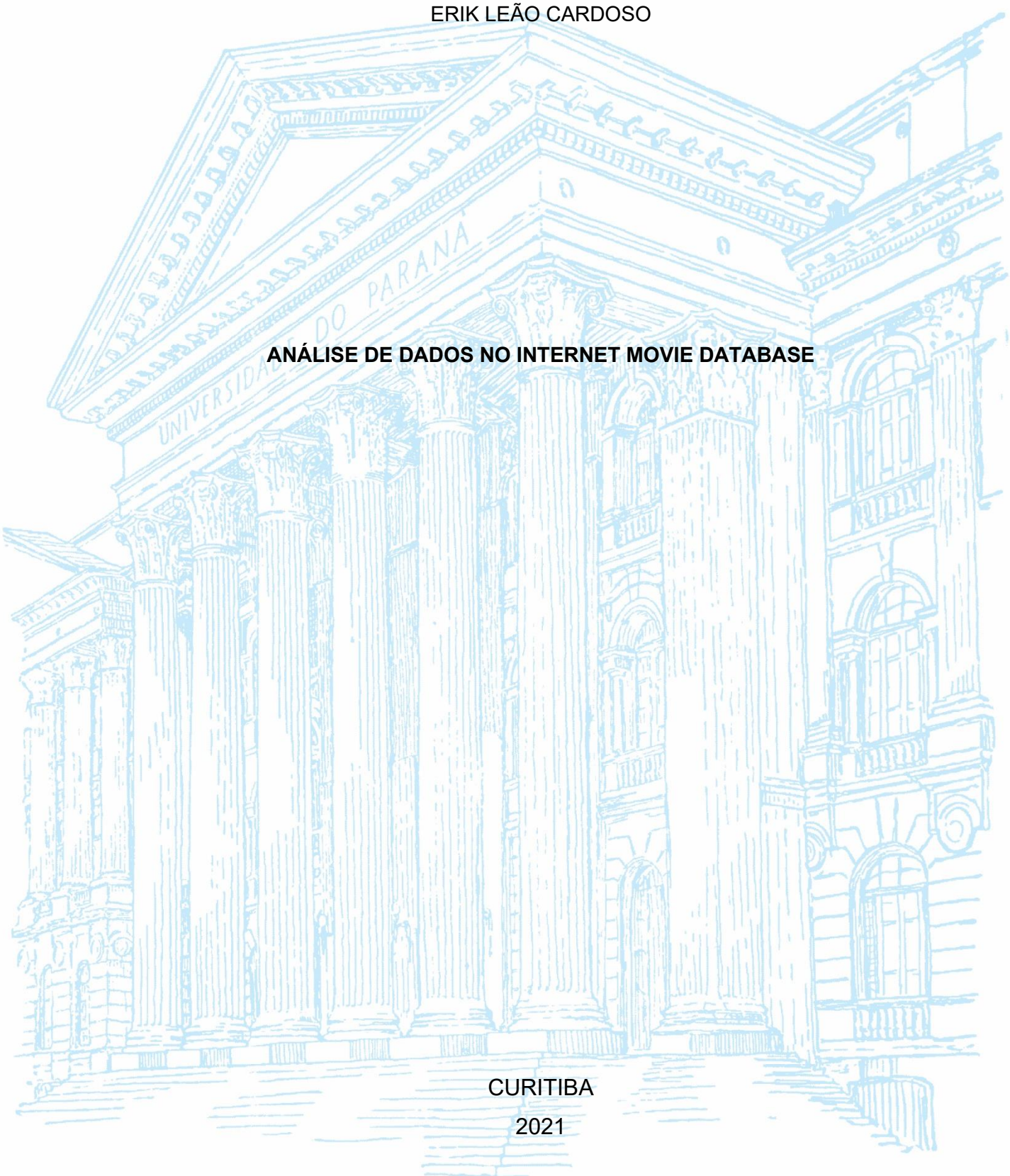
UNIVERSIDADE FEDERAL DO PARANÁ

ERIK LEÃO CARDOSO



ANÁLISE DE DADOS NO INTERNET MOVIE DATABASE

CURITIBA

2021

ERIK LEÃO CARDOSO

ANÁLISE DE DADOS NO INTERNET MOVIE DATABASE

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção de grau de Bacharel no Curso de Gestão da Informação, do Departamento de Ciência e Gestão da Informação, do Setor de Ciências Sociais Aplicadas, da Universidade Federal do Paraná.

Orientadora: Prof.ª Dr.ª Denise Fukumi Tsunoda

CURITIBA
2021

AGRADECIMENTOS

Eu gostaria de agradecer a minha família por me ajudar nessa caminhada difícil que foi durante o curso, por todo o seu apoio e em me dar forças por sempre acreditar em mim.

À professora Denise Fukumi Tsunoda que além e ser minha professora durante o curso em outras disciplinas, também foi minha orientadora neste trabalho de conclusão de curso e que sempre teve paciência e tirava minhas várias dúvidas durante o TCC, me apoiando sempre.

A minha namorada que sempre me incentivou e me apoiou nos momentos difíceis e felizes em minha trajetória.

A família da minha namorada que também sempre me apoiou durante esta trajetória e me auxiliou durante todo esse tempo.

E um muito obrigado a todos(as) as demais pessoas que me ajudaram durante o curso.

RESUMO

Este trabalho tem como objetivo realizar uma análise na base de dados do IMDB com os títulos de estreia de filmes e séries entre 2015 até 2019 a fim de se identificar padrões e tendencias. O levantamento feito sobre o tema mostrou a falta de artigos sobre esse tema, em aplicar o conhecimento de Gestão da Informação na área de entretenimento, e foi o motivador para a realização do presente estudo. O método utilizado nesta pesquisa foi o KDD (Descoberta de Conhecimento em Bases de Dados), por meio da seleção dos dados, pré processamento, transformação, mineração de dados e a avaliação das informações. Primeiramente, para realizar a análise da base de dados foi feita a delimitação do que seria utilizado e analisado na base de dados e assim selecionados quais arquivos do IMDB seriam utilizados. Depois de fazer a seleção foram unificados os dois arquivos, gerando uma base de dados única. Após gerar a base de dados foram realizadas limpezas de dados e normalizações com o intuito de poder analisar com estatísticas e mineração de dados. Para a realização das estatísticas, primeiramente foi identificado que a maior parte dos atributos são qualitativos e utilizou-se como ferramenta o Excel. Na mineração de dados foram selecionados os algoritmos de Árvore de Decisão e Naïve Bayes ambos executados no RStudio. A execução da Árvore de Decisão e do Naïve Bayes na base de dados mostrou que em ambos tiveram uma taxa de acerto de aproximadamente 70%. O trabalho encerra com os objetivos alcançados e com ideias de utilização de outras ferramentas, algoritmos com objetivo de comparar resultados.

Palavras-chave: IMDB. Mineração de dados. Algoritmos. Estatística. Árvore de decisão. Naïve Bayes.

ABSTRACT

This work aims to carry out an analysis in the IMDB database with the debut titles of films and series between 2015 to 2019 in order to identify patterns and trends. The survey done on the topic showed the lack of articles on this topic in applying the knowledge of Information Management in the entertainment area and was motivated to carry out the present study. The method carried out in this research is the KDD (Knowledge Discovery in Databases), performing data selection, pre-processing, transformation, data mining and information evaluation. First, to carry out the analysis of the database, the delimitation of what would be used and analyzed in the database was made and thus selected which IMDB files would be used, after making the selection the two files were unified and a single database was generated. After the database was generated, data cleaning and normalization were performed in order to be able to analyze with statistics and data mining. For the realization of the statistics, it was first identified that most of the attributes are qualitative and Excel was used as a tool. In data mining the Decision Tree and Naïve Bayes algorithms were selected, both executed in RStudio. The execution of the Decision Tree and Naïve Bayes in the database showed that both had a hit rate of approximately 70%. The work ends with the objectives achieved and with ideas of using other tools, algorithms in order to compare results.

Keywords: IMDB. Data mining. Algorithms. Statistic. Decision tree. Naïve Bayes.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE GRÁFICOS

SUMÁRIO

1 INTRODUÇÃO

Desde as primeiras utilizações e criações de bases de dados e a sua crescente disponibilidade e onipresença, de acordo com Rodrigues e Blattmann (2014), percebe-se que as empresas e as organizações realizam processos de busca por informação e conhecimento com o suporte de tecnologias, e que a orientação de uma organização ao conhecimento pode aumentar sua eficiência por meio da transformação dos dados e informações em vantagens competitivas.

Geralmente empresas como Netflix™ e a Marvel™ Studios, produtoras de filmes, séries e entre outros, se baseiam na análise de dados e algoritmos de recomendação. Onde encontram desde os roteiros, personagens, trailers e até mesmo como as imagens de cada série que aparecem são feitas com base nestas análises para realizar as tomadas de decisões. Os projetos de séries e filmes feitos por essas empresas são escritos por um algoritmo de análise de dados e machine learning que escrevem um roteiro bruto com tudo que aquele determinado público quer assistir.

O Internet Movie Database (IMDB®) de acordo com site Canaltech®, foi criada em 1990 por Col Needham, no Reino Unido, é uma das maiores referências quando se fala sobre filmes, séries, TV e celebridades, e foi vendida para a Amazon™ em 1998.

De acordo com Avancha, Kallurkar e Kamdar (2001) o IMDB possui uma base de dados que contém dados sobre filmes desde os mais antigos até os mais atuais. O repositório iniciou como um conjunto de arquivos de dados os quais disponibilizam informações como classificação de filmes, diretores e entre outros.

A base de dados criada pelo IMDB mostra os dados de filmes, séries e as avaliações que os usuários atribuem. E o presente trabalho busca viabilizar a aplicação de métodos e técnicas de análise de dados com a finalidade de descoberta de padrões e tendências para, por exemplo, identificar os melhores filmes e séries por ano relacionado com gênero e diretores.

1.1 PROBLEMATIZAÇÃO

De acordo com Castro e Ferrari (2016, p. 33), uma das consequências dos avanços da tecnologia, é o problema de superabundância de dados, onde a capacidade de coletar e armazenar dados tem superado a habilidade de analisar e extrair o conhecimento. Desta forma se tornou necessário a aplicação de técnicas e ferramentas que transformem os dados coletados em informações úteis para a tomada de decisão estratégica e até no dia a dia das pessoas.

E ainda Castro e Ferrari (2016, p. 33) afirmam que o crescimento de usuários de internet, de artigos publicados na rede, tuítes diários e vídeos assistidos no Youtube© seguem um crescimento exponencial e que por conta disso vivemos em tempos exponenciais onde encontrar e acessar fontes de informação, pessoas, produtos e serviços não é mais nenhuma dificuldade, mas sim gerenciar, armazenar, processar e extrair conhecimento a partir dessa quantidade quase ilimitada de dados.

Por conta disso, a classificação de filmes pelo mundo é algo que aconteceu de forma exponencial e foi alavancado com o surgimento da internet, e com o aumento das avaliações das pessoas para os filmes e séries de TV, acabou gerando uma grande quantidade de dados espalhados pela rede.

O IMDB e suas avaliações tratam de juntar todos esses dados e transformam em um site com um vasto volume de dados, com uma grande coleção de filmes e séries catalogadas e com suas devidas notas, tanto dos usuários quanto às notas gerais do próprio site.

Com a criação do site do IMDB, notou-se uma grande quantidade de volumes de dados, viu-se nesse fato a oportunidade de realizar uma análise de dados para detectar padrões. A temática da presente monografia pretende responder a seguinte questão sobre a classificação de filmes: como prever padrões e tendências das avaliações de filmes e séries da IMDB com a aplicação de métodos de análise de dados?

