que geraram somente um atendimento, em contraponto houve 15.012 atendimentos gerados por um único produto. Também se nota que 75% dos produtos geraram até 2 atendimentos, isto mostra uma redução na chance de usuários possuírem itens comuns. O que evidencia o que já foi visto na TABELA 37, que poucos produtos são responsáveis pela grande maioria dos atendimentos.

TABELA 38 - ATENDIMENTOS GERADOS POR PRODUTO

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
1	1	1	86	2	15.012

FONTE: A autora (2021).

Na TABELA 39 é possível analisar os vinte produtos que mais geraram atendimentos em 2020. Nota-se grande parte dos produtos de 2020 não existiam ou não foram disponibilizados ou consumidos pelos clientes em 2017.

TABELA 39 - ATENDIMENTOS GERADOS POR PRODUTO POR ANO

CÓD PRODUTO*	ANOS 2017	2018	2019	2020
675	0	958	1.694	84
488	0	0	973	83
6733	0	963	1.602	81

6731	0	209	378	29
3089	0	381	440	27
9197	0	296	698	26
7052	0	62	140	14
12932	0	0	78	10
286	1	297	72	10
13069	0	0	146	9
9198	0	0	15.004	8
6732	0	49	132	4
715	0	209	8	4
2331	0	71	199	3
6734	0	82	160	3
13372	0	0	0	2
11558	0	930	21	2
16302	0	0	0	1
16324	0	0	0	1
10809	0	0	103	1

FONTE: A autora (2021).
*Considerado somente os 20 produtos que mais geraram atendimento em 2020.

A quantidade de produtos distintos (TABELA 40), consumidos pelos clientes, alterou-se ao longo do tempo, nota-se que em 2018 e 2019 há uma maior quantidade de produtos e atendimentos pois são anos completos. Em relação a quantidade de clientes, a soma dos anos é maior que dez mil, pois um mesmo cliente pode ser atendido em diferentes anos.

TABELA 40 - PROTUDOS, ATENDIMENTOS E CLIENTES POR ANO

TIPO	ANOS			
	2017	2018	2019	2020
PRODUTOS	4	307	474	23
ATENDIMENTOS	12	23.275	39.541	405
CLIENTES	12	5.526	7.291	263

FONTE: A autora (2021).

É importante destacar que a base utilizada no estudo faz referência a clientes que estavam presentes na base de *score* e tinham alguma pontuação, (conforme definido na metodologia, na etapa de seleção de público-alvo). Desta maneira o tempo de relacionamento e consumo são fatores que influenciam na pontuação (*score*), o que faz com que o estudo privilegie clientes que se relacionaram mais recentemente.

A TABELA 41 mostra que dos clientes atendidos em 2018 e 2019, somente 2.930 consumiram em ambos os anos, evidenciando uma alta renovação de clientes entre os anos. Já com relação aos produtos, do total, somente 49 geraram atendimento em 2018 e 2019, ou seja, dos 307 produtos utilizados em 2018, somente 49 também foram utilizadas em 2019, o que evidência uma possível renovação de produtos, ou alteração de estratégias de atendimento em diferentes anos.

TABELA 41 - PRODUTOS E CLIENTES EM 2018 E 2019

TIPO	2018 a 2019
PRODUTOS	49
CLIENTES	2.930

FONTE: A autora (2021).

4.3.1.3 Tipo de solução

O tipo de solução define o formato do produto, e no QUADRO 8 são apresentadas as definições de cada classificação disponível da variável.

QUADRO 8 - DESCRIÇÃO DOS TIPOS DE SOLUÇÕES

Tipo de Solução	Definição das Classificações
Consultoria	Serviço de diagnóstico de uma situação particular, sobre a qual pode ser elaborado um plano de ação com soluções específicas e adequadas, bem como o acompanhamento de sua implementação. - Duração mínima de 1 hora.
Seminário	Exposição oral voltada para a disseminação de temas e processos de sensibilização, realizada por um ou mais especialistas, destinada a um grupo de pessoas com interesses comuns. Duração: Possui carga horária de 4 a 12 horas.
Orientação	Orientação sobre questões técnicas a partir de interesse do cliente, que podem ser respondidas com conteúdo disponível, sem necessariamente ocorrer um processo de diagnóstico. Contempla envio/entrega de conteúdo técnico (DVD, CD, cartilha, documento digital etc.).
Oficina	É um trabalho em grupo, em que se trabalham temas de interesse comum por meio de estratégias de exposição oral, dinâmicas de grupo, simulações, experimentações etc. Duração: inferior a 12 horas.
Cursos	Capacitação em que se busca, desenvolver e aprimorar conhecimentos, atitudes e habilidades de gestão. Duração mínima de 12 horas.
Palestra	É uma exposição oral de curta duração voltada para a disseminação de um tema e a processos de sensibilização, realizada por um especialista, destinada a um grupo de pessoas com interesses comuns. Carga horária inferior a 4 horas.
Missão/Caravana	São grupos de pessoas, cuja organização e deslocamento são organizados pela instituição, com a finalidade de viabilizar a participação dos clientes em eventos (feiras, exposições, encontros etc.).
Feira	Evento que reúne expositores (empresas) de diversos segmentos, possibilitando a exposição, demonstração e comercialização de seus produtos e serviços.
Rodada de Negócio	São encontros promovidos entre empresas compradoras e vendedoras, tendo como objetivo a criação de parcerias de negócios.
Não contabilizar	Produto não contabilizado.
Informação	Relacionado à disponibilização de informações gerais, de interesse empresarial, podendo ser demandadas pelo cliente.

FONTE: A autora (2021).

Na TABELA 42 nota-se que, embora existam 736 produtos, há alguns que possuem mais de um tipo de solução vinculado, o que faz o total de produtos passar para 780. Também é possível concluir que cerca 40% dos produtos correspondem a Oficinas e Palestras.

TABELA 42 - QUANTIDADE DE PRODUTOS POR TIPO DE SOLUÇÃO

TIPO DE SOLUÇÃO	QTD. PRODUTOS
Oficina	161
Palestra	160
Seminário	118
Consultoria	103
Orientação	60
Curso	42
Informação	38
Rodada de Negócios	32
Sem Categorização	23
Feira	20
Missão/Caravana	19
Não Contabilizar	4
TOTAL	780

FONTE: A autora (2021).

Analisando o tipo de solução por ano (TABELA 43), nota-se que algumas soluções tiveram grande aumento no número de produtos consumidos, considerando 2018 para 2019. Missão, Seminário, Oficina e Feira foram os tipos de soluções que tiveram uma demanda de mais de 85% de novos produtos.

TABELA 43 - TIPO DE SOLUÇÃO POR PRODUTO E ANO

TIPO DE SOLUÇÃO	ANO			
	2017	2018	2019	2020
Consultoria	1	59	58	5
Curso	0	22	23	0
Feira	0	7	13	0
Informação	0	37	5	3
Missão/Caravana	0	5	14	0
Não Contabilizar	0	0	4	0
Oficina	0	57	106	0
Orientação	2	39	44	16
Palestra	1	67	94	2
Rodada de Negócios	0	11	22	0
Sem Categorização	0	1	22	0
Seminário	0	38	80	0

FONTE: A autora (2021).

4.3.1.4 Temas

Os produtos são classificados por temas, que estão relacionados ao conteúdo do produto. Em uma primeira análise (TABELA 44), constatou-se 179 temas, porém muitos tinham acrescido ao nome ano, por exemplo, ‘Planejamento estratégico 2019’.

Assim, para fazer uma análise geral da quantidade de temas, e a quantidade de produtos associados, primeiramente realizou-se uma limpeza nos dados removendo todas as denominações de anos presentes. Após a limpeza restou na

base 159 temas. Além disso, dos 736 produtos, há 422 (57%) que não possuem nenhum tema cadastrado.

TABELA 44 - VINTE TEMAS COM MAIS PRODUTOS

TEMA	QTD. PRODUTOS
NA	422
Planejamento estratégico	28
Atendimento ao cliente	26
Diagnóstico empresarial	26
Liderança	19
Comportamento empreendedor	17
Estratégia de *marketing*	17
Gestão econômico/financeira	14
Turismo	14
Empreendedorismo	13
Gestão da qualidade e produtividade	12
Gestão econômico/ financeira	12
Marketing de relacionamento	12
Marketing digital	12
Marketing estratégico	12
Gestão de vendas	11
Mídias digitais	11
Inovação	10
Modelos de negócios	10
Gestão da inovação	9

FONTE: A autora (2021).

Outro fato relativo ao tema, é que um produto pode ser vinculado a mais de um tema, desta maneira, há produtos com até 17 temas cadastrados, embora 75% dos produtos tenham apenas um tema (TABELA 45).

TABELA 45 - TEMAS POR PRODUTO

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
1	1	1	1,462	1	17

FONTE: A autora (2021).

Já com relação ao ano, a TABELA 46 apresenta os vinte temas com mais produtos vinculados, para esta análise utilizou-se como referência 2019 por ser o ano completo mais recente da base. Nota-se que em 2019 foram incluídos novos temas, como por exemplo *Marketing Digital,* além disso percebe-se que a falta de cadastro do produto no tema aumentou de 2017 para 2019.

TABELA 46 - TEMAS POR ANO

TEMA*	2017	2018	2019	2020
NA	3	159	277	13
Atendimento ao cliente	0	7	20	0
Planejamento estratégico	0	17	16	0
Diagnóstico empresarial	1	18	15	4
Liderança	0	9	15	0

Comportamento empreendedor	0	6	14	1
Estratégia de *marketing*	0	8	12	0
Marketing digital	0	0	12	0
Mídias digitais	0	0	11	0
Gestão econômico/financeira	0	7	10	2
Marketing de relacionamento	0	3	10	0
Modelos de negócios	0	1	9	0
Gestão da inovação	0	1	8	0
Gestão de vendas	0	4	8	0
Redes sociais	0	1	8	0
Startup	0	1	8	0
Turismo	0	7	8	0
Modelos de excelência em gestão	0	2	7	0
Tendências	0	1	7	0
Análise da viabilidade econômica e financeira	0	2	6	1

FONTE: A autora (2021).
*Considerado somente os 20 temas com mais produtos vinculados em 2019.

Nota-se na TABELA 47 que os produtos sem tema foram os que mais geraram atendimentos.

TABELA 47 - TEMAS POR QUANTIDADE DE ATENDIMENOS

TEMA	QTD. ATENDIMENTO
NA	36.398
Atendimento ao cliente	15.689
Diagnóstico empresarial	4.143
Inovação	3.914
Gestão econômico/financeira	2.916
Planejamento estratégico	1.457
Empreendedorismo	1.113
Diagnóstico empresarial	1.046
Estratégia de *marketing*	833
Modelos de excelência em gestão	725
Marketing digital	683
Modelos de negócios	683
Mídias digitais	680
Redes sociais	674
Gestão de processos empresariais	670
Transformação digital	668
Análise mercadológica	667
Gestão de vendas	417
Linhas de crédito	273
Comportamento empreendedor	271

FONTE: A autora (2021).

4.3.1.5 Área do Conhecimento

Outra variável pertencente a base de interações e relativa a produto é a área de conhecimento. Na base havia 36 áreas de conhecimento, porém como várias possuíam em sua nomenclatura a data de criação, foi necessário realizar uma limpeza na base. Assim de 36, restaram 24 áreas de conhecimento únicas.

Porém da mesma forma que tema, existe na base 422 produtos sem classificação de área de conhecimento. Já com relação a quantidade de produtos, as áreas de conhecimento com mais produtos são *Marketing* e Vendas com 92 produtos e Serviços Financeiros com 47.