1.2 OBJETIVO

Através do estabelecimento dos objetivos, com o intuito de se obter respostas para o problema de pesquisa, foram definidos o objetivo geral e específicos.

O objetivo geral do trabalho concentra-se em aplicar métodos de análise de dados em uma base de dados da Internet Movie Database (IMDB) para identificação de tendências e padrões.

Derivados do objetivo geral, os objetivos específicos são:

a) levantamento bibliográfico e estudos sobre o tema;
b) construir uma base de dados composta por um conjunto de arquivos do IMDB;
c) analisar a base de dados e selecionar os atributos e instâncias a serem analisados;
d) analisar as estatísticas descritivas e definir os algoritmos de mineração de dados para análise da base;
e) aplicar os algoritmos escolhidos;
f) analisar os resultados.

1.3 JUSTIFICATIVA

Com o crescimento do mercado global de mídia e entretenimento, e graças a internet tivemos o fácil acesso a esses produtos e podermos realizar críticas sobre os filmes e séries em geral. Além da possibilidade de postar em redes sociais e portais especializados em entretenimento. Um desses portais é o Internet Movie Database (IMDB) que possui um agregador de notas de usuários em relação aos títulos lançados. E através deste portal, observou-se a possibilidade de se realizar um estudo por meio de análise de dados e, portanto, este trabalho visa identificar padrões em base do IMDB.

Dessa forma, realizou-se um levantamento no dia 15 de janeiro de 2021 nas bases bibliográficas, no portal da Universidade Federal do Paraná (EBSCOHOST) e no portal da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) de modo a se fazer uma verificação de pesquisas semelhantes ao tema considerando o acervo bibliográfico.

Para a realização da pesquisa dois trabalhos foram conduzidos: o primeiro utilizou o parâmetro "IMDB" e “analysis" tanto na EBSCOHOST e na CAPES, sendo que no portal da Capes foi utilizado filtros para refinar a busca como mostra a Figura 1.

FIGURA 1 - REFINAMENTO DE BUSCA 1

FONTE: O AUTOR (2021)

Através da Figura 1 observa-se que o refinamento de busca realizado incluiu os seguintes tópicos: Computer Science, Algorithms, Data Mining, Machine Learning, Classification, Film, Internet, Artificial Intelligence, Library & Information Science e Deep Learning.

No Quadro 1 apresenta-se o total de publicações feitas em cada portal e o total de artigos selecionados para o desenvolvimento desta pesquisa.

QUADRO 1 - PESQUISA DE LEVANTAMENTO BIBLIOGRÁFICO 1

Portal bibliográfico	Total de registros	Total de documentos relacionados com o tema
EBSCOHOST	151	10
CAPES	551	17

FONTE: O AUTOR (2021)

Em ambos os portais foram verificados que dos 151 registros do EBSCOHOST e dos 551 da CAPES somente 27 artigos se aproximaram em relação a temática deste estudo, o que comprova a falta de assunto sobre o tema, mostrando a relevância deste estudo.

Foi realizada uma segunda busca para reforçar a relevância da temática desta pesquisa, utilizando os parâmetros “IMDB” AND ”database” AND ”analysis”, e novamente no portal da CAPES foi utilizado filtros para refinar a busca como mostra a Figura 2.

FIGURA 2 - REFINAMENTO DE BUSCA 2

FONTE: O AUTOR (2021)

Na Figura 2 observamos que o refinamento de busca realizado incluiu os mesmos tópicos da pesquisa anterior que são: Computer Science, Algorithms, Data Mining, Machine Learning, Classification, Film, Internet, Artificial Intelligence, Library & Information Science e Deep Learning.

No Quadro 2 apresenta o total de publicações feitas em cada portal e o total de artigos selecionados para o desenvolvimento desta pesquisa.

QUADRO 2 - PESQUISA DE LEVANTAMENTO BIBLIOGRÁFICO 2

Portal bibliográfico	Total de registros	Total de documentos relacionados com o tema
EBSCOHOST	49	3
CAPES	351	3

FONTE: O AUTOR (2021)

A segunda pesquisa traz um resultado menor que a anterior, ela apresenta apenas três artigos relacionados com o tema em relação aos 49 como resultado da EBSCOHOST e da CAPES de 351 artigos, apenas 3 eram relacionados com o tema, ressaltando a relevância em se realizar este presente estudo.

Além do escasso estudo sobre o tema como já visto anteriormente, a ideia de fazer essa pesquisa, deu-se por conta do interesse do autor em filmes e seriados e avaliar o conteúdo depois de assisti-los. A descoberta de uma base de dados disponível para a realização de análises surgiu a motivação de realizar este estudo. A partir da disponibilização desta base de dados veio um dos objetivos em utilizar métodos e ferramentas a fim de analisar os dados identificando possíveis padrões e tendências observadas.

Com relação ao Curso de gestão da informação da UFPR, uma das justificativas vão em direção primeiramente à contribuição deste trabalho para o acervo da biblioteca. Realizou-se uma busca bibliográfica sobre o tema no portal e não foi encontrado nenhuma referência relacionada. E uma segunda contribuição está à possibilidade de aplicar os conhecimentos adquiridos através de várias disciplinas no curso de forma prática no presente trabalho.

1.4 DELIMITAÇÃO DA PESQUISA

A presente pesquisa abrange a base de dados fornecida pela Internet Movie Database (IMDB). A base em questão, compreende dados de filmes, séries de TV,

curtas e outros, cada um com os diretores responsáveis, atores e notas atribuídas (relacionada aos votos) pelos usuários do IMBD.

O objetivo desta pesquisa não é analisar o elenco, diretores e escritores, mas sim a relação das notas com o tipo de título, sendo filmes e séries e seus gêneros, portanto nem todos os arquivos disponíveis no IMDB serão utilizados.

1.5 ESTRUTURA DO DOCUMENTO

O presente trabalho está organizado em 5 (cinco) seções, além desta que contempla problematização, objetivo, justificativa, delimitação da pesquisa e estrutura do documento.

A segunda seção, compõe a revisão de literatura, apresentando os principais conceitos relacionados à pesquisa, quais sejam: gestão da informação, recuperação da informação, tomada de decisão, descoberta de conhecimento em bases de dados com o KDD e história do cinema com a subcategoria Internet Movie Database.

A terceira seção apresenta a metodologia com a Caracterização da pesquisa, ambiente de pesquisa, Dicionário de dados, materiais e métodos e os processos do KDD sendo (pré-processamento, estatística, mineração de dados e pós processamento).

A quarta seção é formada pelos experimentos e análise dos resultados, neste capítulo são realizados o pré-processamento, tratamento da base de dados, estatísticas descritivas, Mineração de dados e o pós processamento

A quinta e última parte contém as considerações finais, a qual contempla o alcance dos objetivos e trabalhos futuros.

2 REVISÃO DE LITERATURA

Esta seção apresenta os principais conceitos relacionados à pesquisa, tais como: gestão da informação, recuperação da informação, tomada de decisão, os processos do KDD e o referencial teórico sobre o IMDB e história do cinema.

2.1 GESTÃO DA INFORMAÇÃO

A Gestão da Informação segundo Nóbrega, Bastos e Araújo (2014, p. 21) se apresenta como um conjunto de atividades que reproduz a forma pela qual uma organização captura, entrega e utiliza a informação. Os autores afirmam também que a gestão da informação é um conjunto de etapas que realiza a produção de dados/informação, organização, armazenamento, recuperação, acesso e uso.

Para se entender a Gestão da Informação, deve-se compreender os conceitos básicos utilizados na GI. De acordo com o autor Moreira (2014, p. 16) os dados são fatos em forma primária e que se organizados ou arranjados se transformam numa informação, e a informação são dados organizados de forma a possuir valor.

E em seguida o Moreira (2014, p. 16) apresenta uma tabela diferenciando os dados, a informação e o conhecimento.

QUADRO 3 - DIFERENÇA ENTRE DADOS, INFORMAÇÃO E CONHECIMENTO

Dados	Informação	Conhecimento
Simples observações sobre o estado do mundo Facilmente estruturado Facilmente obtido por máquinas Frequentemente quantificado Facilmente transferível	Dados dotados de relevância e propósito Requer unidade de análise Exige consenso em relação ao significado Exige necessariamente a mediação humana	Informação valiosa da mente humana Inclui reflexão, síntese, contexto De difícil estruturação De difícil captura em máquinas Frequentemente tácito De difícil transferência

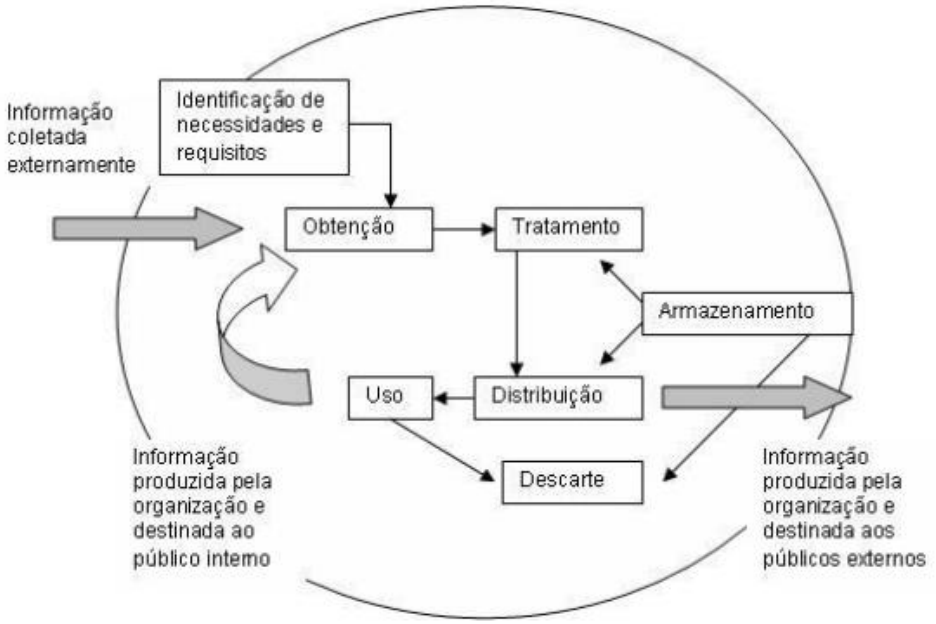
FONTE: DAVENPORT (1998, p. 18).

Como podemos ver no Quadro 3, dados são observações simples sobre algo, informação são dados com relevância e propósito, e o conhecimento é a informação contida na mente humana. Davenport (1998, p. 18) explica que a informação é difícil de ser definida pois ela é imprecisa e que envolve além dela mesma, mas também os dados e o conhecimento e que servem de conexão entre os dados e o conhecimento.

De acordo com Nóbrega, Bastos e Araújo (2014, p. 21), os problemas informacionais surgem em qualquer etapa do fluxo, a necessidade de entender o fluxo informacional além de seus contextos organizacionais, suas condições de produção e seu uso das necessidades de informação do usuário é essencial para conseguir identificá-los.

Na Figura 3, é possível observar o modelo fluxo de informação onde é identificada a necessidade dessa informação como ilustra Beal (2004).

FIGURA 3 - MODELO DE FLUXO DA INFORMAÇÃO



FONTE: BEAL (2004)

O modelo de fluxo da informação, segundo Beal, consiste na identificação das necessidades e requisitos de informação. E é realizado a partir de etapas que começam pela identificação de necessidades e requisitos de informação, onde, segundo o autor, “é fundamental para que possam ser desenvolvidos produtos

informacionais orientados especificamente para cada grupo e necessidade" (BEAL, 2004, p. 30).

A obtenção das informações que Beal (2004, p. 30) descreve no próximo passo, após a identificação de necessidades, "são desenvolvidas as atividades de criação, recepção e captura de informação, provenientes de fonte externa ou interna, em qualquer mídia ou formato" (BEAL, 2004, p. 30).

O tratamento de informação que sucede a obtenção da informação, é definido pelo autor que "antes de estar em condições de ser aproveitada, é comum a informação precisar passar por processos [...] com o propósito de torná-la mais acessível e fácil de localizar pelos usuários" (BEAL, 2004, p. 30).

A distribuição da informação, é onde a informação será conduzida ao usuário que necessita dessa informação e está ligada diretamente ao uso da informação que compõe a quinta etapa, que para o autor é a "O mais importante de todo o processo de gestão da informação, embora seja frequentemente ignorada pelas organizações" (BEAL, 2004, p. 31).

O armazenamento é a sexta etapa do processo, e para Beal (2004, p. 31) é aqui ficam armazenado os dados e informações de modo a estar sempre disponível para seu uso e reuso dentro da organização e sempre se manter organizada.

A última etapa do modelo consiste no seu descarte da informação que segundo Beal "excluir dos repositórios de informação corporativos os dados e informações inúteis melhora o processo de gestão da informação" (BEAL, 2004, p. 34).

Na Gestão da Informação é sempre necessário realizar o estabelecimento de processos e etapas, como mostra a Figura 4 de Davenport.

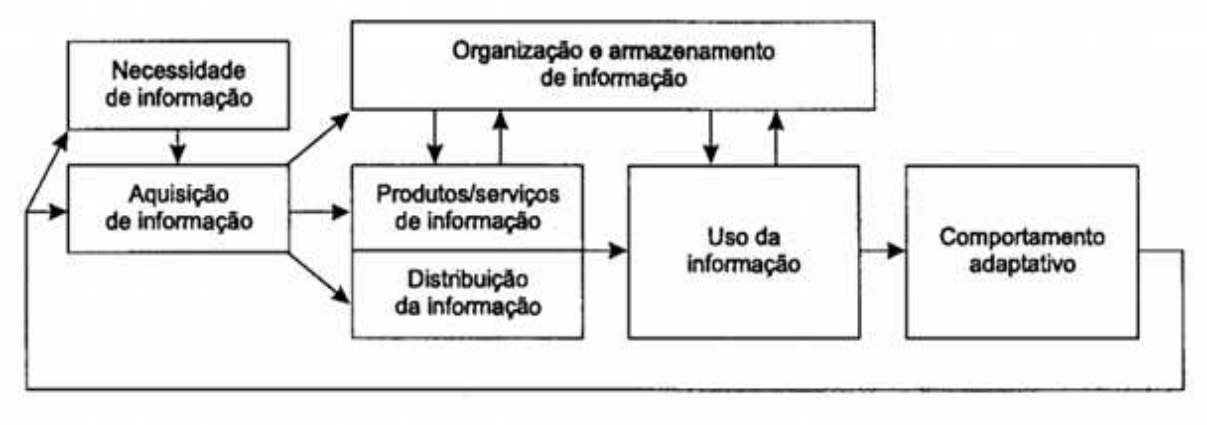
FIGURA 4 - PROCESSO DE GERENCIAMENTO DA INFORMAÇÃO

FONTE: DAVENPORT (1998, p. 175)

Como visto na Figura 4, o primeiro passo consiste em determinar as exigências da informação é um fator problemático, pois demanda que as pessoas da organização percebam o ambiente que cerca as informações. O segundo passo, a obtenção de informações, identifica-se a importância de buscar as informações necessárias. No terceiro passo, do processo de gerenciamento da informação se refere à forma pelas quais as informações são buscadas e comunicadas internamente. E o último passo mostrado, é o uso da informação em que Davenport (1998, p. 194) explica que a maneira como um funcionário processa a informação para tomar uma decisão vai depender de sua própria experiência.

Na Figura 5, Choo (2003) apresenta o ciclo de gestão da informação, esta figura explica cada processo realizado na gestão da informação.

FIGURA 5 - CICLO DE GESTÃO DA INFORMAÇÃO



FONTE: CHOO (2003, p. 404)

A Figura 5 mostra cada etapa do ciclo da Gestão da Informação, e de acordo com Choo (2003, p. 405) a necessidade de informação surge de problemas, incertezas e ambiguidades encontradas em situações e experiências específicas, de modo que não deve haver preocupação com o significado da informação, mas sim com as condições, padrões e regras de uso, que a torna significativa para determinados indivíduos em determinadas situações.

Na aquisição da informação segundo Choo (2003, p. 407), o usuário possui reflexão sobre a extensão e a diversidade de suas preocupações com os acontecimentos e mudanças do ambiente externo. Uma das demandas sugere que as fontes utilizadas para monitorar o ambiente sejam suficientemente numerosas e

variadas para refletir todo o espectro de fenômenos externos. O autor ainda mostra que uma maneira eficaz de administrar a variedade de informações é envolver o maior número possível de pessoas na coleta de informações e que apesar dessa sugestão, deva-se controlar e monitorar essa variedade de fontes. E a seleção e uso das fontes de informação devem ser planejados como qualquer outro recurso vital para a organização.

A organização e armazenamento da informação de acordo com Choo (2003, p. 409) mostra que toda a informação que é adquirida ou criada, é armazenada em arquivos e bancos de dados ou sistemas de informação, de modo a facilitar a sua recuperação e compartilhamento. Choo (2003, p. 409) afirma que toda a informação armazenada representa sua importância na organização e é frequentemente consultado pela organização.

Em produtos e serviços de informação, Choo (2003, p. 412) afirma que produtos e serviços de informação têm como função de garantir que as necessidades de informação dos membros da organização sejam atendidas com uma mistura equilibrada de produtos e serviços. E seus resultados se mostram pela abrangência das circunstâncias específicas que afetam a resolução de cada problema ou cada tipo de problema.

A distribuição da informação de acordo com o autor é o processo pelo qual as informações se disseminam pela organização, de maneira que "a informação correta atinja a pessoa certa no momento, lugar e formato adequados" (CHOO, 2003, p. 414). E que seu objetivo é promover e facilitar a partilha de informações, que é fundamental para a criação de significado, a construção de conhecimento e a tomada de decisões.

O uso da informação é explicado por Choo (2003, p. 415) como um processo social dinâmico de pesquisa e construção que resulta na criação de significado, na construção de conhecimento e na seleção de padrões de ação.

De acordo com Assis (2008, p. 141) um dos objetivos na implantação da Gestão da Informação (GI) em uma organização é garantir que a informação seja administrada como recurso indispensável e valioso. Deve-se ter em mente que o foco do negócio vem em primeiro lugar respeitando o estilo de gestão e a cultura da organização.

De acordo com Assis (2008, p. 141) existem dez pontos importantes para realizar a implementação da gestão da informação nas organizações:

> A gestão da informação deve estar alinhada com a missão e os objetivos estratégicos; Desenvolver um plano de gestão da informação voltado, preferencialmente, para a perspectiva do negócio; Preocupar-se sempre com a máxima: a informação para as pessoas certas, no local correto, no tempo certo, no formato adequado e, se possível, com custo zero; Ter sempre a visão de que a informação deve ser utilizada no seu potencial máximo; Priorizar a qualidade, a disponibilidade, o uso e o valor da informação; O gestor da informação deve estar ligado diretamente à alta administração; Mapear regularmente as necessidades de informação; Considerar a qualidade das fontes de informação e sua disponibilidade; Permanentemente, analisar o custo x benefício das fontes de informação adquiridas; Contextualizar e compartilhar a informação de interesse. (Assis, 2008, p. 141).

Assis (2008, p. 142) também cita os pontos importantes e fundamentais para a sobrevivência e perenidade que envolvem o assunto de produtos de informação e a gestão da informação como o acervo, tratamento e a recuperação da informação, sistemas para administrar a informação e possuir profissionais capazes de gerenciar essas informações.

2.2 RECUPERAÇÃO DA INFORMAÇÃO

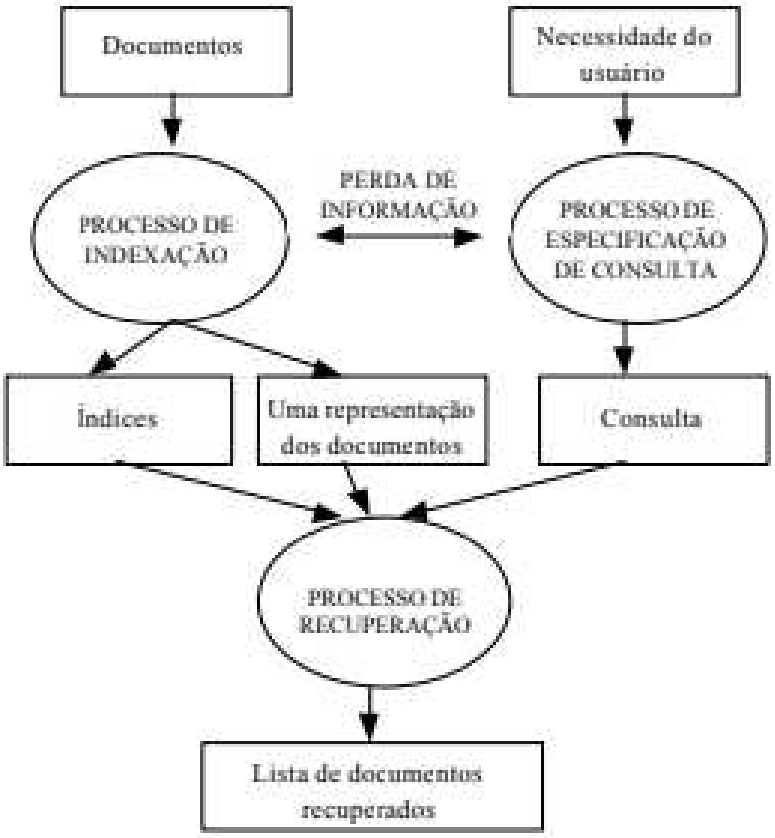
A recuperação de acordo com Assis (2008, p. 143) se dá em decorrência do crescimento e do grande volume de publicações existentes no mundo e por consequência atendendo as necessidades dos usuários, o que exige processos de recuperação cada vez mais desenvolvidos, sendo assim possuindo o objetivo de ter uma gestão da informação que seja capaz de proporcionar uma recuperação eficaz da informação desejada nos documentos existentes no acervo/bancos de dados.

De acordo com Pontes, Carvalho e Azevedo (2013, p. 6) a recuperação da informação é um campo que abrange diversos domínios, desde a ciência da informação até a ciência da computação que possuem ferramentas de organização e recuperação da informação e conhecimento, como classificação, tesauros, taxonomia e ontologias.

A Recuperação da Informação (RI) estuda o armazenamento e recuperação automática de documentos, que são objetos de dados segundo o autor Moreira (2014, p. 16) ele ainda considera a RI como uma subárea da ciência da informação.

Cardoso (2004, p. 1) apresenta um fluxo de níveis hierárquicos organizacionais conforme a Figura 6.

FIGURA 6 - SISTEMA DE RECUPERAÇÃO DA INFORMAÇÃO



FONTE: CARDOSO (SÉC XXI, p.01)

O fluxo de níveis hierárquicos organizacional apresentado na Figura 6 de acordo com Cardoso (2004, p.1) inicia com a inclusão dos documentos e com as necessidades dos usuários seguindo para a fase de processos de indexação e especificação de consulta, através da indexação ele se divide entre índices e representação dos documentos gerando uma lista de arquivos recuperados. A autora complementa que através do grande crescimento informacional tem se ficado cada vez mais complexo recuperar informações e é se exigido mais em processos.