TABELA 48 - QUANTIDADE DE PRODUTOS POR ÁREA DO CONHECIMENTO

ÁREA DO CONHECIMENTO	QTD. PRODUTOS
NA	422
Marketing e vendas	92
Serviços financeiros e contábeis	47
Estratégias empresariais	37
Inovação	36
Planejamento empresarial	31
Recursos humanos e empreendedorismo	28
Desenvolvimento setorial	24
Recursos humanos	18
Empreendedorismo	17
Gestão da produção e qualidade	17
Políticas públicas	12
Associativismo e cooperativismo	9
Startup	9
Legislação	8
Educação	7
Desenvolvimento territorial	6
Design	6
Negócios internacionais	5
Comércio exterior	4
Agronegócios	3
Franquias	3
Tecnologia da informação	3
Legislação aplicada aos pequenos negócios	2
Sustentabilidade	1

FONTE: A autora (2021).

Um produto também pode ter mais de uma área do conhecimento (TABELA 49), podendo chegar a no máximo 11, embora 75% dos produtos tenham somente uma área do conhecimento vinculada.

TABELA 49 – ÁREA DO CONHECIMENTO POR PRODUTO

Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
1	1	1	1,51	1	11

FONTE: A autora (2021).

4.3.1.6 Cluster

Analisando-se a média de atendimentos por cluster (TABELA 50), nota-se que o cluster 11 é o que possui menor média de atendimento, ele é o que possui perfil de

Médias e Grandes Empresas. Outro fato é que há menos produtos únicos consumidos nos clusters 4 e 10, ambos de clientes da Construção civil.

TABELA 50 - ATENDIMENTOS, PRODUTOS E CLIENTES POR CLUSTER

CLUSTER	QTD. ATENDIMENTOS	QTD. PRODUTOS	QTD. CLIENTES	MÉDIA DE ATENDIMENTOS
1	2.139	115	328	6,52
2	16.635	220	2.688	6,19
3	4.661	134	684	6,81
4	3.243	65	553	5,86
5	12.107	366	1.751	6,91
6	11.121	311	1.682	6,61
7	8.260	136	1.274	6,48
8	1.923	152	252	7,63
9	1.797	127	273	6,58
10	1.022	62	182	5,62
11	1.473	252	333	4,42

FONTE: A autora (2021).

Devido haver uma não continuidade na disponibilização de produtos entre 2018 e 2019, em que somente 49 produtos estão em ambos os anos, resolveu-se trabalhar com a recomendação de produtos somente para os clientes que tiveram alguma interação no período de 2019 e 2020 (TABELA 51). Desta forma a quantidade de clientes que terão produtos recomendados reduz de 10 mil para 7.418.

TABELA 51 - ATENDIMENTOS, PRODUTOS E CLIENTES POR CLUSTER A PARTIR DE 2019

CLUSTER	QTD. ATENDIMENTOS	QTD. PRODUTOS	QTD. CLIENTES	MÉDIA DE ATENDIMENTOS
1	1.271	76	227	5,60
2	11.158	146	2153	5,18
3	3.165	93	544	5,82
4	2.174	44	439	4,95
5	7.205	226	1.171	6,15
6	6.902	193	1.205	5,73
7	5.494	88	1.019	5,39
8	1.169	91	186	6,28
9	883	66	176	5,02
10	697	39	124	5,62
11	511	160	174	2,94

FONTE: A autora (2021).

4.3.2 Etapa 2: Preparação da base

Para aplicar os algoritmos utilizou-se as funções do pacote *recommenderlab* (HAHSLER, 2020) do *software R*. Primeiramente converteu-se a base para binária, sendo 1 para quando houve o consumo do produto pelo cliente e 0 caso contrário. A TABELA 52 exemplifica a conversão para base binária.

A recomendação foi gerada por cluster, dado que há um perfil de clientes diferente em cada grupo, assim, direcionar produtos por cluster facilita a comunicação e torna as campanhas de *marketing* mais atraentes.

TABELA 52 - EXEMPLO DE TABELA BINÁRIA DE RECOMENDAÇÃO

		PRODUTOS						
		10006	10057	10809	10855	10870	10968	11100
CLIENTES	10042276	0	0	0	0	0	0	1
	1011012	0	0	0	0	0	0	0
	10150773	0	0	0	0	0	0	0
	10259462	0	0	0	0	0	0	0
	10271358	0	0	1	0	0	0	0
	10274770	0	0	0	0	0	0	0
	10507265	0	0	0	0	0	0	0
	10523872	0	0	0	0	0	0	0
	10539907	0	0	0	0	0	0	0
	10541362	1	0	0	0	0	0	0

FONTE: A autora (2021).

4.3.3 Etapa 3: Definição da forma de avaliação

A avaliação entre os modelos foi realizada de forma *off-line*, que é a forma mais ágil e barata. Este modo simula o processo *on-line*, em que o sistema faz previsões e o usuário interage, utilizando ou não o que foi proposto. Desta forma a avaliação é realizada sobre a base histórica de transações e não sobre os resultados das interações dos clientes. Esta uma forma de avaliar os modelos antes de colocá-los em produção.

Para realizar a avaliação, os dados são separados em treino e teste e em seguida é amostrado n itens que serão ocultos no conjunto de teste, então o algoritmo utiliza os dados de treinamento para prever os dados faltantes.

Nesta divisão foi utilizada a validação cruzada para avaliar a capacidade de generalização de um modelo. Ela faz melhor uso da base de dados, pois ao invés de dividir o conjunto em treinamento e teste, calcula as estimativas sobre todos os dados disponíveis, por meio da realização de várias divisões e alterações sistemáticas das amostras para testes.

A validação cruzada divide a base em k partes, chamadas dobras. Assim repete o treinamento e testes k vezes, e em cada repetição uma dobra diferente é escolhida como teste, e as outras $(k-1)$ são utilizadas para treinamento (PROVOST; FAWCETT, 2016).

Na simulação foram utilizados os seguintes parâmetros:

a) Método para determinar a divisão dos dados em treino e teste: Validação Cruzada;

b) Dobras (k): O número de dobras para a validação cruzada utilizado foi 5;

c) Conjunto de treinamento e teste: 80% treinamento; 20% teste;

d) Dados ocultos (n): 1 (Utilizado para prever o resultado).

É importante observar, que como um produto é oculto, para que se possa prever o resultado, somente foram incluídos na avaliação os clientes que adquiriram mais de um código de produto, assim restaram na base 5.603 clientes (TABELA 53).

TABELA 53 - CLIENTES QUE CONSUMIRAM MAIS DE UM PRODUTO

CLUSTER	QTD. CLIENTES
1	161
2	1.741
3	451
4	396
5	877
6	770
7	813
8	115
9	113
10	90
11	76
TOTAL	5.603

FONTE: A autora (2021).

4.3.4 Etapa 4: Algoritmos e parâmetros testados.

A avaliação testou os seguintes modelos e parâmetros:

a) Apriori:
   - Suporte: 0,2;
   - Confiança 0,75.

b) Filtragem Colaborativa Baseada no Item:
   - Itens semelhantes: 30;
   - Função de similaridade: Jaccard.

c) Filtragem Colaborativa Baseada no Usuário:
   - Usuários semelhantes: 30;
   - Função de similaridade: Jaccard.

Para o modelo Apriori, a definição dos parâmetros baseou-se nos índices de suporte e confiança que foram vistos nos artigos da revisão de literatura. Já na

Filtragem Colaborativa foi utilizado trinta para itens e usuários semelhantes, como parâmetro inicial da análise. Com relação a função de similaridade, foi escolhida a de Jaccard por ser adequada a bases binárias.

Outro ponto definido foi a quantidade de itens a serem recomendados, que neste caso foram três, que são os itens mais similares encontrados.

4.3.5 Etapa 5: Análise dos resultados

A análise dos resultados é realizada com o auxílio de indicadores de desempenho Taxa de Falsos Positivos e Taxa de Verdadeiros Positivos, presentes na Curva ROC, além de Precisão e Recall.

Na FIGURA 17 é possível comparar os modelos aplicados no cluster 1, nela pode-se notar na CURVA ROC, que em relação a TPR, o modelo Filtragem Colaborativa Baseada no Item apresenta melhores resultados se comparado com o Apriori e a Filtragem Colaborativa baseada no usuário, ficando estes, respectivamente, em segundo e terceiro lugar.

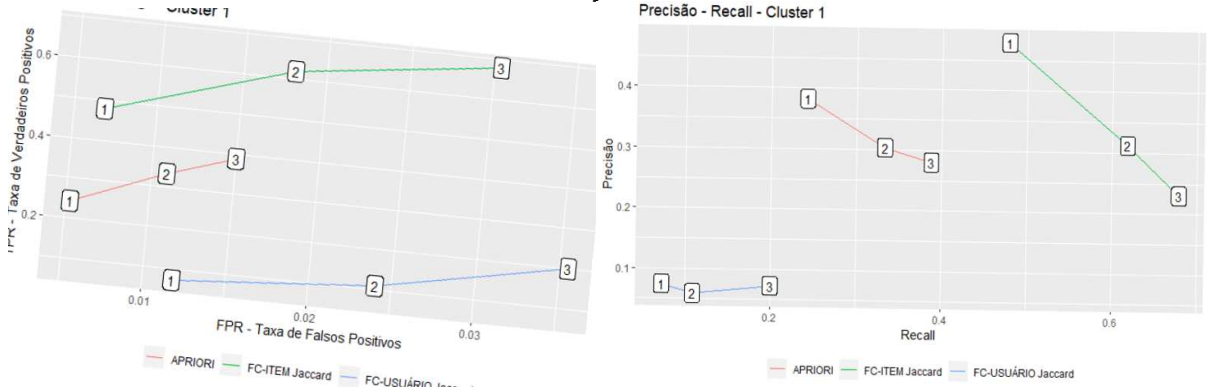
Observa-se que, a TPR sobe com o aumento do número de produtos recomendados, isto demonstra que os modelos estão acertando a classificação positiva, ou seja, há uma redução na classificação errada de produtos relevantes (Falsos Negativos).

É fato que em quanto se aumenta a TPR, também há um aumento na FPR, porém o incremento é mínimo, iniciando em 0,007, na Filtragem Colaborativa Baseada no Item, para a recomendação de um produto, e chegando a 0,030 para três produtos.

Porém, com o aumento do Recall (ou TPR) há uma redução na Precisão, quando se recomenda mais de um produto, o que evidência um aumento de recomendações irrelevantes. Isto justifica-se pela proposição do teste que somente ocultou um produto.

A Filtragem Colaborativa Baseada no Item, também se destaca na métrica Precisão, se comparada aos demais modelos.

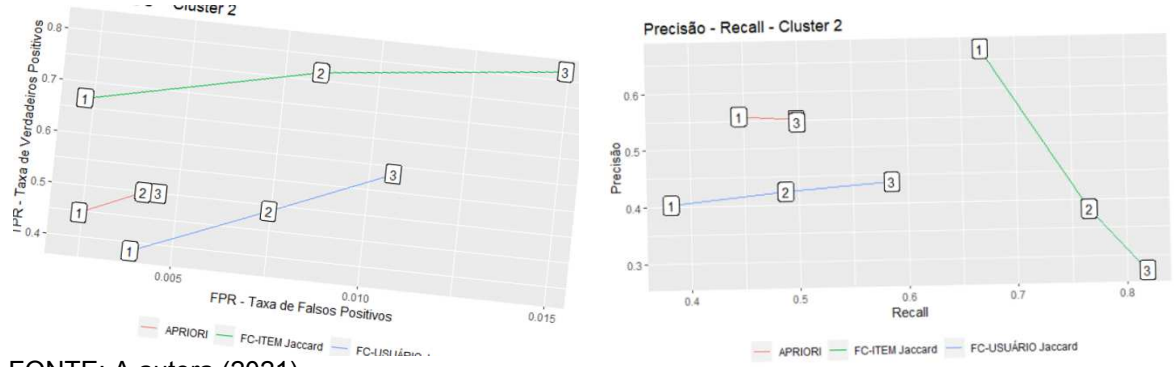
FIGURA 17 - COMPARAÇÃO DE MODELOS NO CLUSTER 1



FONTE: A autora (2021).

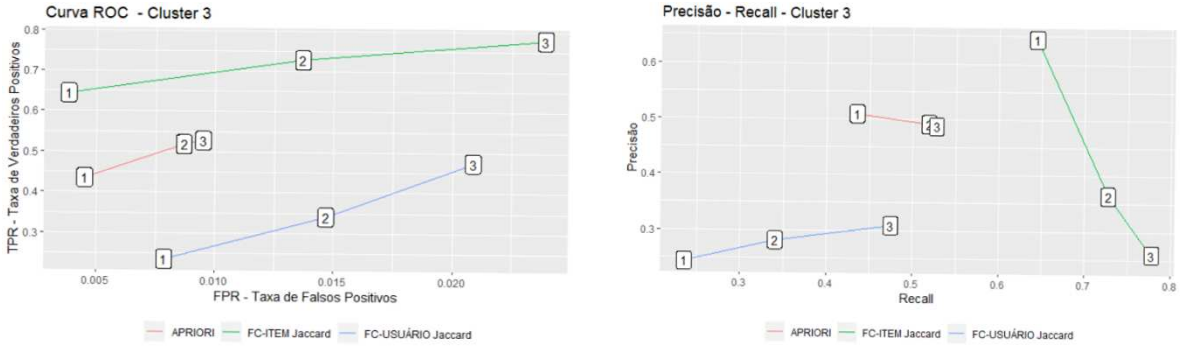
No cluster 2 (FIGURA 18), assim como nos demais, com exceção do cluster 11, poderá ser verificado o mesmo padrão. Em todos eles a Filtragem Colaborativa Baseada no Item mantém melhores resultados se comparado com os demais modelos. Porém nota-se que para o cluster 2 os resultados foram melhores se comparado com o cluster 1, com TPR, para um produto, igual a 0,6681 e Precisão de 0,6689, enquanto o cluster 1 apresentou 0,479 e 0,479 respectivamente para TPR e Precisão.

FIGURA 18 - COMPARAÇÃO DE MODELOS NO CLUSTER 2



FONTE: A autora (2021).

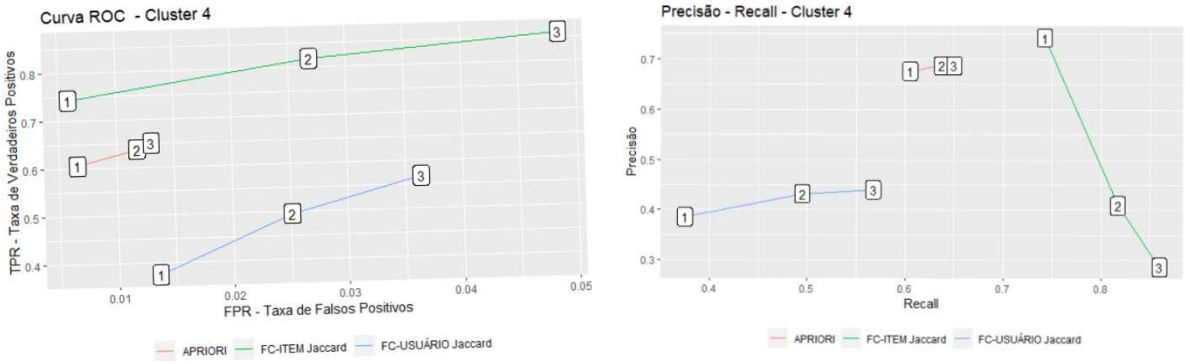
FIGURA 19 - COMPARAÇÃO DE MODELOS NO CLUSTER 3



FONTE: A autora (2021).

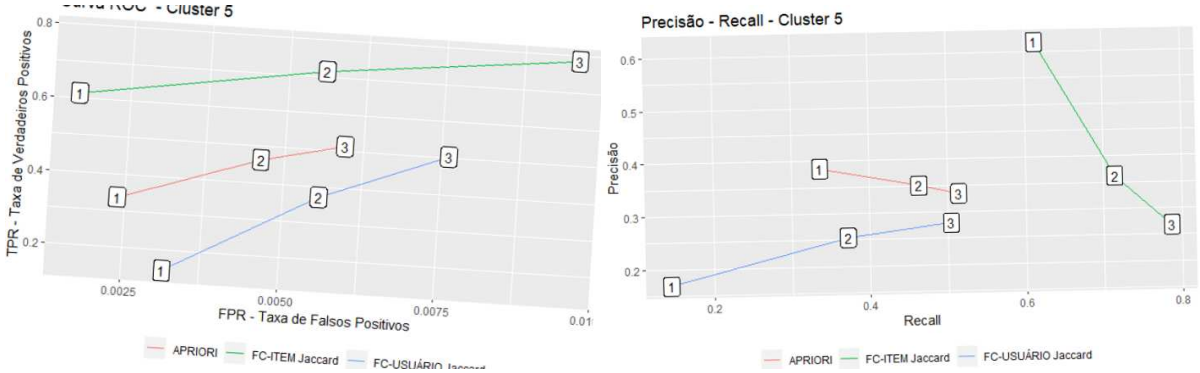
O cluster 4 (FIGURA 20), foi o que apresentou a melhor TPR e Precisão para a recomendação de um produto, com 0,742 e 0,742.

FIGURA 20 - COMPARAÇÃO DE MODELOS NO CLUSTER 4



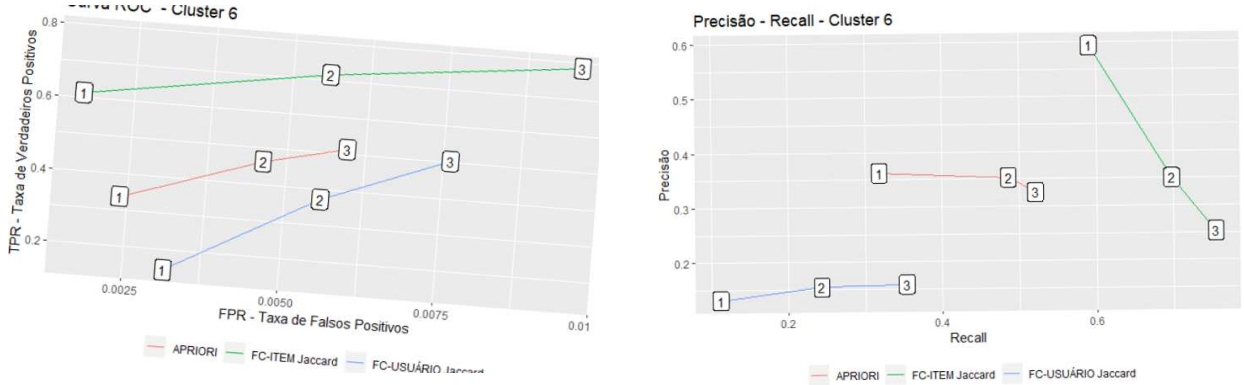
FONTE: A autora (2021).

FIGURA 21 - COMPARAÇÃO DE MODELOS NO CLUSTER 5



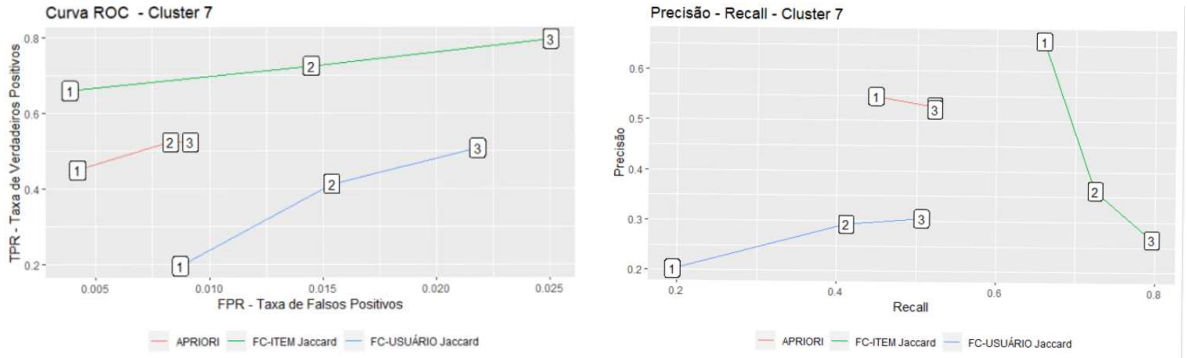
FONTE: A autora (2021).

FIGURA 22 - COMPARAÇÃO DE MODELOS NO CLUSTER 6



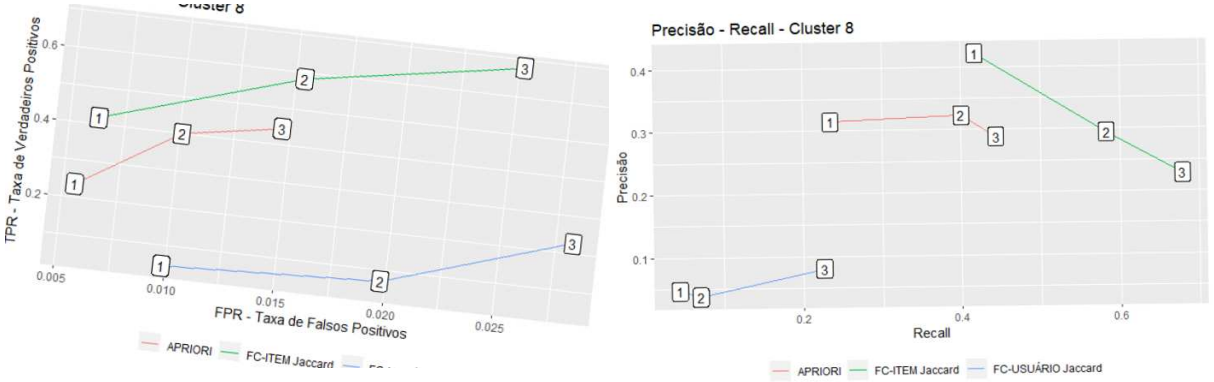
FONTE: A autora (2021).