2.3 TOMADA DE DECISÃO

A tomada de decisão de acordo com Bertoncini, Brito, Leme, Silva (Séc. XX, p. 6), pode ser abordada de várias formas, porém pode seguir dois modelos sendo o racional e comportamental. O modelo racional o “decisor” tem informações perfeitas,

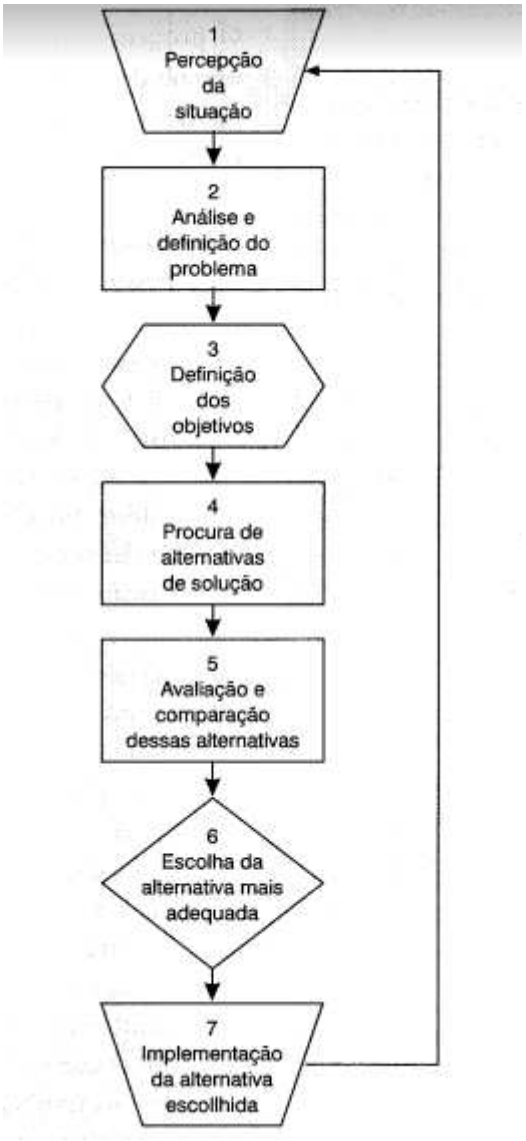
capazes de avaliar sistematicamente e logicamente cada alternativa e tomar melhores decisões para a organização de modo imparcial. Já o modelo comportamental leva em consideração a percepção, experiência, informações e alternativas limitadas quando o gerente terá que tomar decisões.

Chiavenato (2003, p. 347) aborda a origem da tomada de decisão mostrando seu surgimento a partir da Teoria das Decisões que nasceu com Herbert Simon onde utilizou como base para explicar o comportamento humano nas organizações.

Tal teoria tem como concepção de organização um sistema de decisões onde cada pessoa participa racionalmente e conscientemente, escolhendo e tomando decisões individuais a respeito de alternativas racionais de comportamento e assim mostrando que a organização está repleta de decisões e de ações. Para Chiavenato “a teoria comportamental não é somente o administrador quem toma as decisões, mas sim, todas as pessoas na organização, de todas as áreas de atividade e níveis hierárquicos” (CHIAVENATO, 2003, p. 347).

No processo decisório, de acordo com Chiavenato (2003, p. 348), o tomador de decisões depende de suas características pessoais e a maneira em que ele percebe cada situação. Na Figura 7 é apresentado o processo decisório sob a ótica da organização:

FIGURA 7 - PROCESSO DECISÓRIO SOB A ÓTICA DA ORGANIZAÇÃO 1



FONTE: CHIAVENATO (2003, p. 350)

Na Figura 7, o autor mostra que o processo decisório dentro da organização pode possuir várias decorrências, tais como como uma racionalidade limitada, imperfeição na decisão, relatividade nas decisões, hierarquização das decisões, racionalidade administrativa e a Influência organizacional. Cada uma das etapas do processo decisório possui uma influência as outras etapas, porém nem sempre é necessário seguir exatamente todas as etapas, as etapas 3,5 e 7 Chiavenato (2003, p.349) destaca que podem ser abreviadas caso exista necessidade de solução imediata.

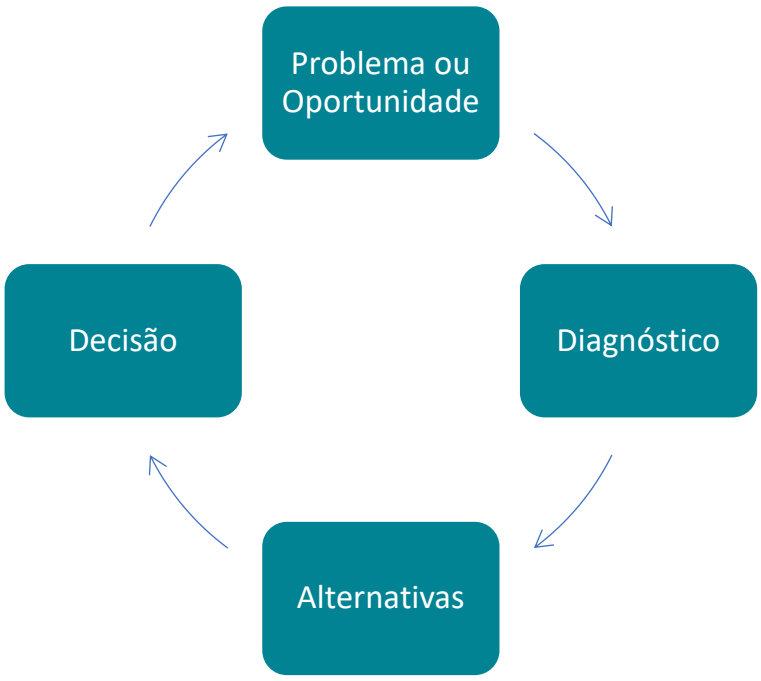
Já segundo Maximiano (2000, p. 139), o processo de tomada de decisão é muito discutido em trabalho de administração e destaca que:

> “Decisões são escolhas que as pessoas fazem para enfrentar problemas e aproveitar oportunidades. Tomar decisões para enfrentar problemas e aproveitar oportunidades é um ingrediente importante do trabalho de administrar. Muito do que os gerentes fazem é resolver problemas e enfrentar outros tipos de situações que exigem escolhas” (Maximiano, 2000, p.139).

Portanto Maximiano (2000, p. 139) conclui que é realizado a tomada de decisão a fim de enfrentar problemas e aproveitar melhor suas oportunidades

O processo decisório é realizado através de quatro etapas segundo Maximiano (2000, p. 131) na Figura 8.

FIGURA 8 - PROCESSO DECISÓRIO SOB A ÓTICA DA ORGANIZAÇÃO 2



FONTE: MAXIMIANO (2000, p. 141)

Maximiano (2000, p. 148) descreve a primeira etapa como uma situação de frustação, desafio ou interesse sobre o problema. Na segunda etapa, são feitas a análise e compreensão do problema, a terceira etapa é a busca por criar alternativas para resolver o problema e a última etapa é feita a avaliação e escolha das alternativas segundo a Figura 8.

O processo decisório se inicia na identificação das escolhas das soluções, quando a decisão é definida o ciclo fecha, enquanto não se define ele passa sempre

pelo ciclo. Quando a decisão é colocada em prática gera outras decisões a fim de resolver os problemas voltando novamente para o ciclo segundo Maximiano (2000, p. 141).

Maximiano (2000, p. 148) apresenta que a identificação do problema é o momento no qual a principal situação é de frustração curiosidade ou aquela vontade de ir ao um desafio. Seguindo para o diagnóstico dessa fase, é realizada a análise do problema tentando compreender a situação. A etapa de geração de alternativas tenta criar soluções para os problemas. E por fim, a última etapa a decisão depois de passar por todas as etapas anteriores realiza-se uma avaliação e julgado se foi a melhor decisão entre as alternativas.

2.4 KDD DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

De acordo com Castro e Ferrari (2016, p. 36) o KDD (*Knowledge Discovery in Database*) ou traduzido como “descoberta de conhecimento em bases de dados” se refere a todo processo de extração de conhecimentos a partir de bases de dados. O processo consiste de quatro fases sendo: a obtenção dos dados, pré processamento, mineração de dados e o pós-processamento como ilustrado na Figura 9.

FIGURA 9 - PROCESSO KDD

FONTE: Adaptado SILVA; PERES; BOSCARIOLI (2016, p. 11)

Através das quatros fases apresentadas por Castro e Ferrari (2016, p. 36) na Figura 9, o KDD em sua primeira fase, a obtenção dos dados, consiste em uma coleção organizada de dados, tanto com valores quantitativos ou qualitativos

referentes a um conjunto de itens onde são os níveis mais básico de abstração a partir do qual a informação e posteriormente os conhecimentos podem ser extraídos.

A segunda fase, o pré-processamento de dados, é a etapa que antecede a mineração de dados que visa realizar a limpeza (dados inconsistentes), integração (combinação de dados obtidos a partir de múltiplas fontes), seleção ou redução (escolha dos dados relevantes à análise) e a transformação (consolidação dos dados em formatos apropriados para a mineração) dos dados para uma análise eficiente e eficaz.

Já a terceira fase do KDD, a mineração de dados, corresponde à aplicação de algoritmos capazes de extrair os conhecimentos a partir dos dados pré-processados. Nessa etapa foram elaboradas as técnicas de análise descritiva, agrupamento, predição, associação e detecção de anomalias.

Por fim, no pós-processamento é realizada a avaliação do conhecimento. Essa etapa visa identificar os conhecimentos úteis e não triviais a partir da avaliação dos resultados.

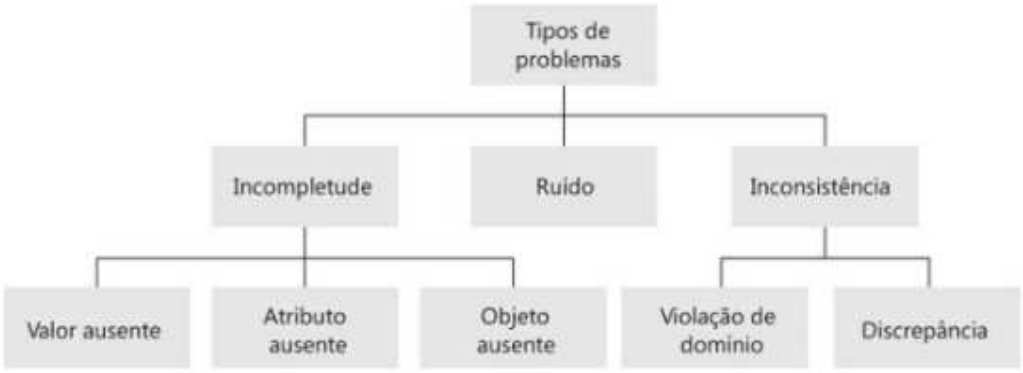
Silva, Peres e Boscarioli (2016, p. 11) apresentam que estes processos do KDD não precisam necessariamente seguir uma sequência, ele pode voltar para uma etapa várias vezes conforme é realizada a aplicação e análise.

2.4.1 Pré-Processamento

De acordo com Neves (2003, p. 30) o pré-processamento engloba a análise inicial dos dados para se extrair definições sólidas como os tipos de dados, formatos, sistema de fonte original, estrutura das tabelas e valores potenciais dos atributos. De acordo com o autor, essa fase também engloba a escolha dos dados relevantes aos objetivos do usuário utilizando a operação necessária e a limpeza e transformação dos dados de forma se tornar possível a mineração de dados.

Através da Figura 10, Castro e Ferrari (2016, p. 65) apresentam os principais problemas com os dados.

FIGURA 10 - PRINCIPAIS PROBLEMAS COM DADOS



FONTE: CASTRO & FERRARI (201 6, p. 67)

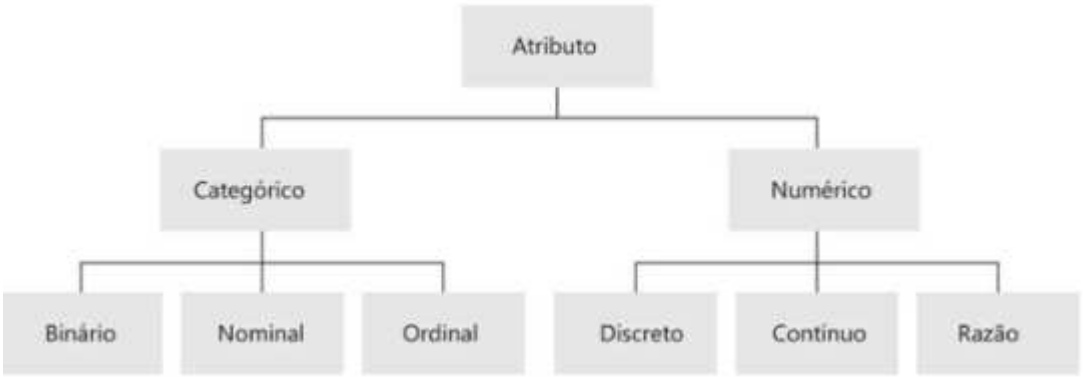
Castro e Ferrari (2016, p. 65) explicam que os dados brutos são aqueles que ainda não foram processados para uso. E deles que podem decorrer três tipos de problemas com os dados, que são: a incompletude, inconsistência de acordo com a Figura 10. A incompletude pode ocorrer de várias formas, como por exemplo a falta de valores de um dado. A inconsistência ocorre quando versões diferentes e conflitantes do mesmo dado aparecem em locais variados, geralmente é aquele cujo valor está fora do domínio do atributo ou apresenta uma grande discrepância em relação aos outros dados. Ruído apresenta vários significados, mas no contexto de mineração de dados é o dado que apresenta uma variação de valor em relação ao seu valor de um dado sem ruído e, portanto, esses ruídos na base de dados podem levar a inconsistências.

De acordo com Neves (2003, p. 31), o que consome a maior parte do tempo na realização do processo do KDD é a tentativa de preparar os dados antes de começar a mineração, e isso deve-se ao fato de que os dados do mundo real podem ser errôneos, incompletos e inconsistentes, e as técnicas de pré-processamento podem melhorar a qualidade do dado e por consequência melhorar a acurácia e eficiência do seu processo de mineração.

Com relação ao processo em si, não há uma sequência obrigatória em relação a subfases de pré-processamento, pois pode-se preferir realizar a limpeza dos dados antes de realizar a seleção, por exemplo. O entendimento dos dados é realizado por especialistas que entendem do que se tratam as tabelas envolvidas, sua relevância, significado, formato, tamanho e tipo de dado dos atributos; identificando os pontos chaves; fazendo os levantamentos estatísticos e verificando a qualidade dos dados fornecidos. Nesta fase de entendimento dos dados é importante entender os atributos da base, como as necessidades do usuário e quando for aplicar os algoritmos de

mineração de dados, pois o pesquisador decidirá os algoritmos depois de ter conhecimento dos atributos porque existem algoritmos que funcionam com atributos categóricos e outros com numéricos, tendo o entendimento da base adianta os processos para dar sequência do KDD. A Figura 11 contém os principais tipos de atributos presentes em bases de dados de acordo com Castro e Ferrari (2016, p. 71).

FIGURA 11 - TIPOS DE ATRIBUTOS



FONTE: CASTRO & FERRARI (2016, p. 71)

Os autores Castro e Ferrari (2016, p. 70), de acordo com a Figura 11, explicam que o atributo pode ser considerado uma entrada ou variável, e pode ter um valor tanto numérico ou categórico, sendo que os valores numéricos podem assumir quaisquer valores numéricos e possuem os atributos discreto, contínuo e razão, já os valores categóricos podem assumir símbolos distintos e possuir atributos binários, nominal e ordinal.

Neves (2003, p. 32) explica que a seleção de dados é a etapa em que ocorre a escolha da(s) tabela(s), atributos e instâncias dela(s) em relação aos objetivos do usuário.

A limpeza dos dados garante a qualidade dos dados utilizando algumas operações como a padronização de dados, tratamento de valores ausentes, eliminação de dados errôneos e de duplicatas segundo (NEVES 2003, p. 32).

Já a transformação de dados, torna a apresentação dos dados de uma forma apropriada para as técnicas de mineração de dados que pode a ser utilizada, onde irá utilizar as operações do tipo normalização de dados, conversões de valores

simbólicos para valores numéricos, discretização e composição de atributos de acordo com Neves (2003, p. 32).

2.4.2 Mineração de Dados

De acordo com Amaral (2016, p. 2) a mineração de dados pode ser definido como processos para explorar e analisar grandes volumes de dados em busca de padrões, previsões, erros, associações e outros tipos de análises. E a mineração de dados está alinhada com o aprendizado de máquina, que é uma área da inteligência artificial onde se desenvolvem algoritmos capazes de fazer com que o computador aprenda utilizando dados de eventos anteriores.

Castro e Ferrari (2016, p. 39), descrevem a mineração de dados como uma área interdisciplinar e multidisciplinar que envolve várias áreas conforme apresentado na Figura 12.

FIGURA 12 - MULTIDISCIPLINARIDADE DA MINERAÇÃO DE DADOS

FONTE: (CASTRO; FERRARI, 2016, P. 39).

Na Figura 12 as áreas envolvidas na mineração de dados e que a multidisciplinaridade da mineração de dados para Castro e Ferrari (2016, p. 39) servem para especificar os tipos de informações a serem obtidas nas tarefas de mineração.

De acordo com Castro e Ferrari (2016, p. 40), essas tarefas de mineração podem ser classificadas em duas categorias: as descritivas e preditivas, onde as descritivas caracterizam as propriedades gerais dos dados e as preditivas fazem inferência a partir dos dados objetivando predições. Na técnica descritiva é possível medir, explorar e descrever características intrínsecas dos dados. Basicamente, essas técnicas permitem investigar a distribuição de frequência, as medidas de centro e variação e as medidas de posição relativa e associação dos dados. Além disso, as técnicas descritivas permitem a sumarização e compreensão dos objetos da base e seus atributos.

Já a predição de acordo com Castro e Ferrari (2016, p. 41) é uma terminologia que se refere à construção e ao uso de modelo para avaliar a classe de um objeto não rotulado ou para estimar o valor de um ou mais atributos de dado objeto.

O Quadro 4 demonstra de forma organizada as técnicas de mineração de dados.

QUADRO 4 - TÉCNICAS DE MINERAÇÃO DE DADOS

Técnicas	***Heurística***	***Descrição***
Descritivas	Agrupamento	O objetivo é criar grupos e atribuir instâncias a estes grupos a partir das características ou atributos destas instâncias.
	Associação	As regras de associação são utilizadas para identificar causas ou conjunto de fatores que levam a ocorrência de um fato ou fenômeno.
	Detecção de Anomalias	Ajuda na identificação de anomalia os dados, removendo dados discrepantes que não

		enquadram no comportamento dos demais registros.
	Sumarização	É a generalização dos dados, ela extrai informações condensadas de uma massa de volume de dados e fornece um resultado com um conjunto menor, tendo uma visão geral dos dados.
Preditivas	Classificação	A classificação tem como objetivo de usar todos os atributos que compõem a relação para tentar prever a classe. Ou seja, descrever ou prever um atributo especial chamado classe.
	Regressão	A regressão tem como classe do tipo numérica e é utilizado como por exemplo para prever altura de uma pessoa a partir do peso.

FONTE: O AUTOR (2020)

Como visto no Quadro 4, cada técnica possui suas heurísticas que, por sua vez, podem atender um caso em específico e que dentro delas há vários algoritmos. A técnica descritiva tem como heurística o agrupamento, associação, detecção de anomalias e sumarização. Já a técnica de predição possui as heurísticas de Classificação e regressão.

2.4.3 Pós Processamento

De acordo com Goldschmidt e Passos (2005, p. 55) a fase de pós-processamento envolve a visualização, a análise e a interpretação do resultado gerado pela etapa de mineração de dados, esta fase é a etapa na qual será avaliado os resultados obtidos do KDD.

Esta etapa de forma em geral analisa todos os resultados e interpreta os mesmos, este processo auxilia no conhecimento e apresentação dos resultados segundo Boente, Oliveira e Rosa (s.d., p. 9).

2.5 HISTÓRIA DO CINEMA

O cinema, de acordo com Mascarello (2006, p. 17), começou por volta de 1895 e não possuía um código próprio e estava misturado a outras formas de entretenimento, como os espetáculos de lanterna mágica, o teatro popular, os cartuns, as revistas ilustradas e os cartões-postais. Esteve em transformação constante entre 1895 a 1915, até que conseguir uma estabilidade em 1915 que caracterizava o cinema hollywoodiano clássico e o início da televisão nos anos de 1950.

De acordo com o autor, não existiu um único descobridor do cinema e que os aparatos que eram envolvidos na invenção não surgiram repentinamente num único lugar, mas sim uma conjunção de circunstâncias técnicas aconteceu quando vários inventores mostraram os resultados de suas pesquisas em relação a projeção de imagens em movimento.

3 METODOLOGIA

Será apresentada neste capítulo a caracterização da pesquisa, ambiente de pesquisa, dicionário de dados, materiais e métodos e os processos do KDD. O método utilizado para a análise da base de dados do IMDB se baseia nos processos da KDD onde foi realizado o pré-processamento, estatística, mineração de dados e pós-processamento.

3.1 CARACTERIZAÇÃO DA PESQUISA

A caracterização da pesquisa de acordo com Gil (2008) é realizada com base nos métodos e técnicas destacadas sob o ponto de vista de sua finalidade, abordagem, objetivos e procedimentos.

Gil (2008, p. 26) define que a pesquisa é um processo formal e sistemático de desenvolvimento do método científico, com o objetivo fundamental de obter respostas utilizando o método científico.

A finalidade da pesquisa de acordo com o autor é descrita como pesquisa pura ou aplicada. Para este trabalho a finalidade da pesquisa é aplicada, a qual apresenta pontos de contato, com o objetivo de responder questões específicas buscando assim soluções concretas e resultados, enriquecendo o trabalho através dos conhecimentos.

Existem três níveis de pesquisa de acordo com Gil (2008, p. 27) que são: as pesquisas descritivas, explicativa e a exploratória. A selecionada para este trabalho foi a pesquisa descritiva que possui o objetivo de descrever determinadas características de uma determinada população, fenômeno ou estabelecimento de relações variáveis, tem como a principal finalidade de observar, registrar e analisar os fenômenos ou sistemas técnicos, sem entrar no mérito dos conteúdos.

Dos oitos procedimento técnicos mostrados por Silva e Menezes (2005, p.21), foi selecionada a pesquisa experimental que determina um objeto de estudo, seleciona variáveis que possam influenciar, definem as formas de controle e verificam os efeitos da variável sobre o objeto.

Seguindo essa classificação de pesquisa, foram definidas as classificações da pesquisa conforme o Quadro 5, destacando de cor laranja aquelas utilizadas na realização deste trabalho.

QUADRO 5 - CLASSIFICAÇÃO DAS PESQUISAS

Finalidade de Pesquisa	Objetivos	Procedimentos técnicos
Pura	***Descritiva***	Pesquisa Bibliográfica
Aplicada	Explicativa	Pesquisa Documental
	Exploratória	***Pesquisa Experimental***
		Levantamento
		Estudo de Caso
		Pesquisa Expost-Facto
		Pesquisa Ação

Pesquisa Participante

FONTE: O AUTOR (2020)

No Quadro 5 é mostrada a classificação das pesquisas neste trabalho, selecionado a finalidades como do tipo aplicada, o objetivo como descritivo e o procedimento como experimental.