FIGURA 23 - COMPARAÇÃO DE MODELOS NO CLUSTER 7



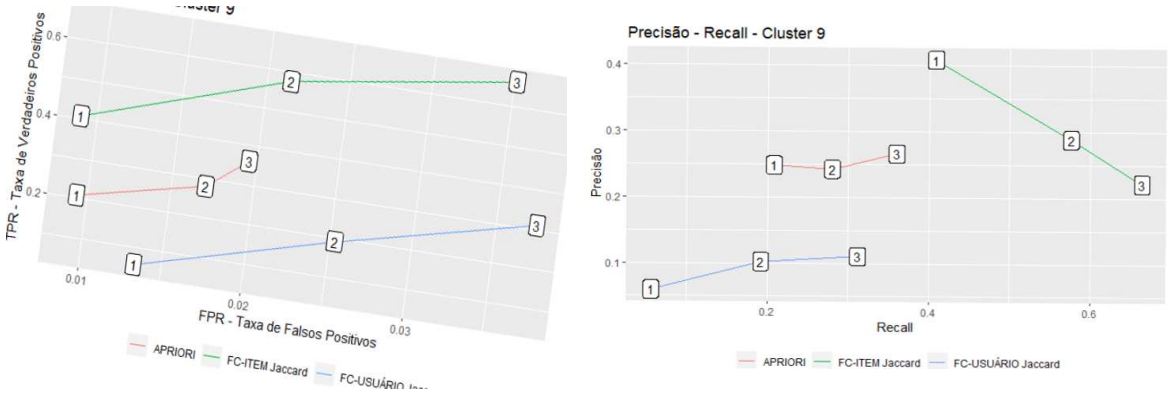
FONTE: A autora (2021).

FIGURA 24 - COMPARAÇÃO DE MODELOS NO CLUSTER 8



FONTE: A autora (2021).

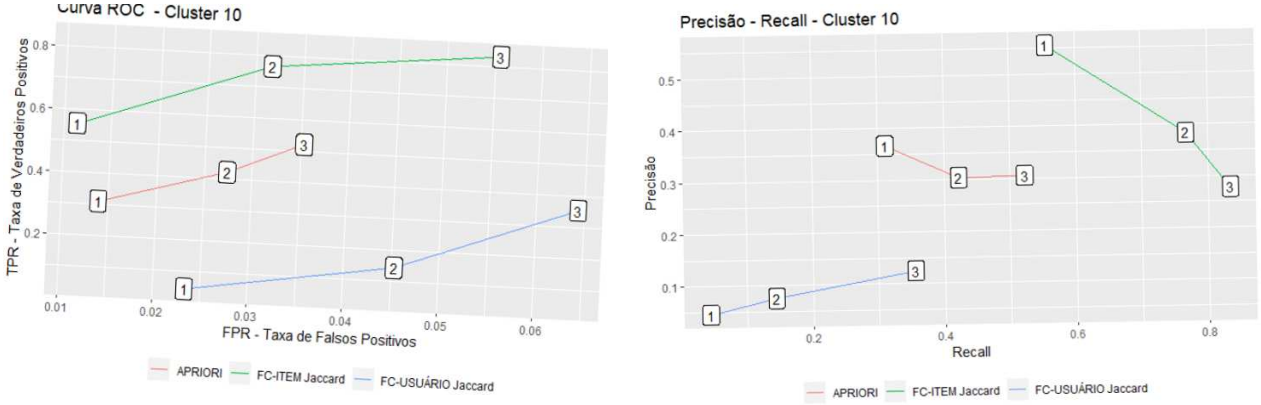
FIGURA 25 - COMPARAÇÃO DE MODELOS NO CLUSTER 9



FONTE: A autora (2021).

O cluster 9 (FIGURA 26), foi o que apresentou a piores TPR e Precisão para a recomendação de um produto considerando a Filtragem baseada no item, com 0,408 para ambos os índices.

FIGURA 26 - COMPARAÇÃO DE MODELOS NO CLUSTER 10



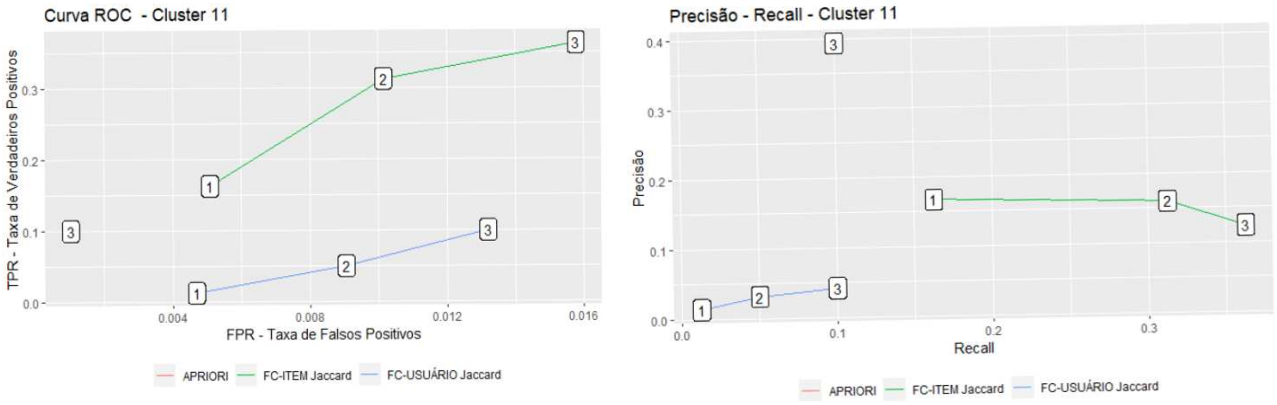
FONTE: A autora (2021).

Já para o cluster 11 (FIGURA 27), composto pelo perfil de Médias e Grandes empresas, o algoritmo Apriori apresentou melhor Precisão, para recomendação de um produto (0,393), se comparado aos modelos de Filtragem Colaborativa, que apresentaram Precisão de 0,167 e 0,013 para Item e Usuário respectivamente.

A Filtragem Colaborativa Baseada no Item apresentou Recall de 0,162, um pouco maior que o Apriori com 0,100, porém, visto que a diferença apresentada é pequena dado a Precisão dos modelos, preferiu-se o Apriori, que possui maior Precisão.

Contudo, para os parâmetros de suporte e confiança configurados, o modelo somente conseguiu recomendar um produto, por isto não há variação nas métricas de Precisão e Recall.

FIGURA 27 - COMPARAÇÃO DE MODELOS NO CLUSTER 11



FONTE: A autora (2021).

Portanto nesta análise foi possível perceber que a Filtragem Colaborativa Baseada no Item foi a mais indicada para os primeiros dez clusters, isto segundo as métricas de Precisão e Recall. Para esses cluster houve o mesmo padrão de recomendação, porém foi evidenciado melhor desempenho no cluster 4 e pior no cluster 8.

Já para o cluster 11, o modelo mais indicado foi o Apriori, por apresentar maior taxa de Precisão, se comparado aos demais.

4.3.6 Etapa 6: Ajuste nos parâmetros

Visto que a Filtragem Colaborativa Baseada no Item foi o melhor modelo para os dez primeiros clusters, nesta fase foi executada uma variação do conjunto de n itens mais próximos (produtos) que o algoritmo utiliza para calcular similaridade e então recomendar, com vistas a melhorar o desempenho do modelo.

Para esta análise não foram considerados conjuntos de itens mais próximos menores que 15, pois poderia deixar o algoritmo muito instável. Assim foram definidos conjuntos contendo de 15 a 60 produtos.

Com o intuito de verificar qual a melhor quantidade de itens, foi utilizada uma medida de desempenho proposta por Gorakala e Usuelli (2015), que é uma média

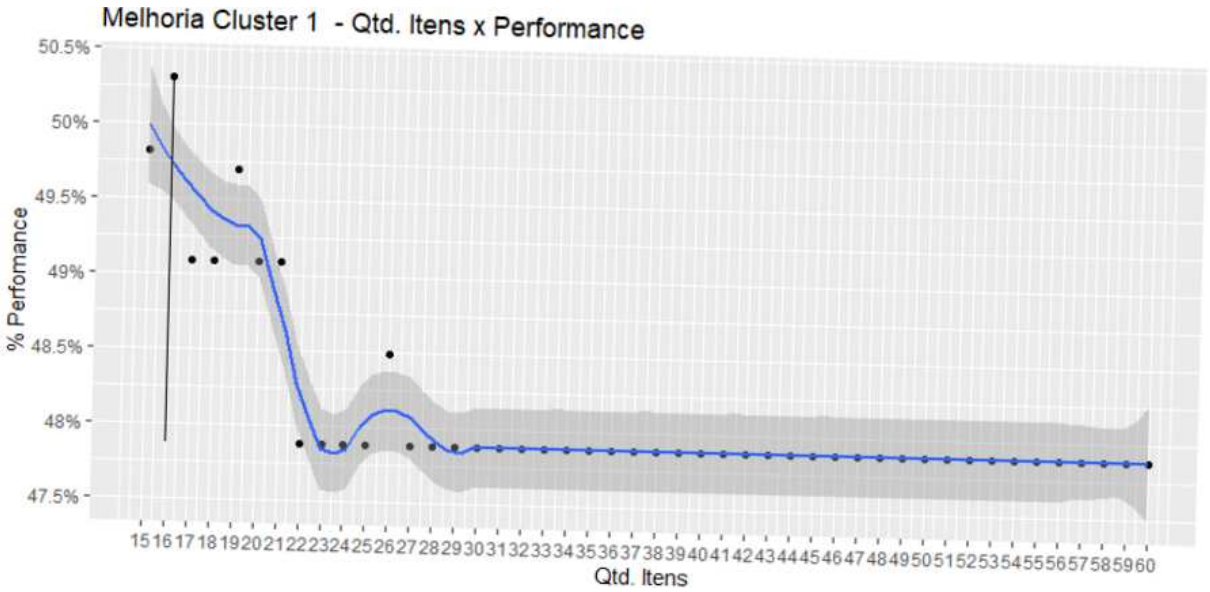
ponderada entre Precisão e Recall, definida com Performance, a qual foi utilizada somente na recomendação de um item.

As figuras a seguir ilustram a variação da Performance considerando a quantidade de itens mais próximos para cada cluster. As faixas em cinza, auxiliam a verificar o comportamento dos dados, e foram geradas pela função *geom_smooth*() do pacote *ggplot2* (HADLEY WICKHAM, 2016).

Já as tabelas comparam o conjunto de itens mais próximo inicial, com o conjunto que teve a melhor performance, demonstrando as diferenças para as métricas de Precisão, Recall, TPR e FPR.

A FIGURA 28 mostra, para o cluster 1, a variação da Performance em relação ao conjunto de itens mais próximos. Nela é possível notar que um conjunto menor de itens gera melhoras no indicador de Performance, sedo o melhor índice atingido para um conjunto de 15 itens.

FIGURA 28 - MELHORIA CLUSTER 1



FONTE: A autora (2021).

Na TABELA 54 - MELHORIA CLUSTER 1 é demonstrado o valor de cada uma das métricas, para o número de itens mais próximos utilizado inicialmente (30 itens) e o apontado na FIGURA 28, que apresenta a melhor performance (16 itens). Nota-se que houve um acréscimo de 0,024 na Precisão e Recall com a utilização de 15 itens.