3.2 AMBIENTE DE PESQUISA

O estudo desse projeto foi realizado por meio das bases de dados da IMDB no seguimento de entretenimento.

O Internet Movie Database (IMDB) de acordo com Topal e Ozsoyoglu (2016, p. 1) é a fonte de conteúdo mais popular do mundo para conteúdo sobre filmes, TV e celebridades com resenhas de mais de 3,5 milhões de filmes, onde os membros do IMDB fornecem uma análise e nota sobre os filmes e séries em geral. A nota geral dos filmes é calculada pelo algoritmo de classificação do IMDB que utiliza a pontuação dos usuários que analisaram.

De acordo com Chmielewski, Dawn C. (2013), Col Needham é o criador, fundador e chefe executivo do Internet Movie Database, a database foi criada em 17 de outubro de 1997 e foi incorporado a World Wide Web em 1996.

3.4 DICIONÁRIO DE DADOS

As bases de dados do IMDB estão disponibilizadas no site[1] do IMDB e são bases separadas, através da junção dos arquivos obteve-se uma base com 6.629.737 títulos do ano de 1874 até o dezembro 2019 e 10 atributos. A base foi convertida para o formato de .CSV seus principais atributos estão descritos no Quadro 6 separados pelo tipo de atributo.

[1] Fonte: https://datasets.imdbws.com/

QUADRO 6 - QUADRO DE ATRIBUTOS DA BASE DE DADOS DO IMDB

Atributo	Valor do atributo	Tipo de atributo	Descrição
tconst	Variável Ex: tt0000001	Alfanumérico	Identificador único alfanumérico do título
titleType	Variável: Ex: short	Nominal	Título criado durante o projeto do filme
primaryTitle	Variável: Ex: Miss Jerry	Nominal	Título conhecido comercialmente
isAdult	Variável: 0 ou 1	Binário	Se o conteúdo é adulto ou não
startYear	Variável Ex: 2010	Inteiro	Representa o ano de lançamento de um título. No caso da série de TV, é o ano de início da série
endYear	Variável Ex: 2010	Inteiro	Série de TV no ano final. "\ N" para todos os outros tipos de título
runtimeMinutes	Variável: Ex: 1	Inteiro	Tempo de execução principal do título em minutos
genres	Variável Ex: Comedy	Nominal	Inclui até três gêneros associados ao título
averageRating	Variável Ex: 5.8	Inteiro	Média ponderada de todas as classificações de usuários individuais (Nota de 1.0 a 10.0)

numVotes	Variável Ex: 1487	Inteiro	Número de votos que o título recebeu

FONTE: IMDB (2020)

No Quadro 6 encontra-se descrito cada atributo e é possível analisar um exemplo de cada atributo, o tipo de dados e a descrição dos mesmos.

3.5 MATERIAIS E MÉTODOS

Para a realização do presente trabalho, as ferramentas utilizadas para o estudo foram o Excel, Rstudio (Versão 1.3.1093) e o Word, assim como descritos em específico no capítulo da metodologia. Em relação ao método utilizado todo o presente trabalho foi aplicado em cima da metodologia do KDD explicada no capítulo 2.4 KDD Descoberta de Conhecimento em Bases de Dados.

Um método que consiste na análise da base dos dados classificadas em atividades como obtenção dos dados para análise, pré-processamento, mineração de dados e o pós processamento.

3.6 PRÉ- PROCESSAMENTO

O pré-processamento neste trabalho foi realizado na limpeza dos dados e no tratamento da base de dados. Para a realização da limpeza e tratamento dos dados foi necessário determinar e escolher quais arquivos seriam utilizados e juntar os arquivos da base de dados. Conforme descrito no capítulo 3.4 Dicionário de dados, a base de dados contém 10 atributos e 6.629.737 títulos registros do ano de 1874 até dezembro de 2019.

3.6.1 Limpeza dos dados

De acordo com Silva, Peres e Boscaroli (2016, p. 42), para a limpeza de dados, são utilizadas quatro análises de validações de dados, os valores ausentes, valores ruidosos, valores inconsistentes e valores redundantes.

Os autores Silva, Peres e Boscaroli (2016) esclarecem que os valores ausentes podem ser identificados de duas formas: quando os atributos de um conjunto de dados não apresentam valores para determinados exemplares, ou quando um conjunto de dados não possui valores para um atributo de interesses ou aparenta valores agregados em relação aquele tributo. Já os valores ruidosos se referem às modificações dos valores originais e que consistem em valores diferentes da maioria dos outros valores do conjunto de dados, também chamados de *outliers*. Os valores inconsistentes são aqueles que apresentam discrepâncias tanto de tipo ou de domínio. E por último, os valores redundantes ocorrem devido a três fatores: uso de nomenclaturas diferentes para atributos equivalentes, inserção de exemplares repetidos e a prática de armazenar atributos do tipo derivado.

Para a realização do pré-processamento foi utilizado o Excel como ferramenta através da tabela dinâmica.

3.6.2 Tratamento da Base de dados

O tratamento da base de dados é a última fase realizada no pré-processamento onde foi realizado o agrupamento, normalizações categorizações e ajuste ou devidas correções na base de dados para que pudessem ser aplicadas as estatísticas e a execução de algoritmos para a mineração de dados. As ferramentas utilizadas para o tratamento da base foram o Excel e o RStudio.

Visto que muitos dos filmes e séries possuíam mais de um tipo de gênero, foram utilizados nessa análise somente os filmes e séries que possuíam gêneros únicos, totalizando 27 gêneros disponíveis no apêndice A.

Portanto todos os filmes, séries e minisséries com mais de um gênero foram resumidos a somente um gênero baseado na maior quantidade de gênero que foi produzido naquele ano. Por exemplo, o ano de 2012 teve um número muito grande de filmes de ação produzidos, e um filme que continha drama, ação e aventura foi classificado como ação. Depois que todos os títulos que possuem ação e mais outros

gêneros se tonarem somente um, e os demais foram sendo baseados no segundo maior gênero produzido naquele ano e assim por diante.

3.7 ESTATÍSTICA DESCRITIVA

De acordo com Guedes et Al (Séc. XXI, p.1), a estatística descritiva tem como objetivo de sintetizar e analisar uma série de valores de uma mesma natureza na base de dados para assim ter uma visão global desses valores tanto por meio de tabelas, gráficos e de medidas descritivas.

Segundo Reis (2002, p. 23), para a realização da estatística descritiva primeiramente é identificado o tipo de variável presente na base classificando em dados qualitativos e quantitativos. As variáveis quantitativas podem classificadas como contínuas e discretas. Contínua segue números em escala e as discretas são números inteiros e únicos. Já as variáveis qualitativas podem ser denominadas por nominais e ordinais, sendo as variáveis nominais categóricos sem ordenação e ordinais seguem uma lógica ordenada. Através desta categorização baseada por tipo de atributo foram selecionados os principais tipos de análise, de forma que pudessem gerar gráficos, tabela de frequências, gráficos, média, mínimo e máximo a fim de visualizar os resultados.

Para a realização da estatística foram utilizados o Excel e o RStudio como ferramentas de análise.

3.8 MINERAÇÃO DE DADOS

Os algoritmos escolhidos para a análise da base de dados do IMDB foram selecionados observando que boa parte dos atributos são qualitativos, sendo assim optaou-se por utilizar a Árvore de Decisão e o Naïve Bayes.

O algoritmo Árvore de Decisão, de acordo Campos (2017), é um método de aprendizado de máquinas utilizado em tarefas de classificação e regressão, e que armazena regras em seus nós, ramos e folhas que representam a decisão a ser tomada.

De acordo com Becker (2019), o algoritmo Naïve Bayes é um classificador probabilístico que categoriza textos baseado na frequência das palavras usadas.

3.9 PÓS PROCESSAMENTO

A última fase do KDD consiste no pós-processamento, que mostra a análise e a interpretação dos resultados considerando os resultados da estatística descritiva e dos algoritmos utilizados na mineração de dados. A análise de resultados foi feita pelo método de comparações de resultados entre os dois algoritmos comparando a taxa de acerto, matriz confusão e a interpretação da árvore de decisão e as probabilidades do Naive Bayes.

4 EXPERIMENTOS E ANÁLISE DOS RESULTADOS

Diferente da sessão anterior que buscou esclarecer a metodologia, na presente sessão serão esclarecidas as abordagens deste estudo através do pré-processamento da base, tratamento da base de dados, estatística descritiva, algoritmos de mineração de dados e o pós-processamento.

4.1 PRÉ PROCESSAMENTO

O pré-processamento dos dados tem como objetivo adaptar a base de dados para atender melhor os algoritmos de mineração de dados e as estatísticas descritivas. Através do pré-processamento foram realizados a limpeza dos dados e o tratamento das bases de dados.

A base do IMDB (2020) estava dividida entre sete arquivos em formato .TSV (Valores Separados por Tabulações) conforme descritos no QUADRO 7, sendo necessário converte-los para .CSV (*Comma-separated values*) para a utilização do script de junção da base no RStudio.

QUADRO 7 - ARQUIVOS DE BASES DE DADOS DO IMDB

ARQUIVO	DESCRIÇÃO
title.akas.tsv.gz	informações dos títulos
title.basics.tsv.gz	informações extras dos títulos
title.crew.tsv.gz	informações sobre os diretores e os escritores
title.episode.tsv.gz	informações sobre os episódios transmitidos na TV
title.principals.tsv.gz	informações sobre o elenco principal dos títulos
title.ratings.tsv.gz	informações sobre as notas dadas pelos usuários e a nota geral
name.basics.tsv.gz	informações dos nomes das pessoas envolvidas no títulos

FONTE: IMDB (2020)

Para juntar as bases em um único arquivo utilizou-se o programa RStudio o qual emprega a linguagem R. Para juntar foi utilizado o atributo em comum em todas "*tconst*", unidos pelo comando "MERGE", conforme a Figura 13, o qual junta os dados de forma horizontal na tabela. Porém nem todos os arquivos foram utilizados pois o objetivo desta pesquisa não é analisar o elenco, diretores e escritores mais sim a relação das notas com o tipo de título sendo filme e series além do gênero. Os arquivos que não foram utilizados para montar a base de dados foram: title.crew.tsv.gz (informação do diretor e escritor), title.akas.tsv.gz (região, linguagem e identidade), title.principals.tsv.gz (contém os dados do elenco), name.basics.tsv.gz (dados de cada indivíduo do elenco) e title.episode.tsv.gz (título, temporada e número de episódios.

Para a junção desta base foram utilizados dois arquivos, o primeiro arquivo é o "title.basics.tsv.gz" o qual foram selecionados os atributos como mostra o Quadro 8.

QUADRO 8 - ATRIBUTOS UTILIZADOS DA BASE TITLE BASICS

Campos (Atributos)	Descrição
tconst	Id de identificação
TitleType	O tipo do título presente na base
primaryTitle	Nome primário do título
OriginalTitle	O nome original sem tradução
isAdult	Se o filme é adulto ou não sendo adulto 1 e não adulto 0
startYea	Ano de lançamento do título
endYear	Ano de encerramento do título, geralmente é utilizado para término de séries
gêneros	Os tipos de gêneros de cada título podendo ser mais de um

FONTE: IMDB (2020)

Os 8 atributos presentes no Quadro 8 são os campos disponíveis dentro do arquivo do IMDB "title.basics.tsv.gz".

O segundo arquivo selecionado é o "title.ratings.tsv.gz" e foram utilizados os seguintes atributos conforme ilustrado no Quadro 9.

QUADRO 9 - ATRIBUTOS UTILIZADOS DA BASE RATINGS

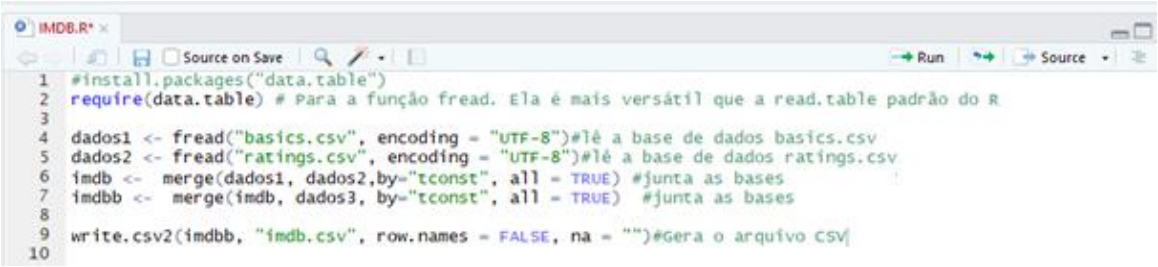
Campos (Atributos)	Descrição
tconst	Id de identificação
averageRating	média ponderada de todas as avaliações individuais do usuário
numVotes	Número total de votos

FONTE: IMDB (2020)

A união dos arquivos presentes no Quadro 8 e 9 resultou em uma base com (6.629.737) linhas e 10 variáveis sendo elas: tconst, titleType, primaryTitle, originalTitle, isAdult, startYear, runtimeMinutes, genres, averageRanting, numVotes.

Para a junção dos arquivos do IMDB foram usados os script no RStudio conforme a Figura 13.

FIGURA 13 - SCRIPT DE JUNÇÃO DE BASES



Fonte: O AUTOR (2020)

Observando o código na Figura 13, a primeira função executada é o “data.table” a qual disponibiliza mais funções no RStudio, em sequência são inseridos os dois arquivos de dados e para a junção das bases foi utilizado o comando “merge” o qual junta as bases através das colunas em comuns.

O arquivo “title.akas.tsv.gz” não foi utilizado porque os dados presentes no seu conteúdo já estavam presentes em outros arquivos. Além de que o título escrito não foi utilizado na análise porque a mineração não vai ser aplicada em texto, mas sim através do seu ID, e os dados presente neste arquivo são: titleId; ordering; title; region; language; types; atributos; e isOriginalTitle.

O arquivo “title.crew.tsv.gz” contém informações sobre o diretor e escritor de cada título, ele não vai ser utilizado pelo objetivo deste projeto não constituir em uma avaliação de diretor e escritor, portanto, foram retirados da análise. Os dados

presentes neste arquivo é *tconst* (identificador único alfanumérico do título); *directors* (diretor do título fornecido) e por fim o campo *writer* (escritor do título fornecido).

O arquivo “name.basics.tsv.gz” não foi utilizado por conter informações específicas de séries, sendo o número de episódios e temporadas. O objetivo deste trabalho não foi analisar somente séries, mas também filmes e utilizar estes campos comparando com filmes não possibilitaria um resultado comparativo.

O arquivo “title.principals.tsv.gz” também não foi utilizado pelas mesmas razões do arquivo anterior, o intuito deste projeto não é uma análise do elenco, diretores e escritores, mas sim a relação das notas com o tipo de título sendo filme e series além do gênero. Este arquivo possui os seguintes dados: tconst (identificador único alfanumérico do título); ordering (número para identificar exclusivamente as linhas de um determinado titleId); nconst (identificador único alfanumérico do nome / pessoa); categoria (categoria de trabalho em que a pessoa estava); job (cargo).

Por fim, o último arquivo “name.basics.tsv.gz” não foi utilizado pelas mesmas razões dos anteriores e seus dados são: nconst (identificador único alfanumérico do nome / pessoa); primaryName (nome pelo qual a pessoa é mais frequentemente creditada); birthYear; deathYear; primaryProfession (as três principais profissões da pessoa); knownForTitles (títulos pelos quais a pessoa é conhecida).

Entre os 10 atributos existentes na base, foram utilizados somente 8 atributos, por serem relevantes para o objetivo da pesquisa, conforme o Quadro 10.

QUADRO 10 - ATRIBUTOS DELETADOS

Atributo	Motivo
endYear	Ano em que terminou o seriado. Não foi necessário
primaryTitle	Título de produção, antes do lançamento. Não foi necessário

FONTE: O AUTOR (2020)

Através do Quadro 10 percebe-se que a maior parte dos atributos não utilizados estavam vazios, ou campos que não eram relevantes para a pesquisa como runtimeMinutes, endYear e primaryTitle.

O Quadro 11 contêm os 8 atributos que foram utilizados para análise do presente trabalho.

QUADRO 11 - ATRIBUTOS UTILIZADOS

Atributo	Descrição
tconst	Identificador único alfanumérico do título
titleType	Título criado durante o projeto do filme
originalTitle	Título conhecido comercialmente
isAdult	Se o conteúdo é adulto ou não
startYear	Representa o ano de lançamento de um título. No caso da série de TV, é o ano de início da série
genres	Inclui até três gêneros associados ao título
averageRating	Média ponderada de todas as classificações de usuários individuais (Nota de 1.0 a 10.0)
numVotes	Número de votos que o título recebeu

FONTE: O AUTOR (2020)

De forma em geral foram retirados da base de dados 2 atributos apresentados no Quadro 10 e analisados 8 atributos no Quadro 11.

4.1.1 Limpeza de dados

Para a análise desta pesquisa realizou-se um corte no campo “startYear” referente aos anos acima e igual a 2015, o qual resultou em uma base de dados com (1.755.703) linhas, já que o objetivo é analisar os anos de 2015 até 2019.

A primeira análise realizada nos dados foram a dos valores ruidosos, através deste exame foi executado um levantamento com o total de títulos por tipo de título de acordo com o Quadro 12.

QUADRO 12 – TOTAL DE TÍTULOS POR TIPO DE CONTEÚDO

Nome da Base	Nº de Linhas	Motivo do corte na base
Dados	6.629.737	Base original, não foi realizada modificação.

DadosAcima2015	1.755.703	Nesse corte foram mantidos os filmes e séries a partir de 2015.
Filmes	95.250	A partir daqui foram realizadosos cortes por tipo de título, este corte são os filmes lançados a partir de 2015.
Séries	45.978	Séries lançadas a partir de 2015.
tcurtas	225.485	Curtas lançados a partir de 2015.
tvEpisode	1.285.816	Episódios de séries a partir de 2015.
tvfilmes	16.543	Filmes lançados diretamente lançados para TV a partir de2015.
tvMiniseries	12.090	Mini-Séries a partir de 2015.
tvShort	2.181	Curtas lançado diretamente para TV a partir de 2015.
tvSpecial	6.161	Especiais lançados a partir de 2015.
video	60.928	Videos lançados a partir de 2015.
videoGame	5.271	Vídeos de vídeo game lançados a partir de 2015.

FONTE: O AUTOR (2020)

A partir do Quadro 12 é possível observar o total de títulos dentro de cada tipo variando entre filmes, séries, curtas e entre outros. Para análise deste trabalho foi considerado o titleType: “movie, TvMovie, TvMiniseries e tvseries”. Utilizando o filtro no campo “TitleType” foram retirados da análise 1.585.842 títulos que eram “tvShort”, “tvEpisode”, “tvSpecial”, “short”, “vídeo” e “videoGame", “tvShort”, “short”, “vídeo” e “videoGame". Foram retirados pois o foco é somente filmes e séries, “tvEpisode” foi retirado pois mostrava os dados dos episódios de séries pois o foco da análise é a

série como um todo e não cada episódio, e realizado esses cortes resultou em uma base de dados com (169.861) linhas.

No Quadro 13 contém os tipos de títulos utilizados nesta análise:

QUADRO 13 - DESCRIÇÃO DOS TIPOS DE TÍTULOS

TitleType	Análise	Descrição
movie	Utilizado para a Análise	Filmes lançados no cinema.
tvMovie	Utilizado para a Análise	Filmes lançados direto para a televisão.
tvSeries	Utilizado para a Análise	Séries lançadas tanto pela emissora quanto por canais de streaming.
tvMiniSeries	Utilizado para a Análise	Minisséries lançadas na televisão.

FONTE: O AUTOR (2020)

A segunda análise empregada foi acerca dos valores inconsistentes e ao gerar a base identificou-se que vários dados em relação ao ano de lançamento estavam errados ou que não se encaixavam nos anos entre 2015 e 2019. Dessa forma, foi criada uma tabela dinâmica ilustrando os anos inconsistentes utilizando como colunas o “titleType” e as linhas e a contagem utilizando o “startYear”, conforme na Tabela 1 mostra a quantidade de anos de cada tipo de título entre 2020 e 2115:

TABELA 1 - ANOS INCONSISTENTES DA BASE DE DADOS

Anos	movie	tvMiniSeries	tvMovie	tvSeries	Total Geral
2020	8508	519	530	2140	11697
2021	970	15	21	77	1083
2022	181	2	3	9	195
2023	41	0	0	2	43
2024	12	1	0	0	13
2025	2	0	0	0	2
2026	4	0	0	1	5
2027	4	0	0	0	4
2115	1	0	0	0	1
Total Geral	**9723**	**537**	**554**	**2229**	**13043**

FONTE: O AUTOR (2020)

Os anos entre 2021 e 2115 não serão utilizados pois os títulos não foram lançados ainda e o ano de 2020 por não ter um ciclo fechado também foi eliminado.

Como podemos ver na Tabela 1 a coluna “*movie*” no ano 2020 possui (8.508) linhas, apesar de estarmos em 2020 o ciclo do ano não está completo, por este motivo não foi utilizado na base de dados e somando os anos seguintes foi eliminado 9.723 títulos.

Em relação a “tvMiniSeries”, o ano 2020 possui (519) linhas e somado aos anos seguintes o total de anos que não serão utilizados esta análise são de (537) linhas.

E de acordo com a “tvMovie” o ano 2020 possui (530) linhas e somado aos anos seguintes o total de anos que não serão utilizados esta análise são de (554) linhas.

E por fim em relação a tvSeries o ano 2020 possui (2.140) linhas e somado aos anos seguintes o total de anos que não serão utilizados esta análise são de (2.229) linhas.

Por meio desta avaliação foram identificados (13.043) registros com dados iguais ou acima de 2020, reduzindo a base para (156.818) linhas de (169.861) titulos.

A terceira análise realizada na base de dados são os valores ausentes, através dos atributos “averageRating”, “numVotes” e “genres” foi identificado que muitos títulos não possuíam registros. No caso do atributo “genres” viu-se que cerca de 7.262 filmes e séries não continham dados no gênero, já no atributo “averageRating” e “numVotes” foram eliminadas (84.822) linhas, pois em ambos os mesmos títulos não tinham registros, porém um título tinha dados no atributo “numVotes” e não tinha dados no “averageRating”, portanto este título foi eliminado da pesquisa por estar incompleto.

Através desta validação foram identificados (92.085) registros com dados ausentes, reduzindo a base de (156.818) linhas para (64.733) títulos.

A quarta e última análise corresponde aos valores redundantes e na base de dados não foi encontrado nenhum registro. Lembrando que podem existir dois títulos com o mesmo nome, sendo filme ou série por exemplo.

Através dessa sessão de limpeza dos dados foram eliminados (6.565.004) títulos de (6.629.737) e a base de dados ficando assim com (64.733) linhas.

O Quadro 14 contém o total de valores retirados nas quatro validações.