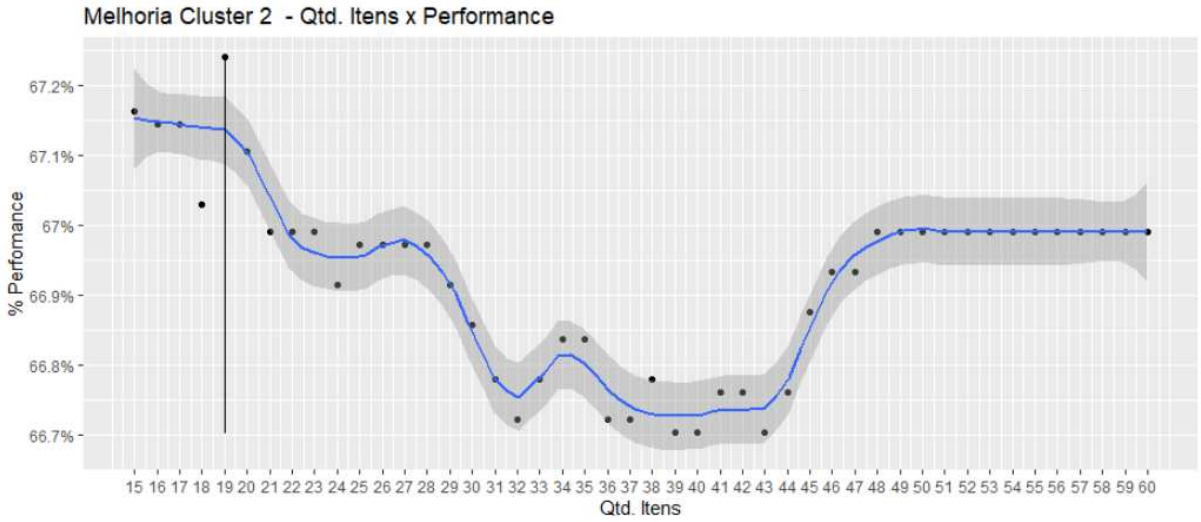
TABELA 54 - MELHORIA CLUSTER 1

	CLUSTER 1			
ITENS	PRECISÃO	RECALL	TPR	FPR
30	0,479	0,479	0,450	0,007
16	0.503	0.503	0.503	0.006
Diferença	0,024	0,024	0,024	0

FONTE: A autora (2021).

A FIGURA 29 mostra que o conjunto contendo 19 de itens é o qual se tem a maior Performance.

FIGURA 29 - MELHORIA CLUSTER 2



FONTE: A autora (2021).

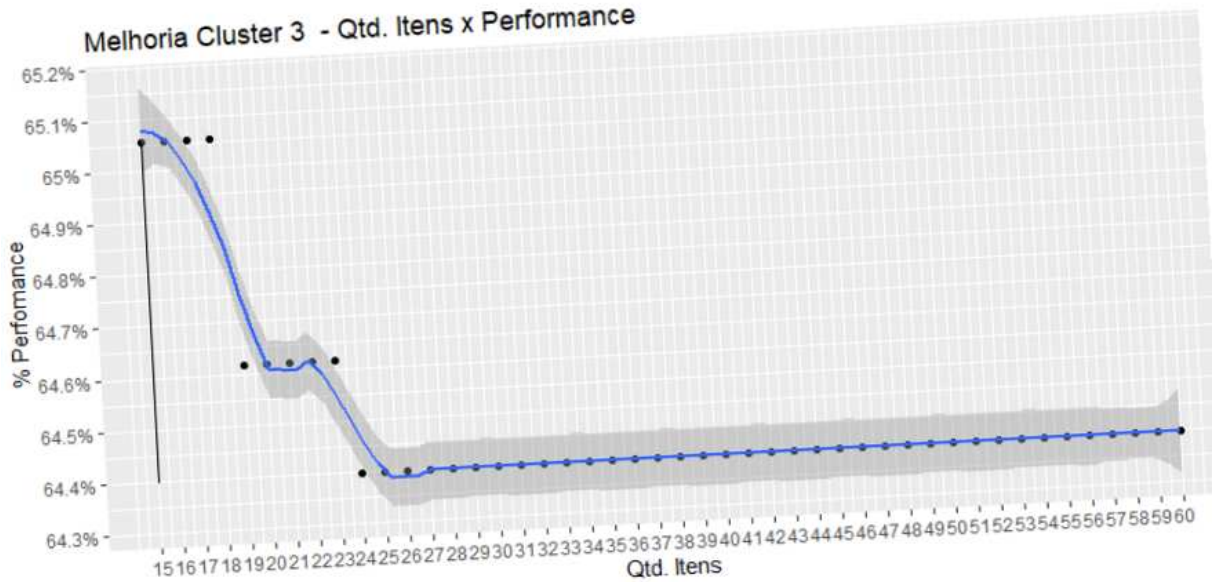
Na TABELA 55 é evidenciado que embora há uma melhoria nas métricas ela é pequena.

TABELA 55 - MELHORIA CLUSTER 2

	CLUSTER 2			
ITENS	PRECISÃO	RECALL	TPR	FPR
30	0,669	0,668	0,668	0,002
19	0,673	0,672	0,672	0,002
Diferença	0,004	0,003	0,003	0

FONTE: A autora (2021).

FIGURA 30 - MELHORIA CLUSTER 3



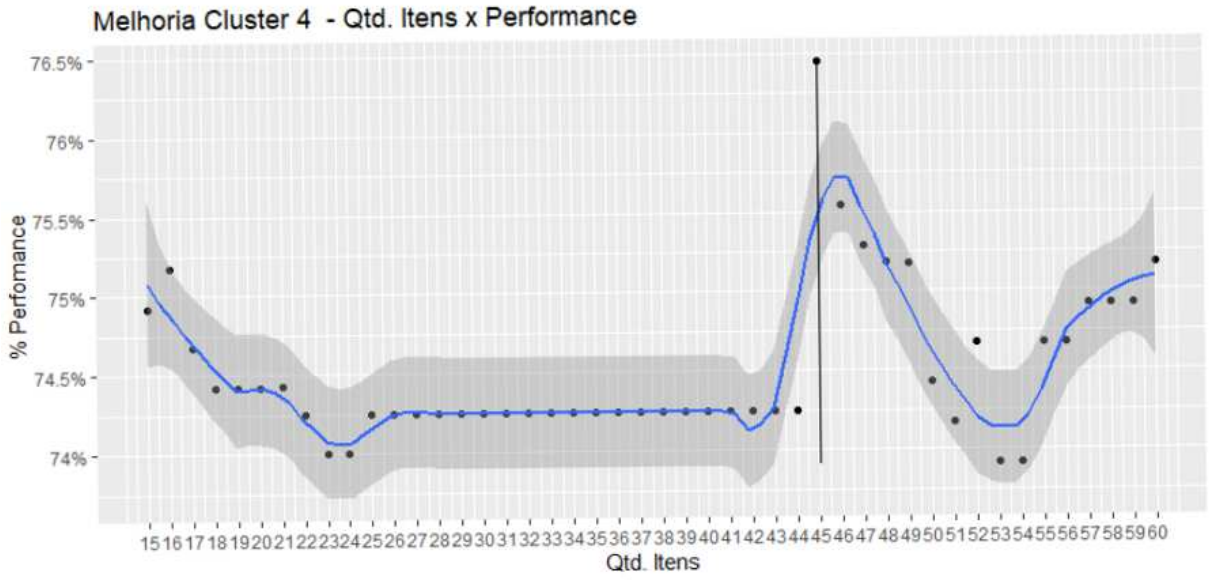
FONTE: A autora (2021).

TABELA 56 - MELHORIA CLUSTER 3

CLUSTER 3				
ITENS	PRECISÃO	RECALL	TPR	FPR
30	0,644	0,644	0,644	0,003
15	0.672	0.670	0.670	0,002
DIFERENÇA	0,028	0,026	0,026	0

FONTE: A autora (2021).

FIGURA 31 - MELHORIA CLUSTER 4



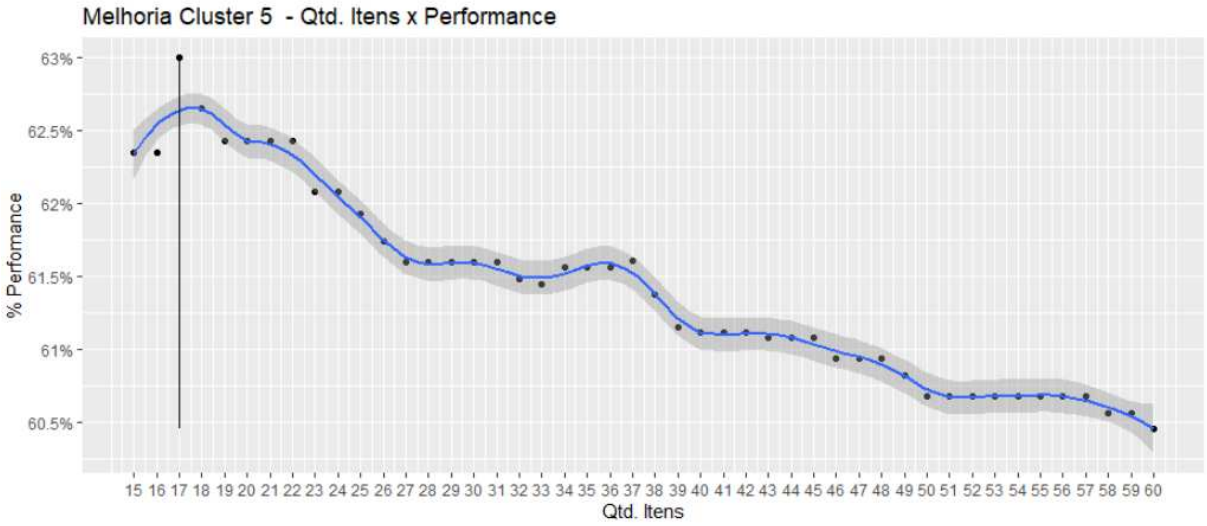
FONTE: A autora (2021).

TABELA 57 - MELHORIA CLUSTER 4

	CLUSTER 4			
ITENS	PRECISÃO	RECALL	TPR	FPR
30	0,742	0,742	0,742	0,005
45	0.766	0.745	0.745	0,004
Diferença	0,024	0,003	0,003	-0,001

FONTE: A autora (2021).

FIGURA 32 - MELHORIA CLUSTER 5



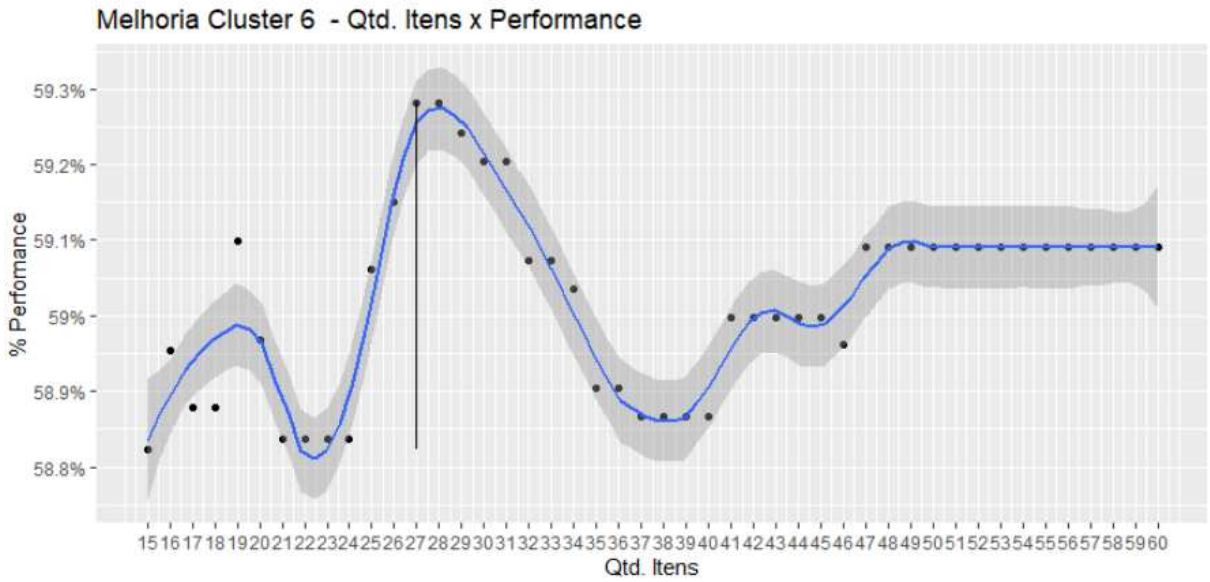
FONTE: A autora (2021).

TABELA 58 - MELHORIA CLUSTER 5

	CLUSTER 5			
ITENS	PRECISÃO	RECALL	TPR	FPR
30	0,618	0,614	0,614	0,002
17	0,634	0,626	0,626	0,002
Diferença	0,016	0,012	0,012	0

FONTE: A autora (2021).

FIGURA 33 - MELHORIA CLUSTER 6



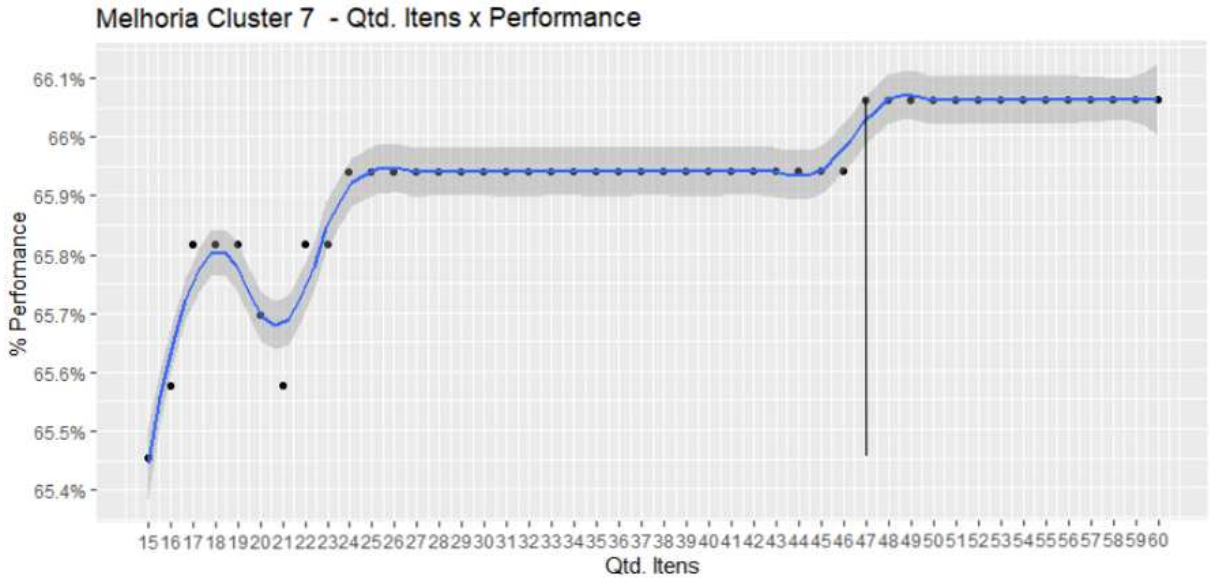
FONTE: A autora (2021).

TABELA 59 - MELHORIA CLUSTER 6

	CLUSTER 6			
ITENS	PRECISÃO	RECALL	TPR	FPR
30	0,593	0,591	0,591	0,002
27	0,595	0,591	0,591	0,002
Diferença	0,002	0,000	0,000	0,000

FONTE: A autora (2021).

FIGURA 34 - MELHORIA CLUSTER 7



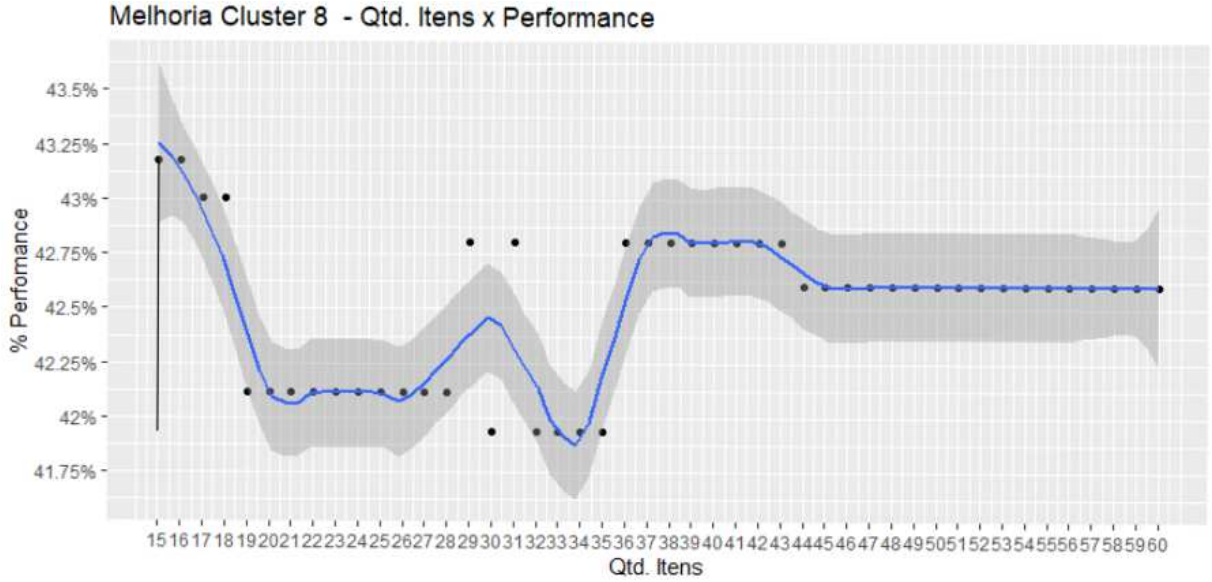
FONTE: A autora (2021).

TABELA 60 - MELHORIA CLUSTER 7

CLUSTER 7				
ITENS	PRECISÃO	RECALL	TPR	FPR
30	0,659	0,659	0,659	0,003
47	0,661	0,661	0,661	0,004
DIFERENÇA	0,002	0,002	0,002	0,001

FONTE: A autora (2021).

FIGURA 35 - MELHORIA CLUSTER 8



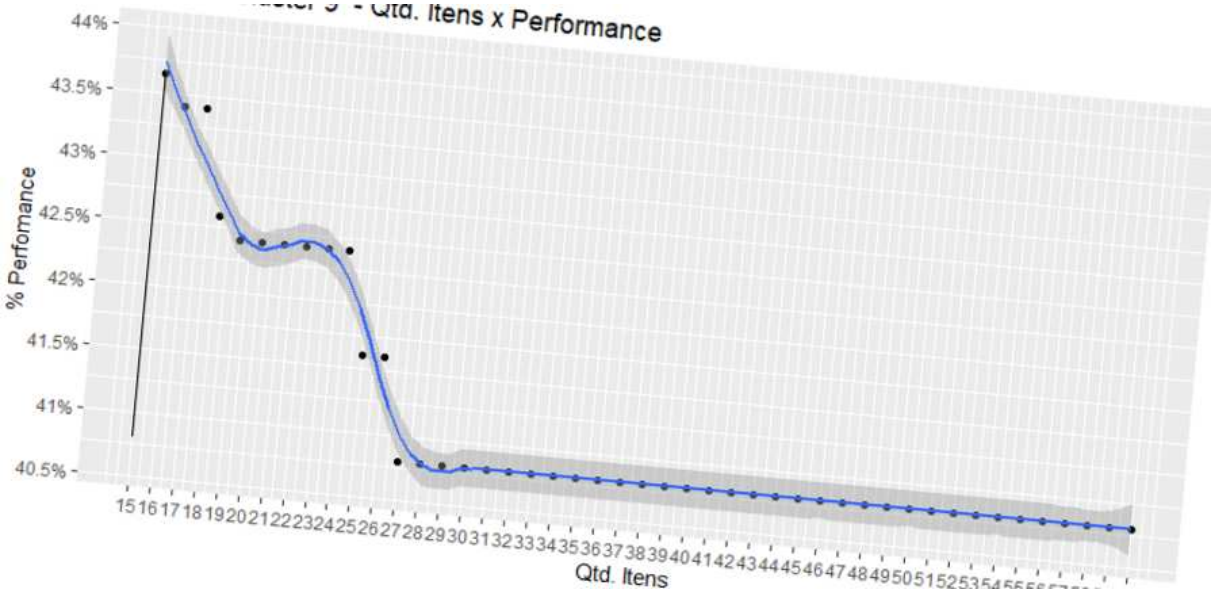
FONTE: A autora (2021).

TABELA 61 - MELHORIA CLUSTER 8

CLUSTER 8				
ITENS	PRECISÃO	RECALL	TPR	FPR
30	0,421	0,417	0,417	0,006
15	0.437	0.426	0.426	0,006
Diferença	0,016	0,009	0,009	0

FONTE: A autora (2021).

FIGURA 36 - MELHORIA CLUSTER 9



FONTE: A autora (2021).

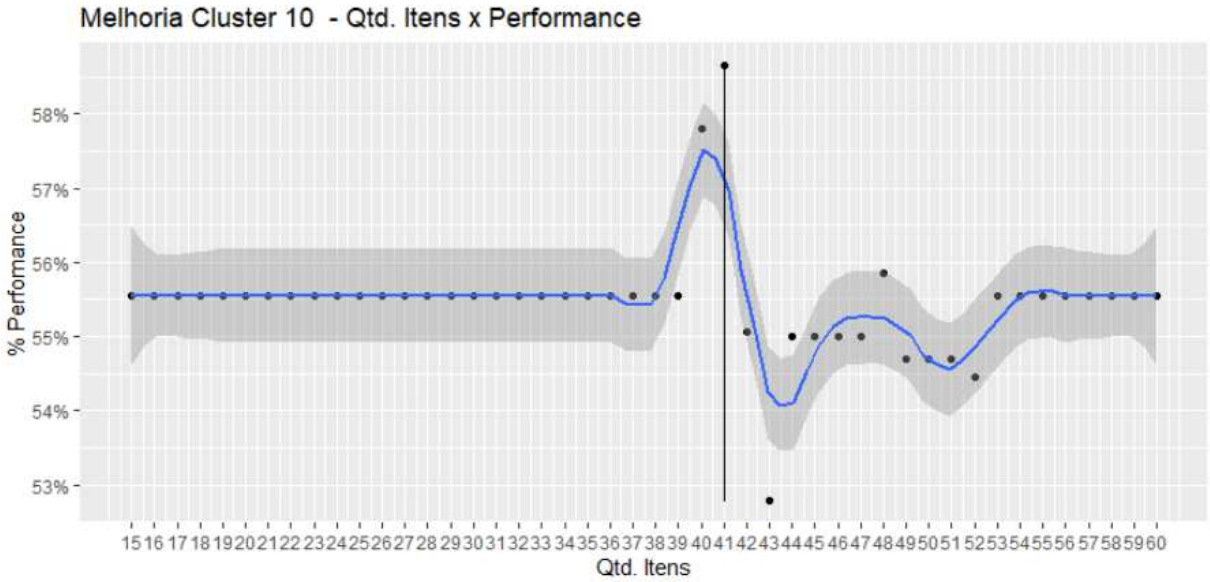
O cluster 9, foi o que teve a pior desempenho, comparado com os demais clusters, mas com a melhoria no parâmetro, utilizando quinze itens mais próximos, teve um acréscimo de 0,033 (8,09%) na Precisão, como pode ser observado na TABELA 62.

TABELA 62 - MELHORIA CLUSTER 9

	CLUSTER 9			
ITENS	PRECISÃO	RECALL	TPR	FPR
30	0,408	0,408	0,408	0,009
15	0,441	0,432	0,432	0,008
Diferença	0,033	0,024	0,024	-0,001

FONTE: A autora (2021).

FIGURA 37 - MELHORIA CLUSTER 10



FONTE: A autora (2021).

TABELA 63 - MELHORIA CLUSTER 10

	CLUSTER 10			
ITENS	PRECISÃO	RECALL	TPR	FPR
30	0,556	0,556	0,556	0,114
41	0.595	0.578	0.577	0,010
Diferença	0,034	0,011	0,011	-0,001

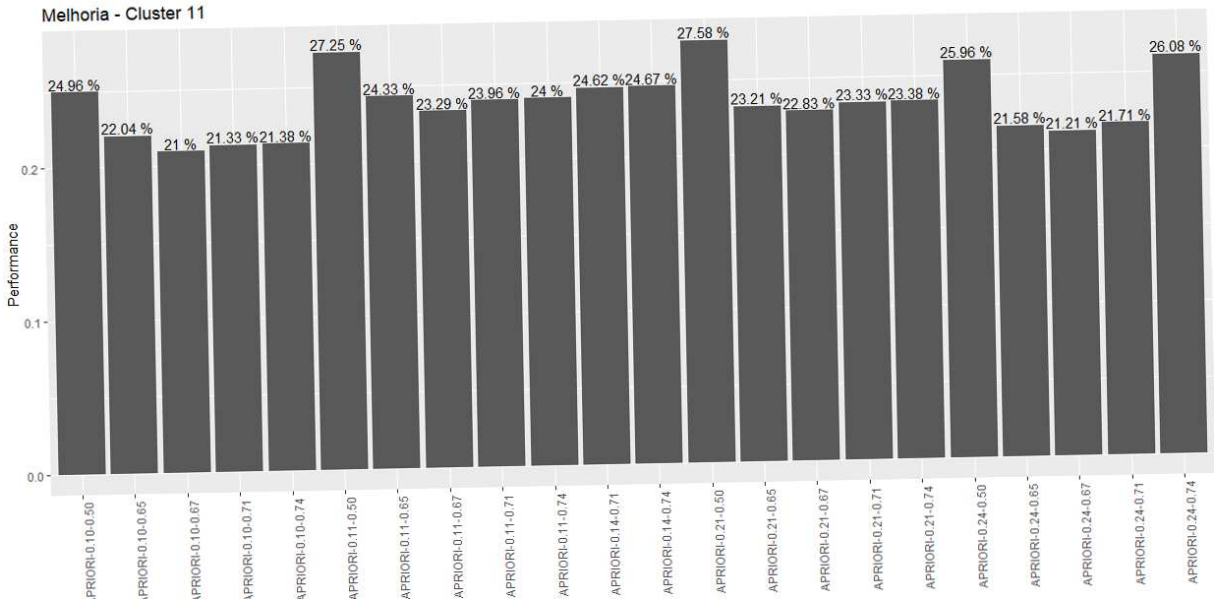
FONTE: A autora (2021).

Pode-se notar que a alteração no conjunto de itens próximos trouxe pequenas melhorias nas métricas, sendo o cluster 9 foi um dos que mais apresentou diferença positiva nas métricas de Precisão e Recall, com aumento de 0,033 e 0,024, respectivamente. Já o cluster 6 foi o que menos apresentou melhora, tendo somente um acréscimo de 0,002 na Precisão.

No cluster 11, em que o modelo Apriori teve melhor desempenho comparado aos demais, foi mensurada a performance para 697 modelos Apriori que tiveram alterações no parâmetro, combinando-se Suporte e Confiança, iniciando com Suporte de 0,1 até 0,26 e Confiança de 0,5 até 0,9. Na FIGURA 38, foi plotado somente os modelos com menores parâmetros que tiveram alguma variação na Performance. A utilização de Suporte final igual a 0,26 foi definida pois índices maiores geravam erro no algoritmo, demonstrando que era o suporte máximo do conjunto analisado.

Percebe-se que os modelos Apriori com suporte 0,21 e confiança 0,50 gerou a maior performance, 27,58%.

FIGURA 38 - MELHORIA CLUSTER 11



FONTE: A autora (2021).

Na TABELA 64 é possível comparar as métricas para o modelo inicial e o que apresentou a melhor performance, conforme FIGURA 38. Nota-se que a alteração nos parâmetros promoveu um aumento de 0,09 na Precisão e de 0,05 no Recall.

TABELA 64 - MELHORIA CLUSTER 11

	CLUSTER 11			
PARÂMETROS	PRECISÃO	RECALL	TPR	FPR
SUP=0,2 CONF = 0,75	0,393	0,100	0,100	0,001
SUP=0,21 CONF=0,50	0,402	0,150	0,150	0,001
Diferença	0,009	0,050	0,050	0,000

FONTE: A autora (2021).

A melhoria demostrou que ajustes simples nos parâmetros podem aprimorar a eficiência da recomendação de produtos.

5 CONSIDERAÇÕES FINAIS

O objetivo desta pesquisa foi criar um método combinando técnicas de agrupamento de clientes e regras de associação para recomendar produtos. Isto foi proposto pois a combinação delas pode auxiliar as empresas a estruturar uma estratégia de comunicação mais eficiente com os clientes de cada grupo, além de indicar produtos relevantes, o que melhora o relacionamento com o público.

Isto posto, o primeiro desafio foi clusterizar os clientes, e embora a literatura tenha apontado uma grande de utilização do K-means, este algoritmo se mostrou inapropriado para o conjunto de variáveis disponíveis no estudo de caso. Pois ao analisar os dados constatou-se que a base era mista, ou seja, possuía tanto atributos quantitativos quanto qualitativos. Contratempo que foi superado com a utilização do algoritmo K-medoid e medidas de distância compatíveis com os tipos de variáveis.

A pesquisa, além de contribuir com um apanhado teórico sobre clusterização, evidenciou aspectos técnicos como a importância do preparo da base para aplicação dos algoritmos, além disso evidenciou que os tipos de variáveis em uma base são cruciais para a escolha da técnica a ser utilizada. Também apresentou comparações de duas medidas de distâncias, o que corrobora na apresentação e demonstração do método utilizado.

Já no que está tange a recomendação, o desafio foi a aplicação dos algoritmos em uma base de transações esparsa, em que poucos produtos geraram a grande maioria dos atendimentos. Nesta fase a análise descritiva da base foi fundamental, pois permitiu observar a disposição dos dados, o que fundamentou a decisão de usar parte da base de transações, reforçando a importância da utilização de análise descritiva antes da aplicação dos algoritmos.

Além disso, a pesquisa apresenta os métodos e detalha as métricas utilizadas para a comparação de algoritmos de recomendação, bem como mostra a influência de vizinhos próximo (produtos) na melhora da performance do modelo.

Desta maneira esta pesquisa demonstra de maneira clara todo o processo a ser seguido na implantação de técnicas de clusterização e recomendação de produtos, objetivando ser um norte para as empresas que buscam a implantação destas técnicas.

Tudo que foi exposto demonstra que o trabalho conseguiu cumprir o seu proposito na elaboração de método que combinasse a clusterização de clientes e a

recomendação de produtos, partindo desde a análise descritiva dos dados, escolha dos algoritmos e melhora de performance.

Para trabalhos futuros, sugere-se incluir na análise de modelos o algoritmo FP-Growth, dado sua expressividade na revisão de literatura.

REFERÊNCIAS

AHMAD, A.; DEY, L. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. **Pattern Recognition Letters**, v. 28, n. 1, p. 110–118, 2007a.

AHMAD, A.; DEY, L. A k-mean clustering algorithm for mixed numeric and categorical data. **Data and Knowledge Engineering**, v. 63, n. 2, p. 503–527, 2007b.

BABU, G.; BHUVANESWARI, T. Association rule mining for identifying optimal customers using MAA algorithm. **Journal of Theoretical and Applied Information Technology**, v. 66, n. 3, p. 829–838, 2014.

BAFGHI, E. P. Clustering of Customers Based on Shopping Behavior and Employing GeneticAlgorithms. **ENGINEERING TECHNOLOGY & APPLIED SCIENCE RESEARCH**, v. 7, n. 1, p. 1420–1424, fev. 2017.

BEHESHTIAN-ARDAKANI, A.; FATHIAN, M.; GHOLAMIAN, M. A novel model for product bundling and direct marketing in e-commerce based on market segmentation. **Decision Science Letters**, v. 7, n. 1, p. 39–54, 2018.

BEKTAS, A.; SCHUMANN, R. How to Optimize Gower Distance Weights for the k-Medoids Clustering Algorithm to Obtain Mobility Profiles of the Swiss Population. **Proceedings - 6th Swiss Conference on Data Science, SDS 2019**, p. 51–56, 2019.

BI, W. et al. A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn. **IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS**, v. 12, n. 3, p. 1270–1281, jun. 2016.

BISCHL, B. et al. mlr : Machine Learning in R. **Journal of Machine Learning Research**, v. 17, n. 170, p. 1–5, 2016.

BOSE, I.; CHEN, X. Exploring business opportunities from mobile services data of customers: An inter-cluster analysis approach. **ELECTRONIC COMMERCE RESEARCH AND APPLICATIONS**, v. 9, n. 3, SI, p. 197–208, 2010.

BRENTARI, E.; DANCELLI, L.; MANISERA, M. Clustering ranking data in market segmentation: a case study on the Italian McDonald's customers' preferences. **JOURNAL OF APPLIED STATISTICS**, v. 43, n. 11, p. 1959–1976, 2016.

BUDIAJI, W.; LEISCH, F. Simple k-medoids partitioning algorithm for mixed variable data. **Algorithms**, v. 12, n. 9, p. 1–15, 2019.

CHA, S.-H. Taxonomy of nominal type histogram distance measures. **City**, v. 1, n. 2, p. 1, 2008.

CHANG, C.-I.; HO, J.-C. A Two-Layer Clustering Model for Mobile Customer Analysis. **IT Professional**, v. 19, n. 3, p. 38–44, 2017.

CHEN, T. The RFM-FCM approach for customer clustering. **International Journal of Technology Intelligence and Planning**, v. 8, n. 4, p. 358–373, 2012.

CHEN, Y. L. et al. Market basket analysis in a multiple store environment. **DECISION SUPPORT SYSTEMS**, v. 40, n. 2, p. 339–354, 2005.

CHIANG, D.-A. et al. Mining disjunctive consequent association rules. **APPLIED**

SOFT COMPUTING, v. 11, n. 2, p. 2129–2133, mar. 2011.

DE ASSIS, E. C.; DE SOUZA, R. M. C. R. A K-medoids clustering algorithm for mixed feature-type symbolic data. **Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics**, p. 527–531, 2011.

DELLAERT, B. G. C.; HÄUBL, G. Searching in choice mode: Consumer decision processes in product search with recommendations. **Journal of Marketing Research**, v. 49, n. 2, p. 277–288, 2012.

DRESCH, A.; LACERDA, D. P. **Design science research: método de pesquisa para avanço da ciência e tecnologia**. Bookman ed. [s.l: s.n.].

FÁVERO, L. P.; BELFIORE, P. **HARIKUMAR SURYA, 2015**. [s.l: s.n.].

GANMAWU, S. A.; WELLS, M. T. **Data Clustering**. [s.l: s.n.].

GOHR, C. F. et al. UM MÉTODO PARA REVISÃO SISTEMÁRITCA DA LITERATURA EM PESQUISAS DE ENGENHARIA DE PRODUÇÃO. **produção. Encontro Nacional de Engenharia de Produção**, v. 33, 2013.

GORAKALA, S. K.; USUELLI, M. **Building a Recommendation System with R**. [s.l: s.n.].

GUCDEMIR, H.; SELIM, H. Integrating multi-criteria decision making and clustering for business customer segmentation. **INDUSTRIAL MANAGEMENT & DATA SYSTEMS**, v. 115, n. 6, p. 1022–1040, 2015.

GUJARATI, D. N.; PORTER, D. C. **ECONOMETRIA BÁSICA**. Quinta Edi ed. [s.l.] Amgh Editora, 2011.

GUNAWARDANA, A.; SHANI, G. A survey of accuracy evaluation metrics of recommendation tasks. **Journal of Machine Learning Research**, v. 10, p. 2935–2962, 2009.

GUPTA, G.; AGGARWAL, H.; RANI, R. Segmentation of retail customers based on cluster analysis in building successful CRM. **International Journal of Business Information Systems**, v. 23, n. 2, p. 212–228, 2016.

HADLEY WICKHAM. **ggplot2: Elegant Graphics for Data Analysis**. [s.l.] Springer-Verlag New York, 2016.

HAHSLER, M. recommenderlab: A Framework for Developing and Testing Recommendation Algorithms. **Nov**, p. 1–37, 2011.

HAHSLER, M. **recommenderlab: Lab for Developing and Testing Recommender Algorithms**, 2020. Disponível em: <https://cran.r-project.org/package=recommenderlab>

HAN, J.; KAMBER, M.; PEI, J. **Data Mining Concepts and Techniques**. Third Edit ed. [s.l.] Morgan Kaufmann, 2012.

HARIKUMAR, S.; SURYA, P. V. K-Medoid Clustering for Heterogeneous DataSets. **Procedia Computer Science**, v. 70, p. 226–237, 2015.

HEMALATHA, M. Market basket analysis - A data mining application in Indian retailing. **International Journal of Business Information Systems**, v. 10, n. 1, p. 109–129, 2012.

HENNIG, C. Clustering strategy and method selection. **Handbook of Cluster**

Analysis, p. 703–730, 2015.

HIZIROGLU, A.; SENBAS, U. D. An application of fuzzy clustering to customer portfolio analysis in automotive industry. **International Journal of Fuzzy System Applications**, v. 5, n. 2, p. 13–25, 2016.

HOSSEINI, S. M. S.; MALEKI, A.; GHOLAMIAN, M. R. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. **EXPERT SYSTEMS WITH APPLICATIONS**, v. 37, n. 7, p. 5259–5264, jul. 2010.

HUANG, Y.-P. et al. Using back-propagation to learn association rules for service personalization. **Expert Systems with Applications**, v. 35, n. 1–2, p. 245–253, 2008.

HUNG, L. P. A personalized recommendation system based on product taxonomy for one-to-one marketing online. **EXPERT SYSTEMS WITH APPLICATIONS**, v. 29, n. 2, p. 383–392, 2005.

JALILI, M. et al. Evaluating Collaborative Filtering Recommender Algorithms: A Survey. **IEEE Access**, v. 6, p. 74003–74024, 2018.

KARA, A.; KAYNAK, E. Markets of a single customer: exploiting conceptual developments in market segmentation. **European Journal of Marketing**, v. 31, n. 11/12, p. 873–895, 1997.

KUO, R. J. et al. An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation. **NEUROCOMPUTING**, v. 205, p. 116–129, 2016.

LI, D.-C.; DAI, W.-L.; TSENG, W.-T. A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business. **EXPERT SYSTEMS WITH APPLICATIONS**, v. 38, n. 6, p. 7186–7191, jun. 2011.

LIN, Z.; GOH, K.-Y.; HENG, C.-S. The Demand Effects of Product Recommendation Networks: An Empirical Analysis of Network Diversity and Stability. **MIS Quarterly**, v. 41, n. 2, p. 397–426, 1 fev. 2017.

LINGRAS, P. et al. Temporal analysis of clusters of supermarket customers: conventional versus interval set approach. **INFORMATION SCIENCES**, v. 172, n. 1–2, p. 215–240, jun. 2005.

LIU, D. R.; SHIH, Y. Y. Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences. **JOURNAL OF SYSTEMS AND SOFTWARE**, v. 77, n. 2, p. 181–191, 2005a.

LIU, D. R.; SHIH, Y. Y. Integrating AHP and data mining for product recommendation based on customer lifetime value. **INFORMATION & MANAGEMENT**, v. 42, n. 3, p. 387–400, mar. 2005b.

LUO, Y. et al. Customer segmentation for telecom with the k-means clustering method. **Information Technology Journal**, v. 12, n. 3, p. 409–413, 2013.

MAECHLER, M. et al. **cluster: Cluster Analysis Basics and Extensions**, 2018.

MITRA, A. et al. Recommendation system based on product purchase analysis. **INNOVATIONS IN SYSTEMS AND SOFTWARE ENGINEERING**, v. 12, n. 3, SI, p. 177–192, 2016.

MOSTAFA, M. M. Knowledge discovery of hidden consumer purchase behaviour: A market basket analysis. **International Journal of Data Analysis Techniques and Strategies**, v. 7, n. 4, p. 384–405, 2015.

MUNUSAMY, S.; MURUGESAN, P. Modified dynamic fuzzy c-means clustering algorithm – Application in dynamic customer segmentation. **Applied Intelligence**, 2020.

MUSALEM, A.; ABURTO, L.; BOSCH, M. Market basket analysis insights to support category management. **EUROPEAN JOURNAL OF MARKETING**, v. 52, n. 7–8, p. 1550–1573, 2018.

PARK, H. S.; JUN, C. H. A simple and fast algorithm for K-medoids clustering. **Expert Systems with Applications**, v. 36, n. 2 PART 2, p. 3336–3341, 2009.

PROVOST, F.; FAWCETT, T. **Data science para negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados**. Rio de Janeiro: [s.n.].

R CORE TEAM. **R: A Language and Environment for Statistical Computing**Vienna, Austria, 2018. Disponível em: <https://www.r-project.org/>

RAJAGOPAL, S. Customer Data Clustering using Data Mining Technique. v. 3, n. 4, 2011.

RENUKA DEVI, V.; BHARATHI, G.; PRASAD, G. V. S. N. R. V. Prediction of customer churn in telecom sector using clustering technique. **International Journal of Engineering and Advanced Technology**, v. 8, n. 6 Special Issue 2, p. 826–832, 2019.

REYNOLDS, A. P.; RICHARDS, G.; RAYWARD-SMITH, V. J. The application of K-medoids and PAM to the clustering of rules. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 3177, p. 173–178, 2004.

ROUSSEEUW, PETER J.; KAUFMAN, L. **Finding groups in data**. [s.l.] Hoboken: Wiley Online Library, 1990.

SAJJADI, K. et al. A developing model for clustering and ranking bank customers. **International Journal of Electronic Customer Relationship Management**, v. 9, n. 1, p. 73–86, 2015.

SERET, A. et al. A dynamic understanding of customer behavior processes based on clustering and sequence mining. **EXPERT SYSTEMS WITH APPLICATIONS**, v. 41, n. 10, p. 4648–4657, 2014.

SHEIKH, A.; GHANBARPOUR, T.; GHOLAMIANGONABADI, D. A Preliminary Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting. **Journal of Business-to-Business Marketing**, v. 26, n. 2, p. 197–207, 2019.

SHIH, Y.-Y.; LIU, D.-R. Product recommendation approaches: Collaborative filtering via customer lifetime value and customer demands. **EXPERT SYSTEMS WITH APPLICATIONS**, v. 35, n. 1–2, p. 350–360, 2008.

SOHN, S. Y.; KIM, Y. Searching customer patterns of mobile service using clustering and quantitative association rule. **EXPERT SYSTEMS WITH APPLICATIONS**, v. 34, n. 2, p. 1070–1077, 2008.

SOLNET, D.; BORTUG, Y.; DOLNICAR, S. An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue. **INTERNATIONAL JOURNAL OF HOSPITALITY MANAGEMENT**, v. 56, p. 119–125, jul. 2016.

TAN, P.-N.; MICHAEL, S.; VIPIN, K. **Introduction to Data Mining**. [s.l.] Pearson Education India., 2016.

TSAI, C.-F.; HU, Y.-H.; LU, Y.-H. Customer segmentation issues and strategies for an automobile dealership with two clustering techniques. **EXPERT SYSTEMS**, v. 32, n. 1, p. 65–76, 2015.

TURČÍNEK, P.; TURČÍNKOVA, J. Exploring consumer behavior: Use of association rules. **Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis**, v. 63, n. 3, p. 1031–1042, 2015.

UMAMAHESWARI, M.; DEVI, P. I. Prediction of Myocardial Infarction Using K-Medoid Clustering Algorithm. **Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2017**, v. 2018- Febru, p. 1–6, 2018.

UNVAN, Y. A. Market basket analysis with association rules. **COMMUNICATIONS IN STATISTICS-THEORY AND METHODS**, 2020.

VALLE, M. A.; RUZ, G. A.; MORRAS, R. WOS:000425074100012 Market basket analysis: Complementing association rules with minimum spanning trees. **EXPERT SYSTEMS WITH APPLICATIONS**, v. 97, p. 146–162, 2018.

VIDELA-CAVIERES, I. F.; RIOS, S. A. Extending market basket analysis with graph mining techniques: A real case. **EXPERT SYSTEMS WITH APPLICATIONS**, v. 41, n. 4, 2, p. 1928–1936, mar. 2014.

WANG, C.-H. A market-oriented approach to accomplish product positioning and product recommendation for smart phones and wearable devices. **INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH**, v. 53, n. 8, p. 2542–2553, 2015.

WANG, C.-H.; CHIN, H.-T. Integrating affective features with engineering features to seek the optimal product varieties with respect to the niche segments. **ADVANCED ENGINEERING INFORMATICS**, v. 33, p. 350–359, 2017.

WENG, C.-H. Identifying association rules of specific later-marketed products. **APPLIED SOFT COMPUTING**, v. 38, p. 518–529, jan. 2016.

WICKHAM, A. et al. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019.

WITTEN, I. H. et al. **Data Mining: Practical Machine Learning Tools and Techniques**. Fourth Edi ed. [s.l.] Todd Green, 2017.

ZEKIC-SUSAC, M.; HAS, A. Discovering market basket patterns using hierarchical association rules. **CROATIAN OPERATIONAL RESEARCH REVIEW**, v. 6, n. 2, p. 475–487, 2015.

ZHANG, C.; ZHANG, S. **Association Rule Mining Models and Algorithms**. [s.l.] Springer-Verlag, 2002.

ZHENG, D. Application of silence customer segmentation in securities industry based on fuzzy cluster algorithm. **Journal of Information and Computational Science**, v. 10, n. 13, p. 4337–4347, 2013.

ZOERAM, A. A.; MAZIDI, A. K. A New Approach for Customer Clustering by IntegraZOERAM, A. A.; MAZIDI, A. K. A New Approach for Customer Clustering by Integrating the LRFM Model and Fuzzy Inference System. IRANIAN JOURNAL OF MANAGEMENT STUDIES, v. 11, n. 2, p. 351–378, ago. 2018. ting . **IRANIAN JOURNAL OF MANAGEMENT STUDIES**, v. 11, n. 2, p. 351–378, ago. 2018.