QUADRO 14 - RELAÇÃO DE DADOS ELIMINADOS APLICADOS EM QUATRO VALIDAÇÕES

Tipos de validações	Valores retirados
Valores Ruidosos	6.459.876
Valores Inconsistentes	13.043
Valores Ausentes	92.085
Valores Redundantes	0
TOTAL	6.565.004

FONTE: A AUTOR (2020)

Observando o Quadro 14, verifica-se que a maior concentração de dados eliminados da tabela foi correspondente aos valores ruidosos e em sequência pelos valores ausentes. O total de valores eliminados da base de dados para a realização desta pesquisa foi de (6.565.004) títulos e a base de dados ficou com (64.733) títulos.

4.1.1 Tratamento da Base de Dados

Através desta sessão foram realizados a normalização, categorização agrupamento e classificação dos dados. Através destes ajustes e adaptação na base de dados foi necessário para a aplicação do algoritmo de mineração de dados e análise da estatística.

A primeira análise consiste no agrupamento do atributo “averageRating”, as classificações atribuídas variam de 1 a 10 e, para fins desta pesquisa, foram agrupadas conforme mostra o QUADRO 15.

QUADRO 15 - AGRUPAMENTO DA CLASSIFICAÇÃO MÉDIA DOS TITULOS

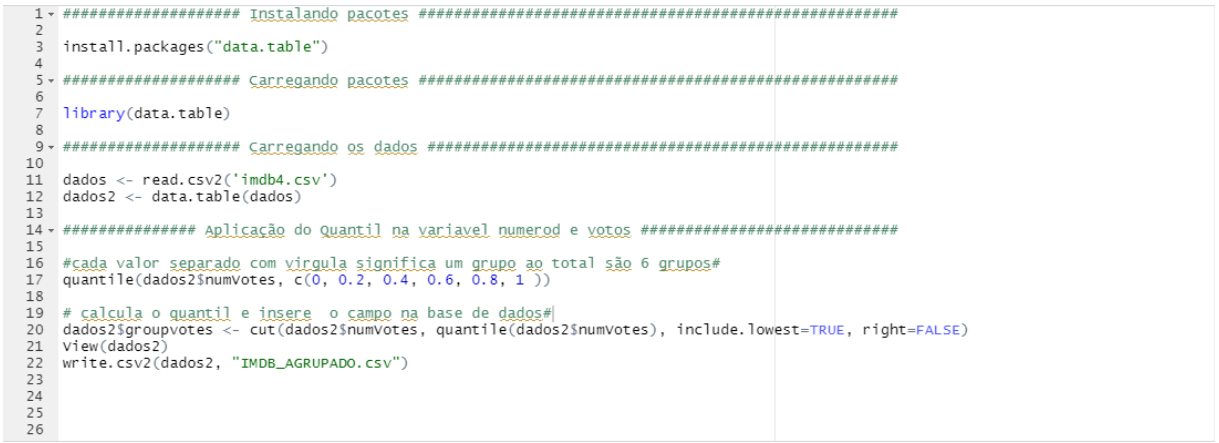
Classificação	Agrupamento
1 a 2,9	E
3 a 4,9	D
5 a 6,9	C
7 a 8,9	B
9 a 10	A

FONTE: O AUTOR (2020)

Através do Quadro 15 observa-se que as melhores classificações estão agrupadas na letra A e a letra E obteve os piores resultados.

O segundo agrupamento realizado nesta pesquisa foi no atributo numVotes. A variação neste atributo foi bem alta, os valores de números de votos variam de 5 até 860.692, para o seu agrupamento foi necessário criar um script no RStudio apresentado na Figura 14.

FIGURA 14 - SCRIPT DE APLICAÇÃO DO AGRUPAMENTO



```
1  ################### Instalando pacotes ######################################################
2
3  install.packages("data.table")
4
5  ################### Carregando pacotes ######################################################
6
7  library(data.table)
8
9  ################### Carregando os dados #####################################################
10
11 dados <- read.csv2('imdb4.csv')
12 dados2 <- data.table(dados)
13
14 ############### Aplicação do Quantil na variavel numerod e votos ############################
15
16 #cada valor separado com virgula significa um grupo ao total são 6 grupos#
17 quantile(dados2$numVotes, c(0, 0.2, 0.4, 0.6, 0.8, 1 ))
18
19 # calcula o quantil e insere  o campo na base de dados#
20 dados2$groupvotes <- cut(dados2$numVotes, quantile(dados2$numVotes), include.lowest=TRUE, right=FALSE)
21 view(dados2)
22 write.csv2(dados2, "IMDB_AGRUPADO.csv")
23
24
25
26
```

FONTE: O AUTOR (2020)

Explicando detalhadamente a Figura 14, pode-se observar que primeiramente foi instalado o pacote data.table, um pacote que disponibiliza mais atribuições do que o R disponibiliza no básico. Após a instalação do pacote foi carregado o arquivo da base de dados IMDB4. Na linha 17 foi executado o código “quantile” para gerar os grupos com as seguintes definições c(0, 0.2, 0.4, 0.6, 0.8, 1). Após ser gerado os grupos na linha 20 foi inserido o resultado do agrupamento realizado pelo quantile e na linha 22 é transformado para uma base de dados .CSV já com os agrupamentos realizados na coluna.

Ao final do código o atributo do agrupamento ficou como “groupvotes” e o arquivo da base como “IMDB_AGRUPADO.CSV”. Após realizado o agrupamento com o quantile obteve-se como resultado os grupos conforme descritos na Tabela 3.

TABELA 2 - TABELA DE ROTULAÇÃO

Agrupamento no R	Variação do agrupamento	Total de títulos por grupo	Rotulação dos grupos
Grupo [5,13)	De 5 a 13 nº de votos	15.581	G1
Grupo [13,43)	De 13 a 43 nº de votos	16.621	G2
Grupo [43,233);	De 43 a 233 nº de votos	16.342	G3
Grupo [233,8.61e+05];	De 233 a 8.61e+05 nº de votos	16.189	G4

FONTE: O AUTOR (2020)

De acordo com a Tabela 3, pode-se observar que o quantile agrupou proporcionalmente os 4 grupos com uma média de 16.183 títulos distribuídos conforme a variação descrita. Os grupos foram rotulados para melhor análise na aplicação dos algoritmos no capítulo de mineração de dados.

O terceiro tratamento realizado nas bases de dados consistiu na categorização no atributo “startYear”. Isso foi necessário pois o algoritmo utilizado possui problemas para analisar dados numéricos tendo que ser todos qualitativos ou binários.

Sendo assim, os anos 2015, 2016, 2017, 2018 e 2019 foram substituídos conforme apresentado na Tabela 3.

TABELA 3 - RENOMEAÇÃO DOS ANOS

Ano	Nome
2015	A1
2016	A2
2017	A3
2018	A4
2019	A5

FONTE: O AUTOR (2020)

Através da Tabela 3 observa-se que a categorização foi realizada de forma crescente sendo 2015 (A1) e 2019 (A5).

4.2 ESTATÍSTICA

Neste capítulo através da identificação dos atributos presente no QUADRO 16, definiu-se a análise a ser empregada. No presente capítulo serão analisadas através de tabela de frequências, gráficos de barras, mínimo e máximo.

QUADRO 16 - CLASSIFICAÇÃO DOS ATRIBUTOS

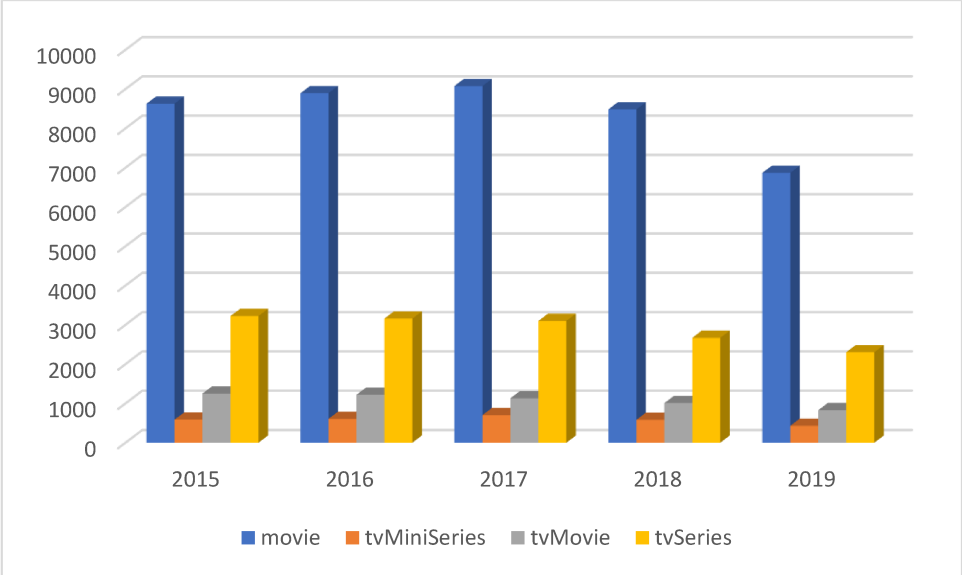
Atributo	Descrição	Tipo	Variação
titleType	Tipos de títulos, se é filme, série, minissérie	Variável Qualitativa Nominal	4 tipos de título
isAdult	Se o filme é para maiores de idade ou não	Variável Booleana	0 ou 1, 0 para não adulto e 1 para adulto
startYear	Ano de estreia do filme ou do seriado	Variável Qualitativa Ordinal	Ano de 2015 até 2019
Genres	O gênero do filme ou seriado, comédia ou drama por exemplo	Variável Qualitativa Nominal	27 tipos de gênero
averageRating	Nota que o título recebeu	Variável Qualitativa Ordinal	Média das notas
groupvotes	Quantidade de votos que o título recebeu	Variável Qualitativa Ordinal	Quantidade de votos

FONTE: O AUTOR (2020)

Através do Quadro 16 observa-se que os tipos das variáveis são qualitativas concentração a maior parte dos tipos de variável são qualitativas ordinal, sendo 3 atributos ordinal, e 2 nominal e 1 booleana.

A primeira análise realizada é a variável “titleType”, a qual está classificada como variável qualitativa nominal. O Gráfico 1 mostra o resultado que gerou através de uma tabela de frequência absoluta a qual cruzou os atributos “titleType” e “startYear”.

GRÁFICO 1 - QUANTIDADE DE TIPOS DE TÍTULOS



FONTE: O AUTOR (2020)

Através do Gráfico 1, percebe-se que os filmes são os mais produzidos durante os anos 2015 a 2019 em comparação aos demais tipos. Enquanto as minisséries são as menos produzidas comparado aos mesmos. Além disso, pôde-se observar que a quantidade de filmes, séries, minisséries e filmes lançados para a TV foram aumentando até 2017 e começaram a reduzir em 2018.

Para melhor análise, a Tabela 4 mostra a quantidade de cada tipo de título por ano de forma detalhada.

TABELA 4 - PRODUÇÕES ANUAIS DO TIPOS DE TÍTULOS

titleType	2015	2016	2017	2018	2019	Total Geral
movie	8628	8894	9076	8482	6867	41947
tvMiniSeries	586	605	700	580	430	2901
tvMovie	1250	1219	1130	1008	828	5435
tvSeries	3223	3158	3098	2669	2302	14450
Total Geral	**13687**	**13876**	**14004**	**12739**	**10427**	**64733**

FONTE: O AUTOR (2020)

Verificando a Tabela 4, o tipo de título mais produzido foi o “movie” durante esses 5 anos analisados com um total de (41947) títulos, e o menos produzido foram as minisséries com um total de (2901) títulos.

A segunda análise realizada neste trabalho o foi no atributo “isAdult”, essa variável serve para definir se um “titleType” é de conteúdo para adultos ou não. Se for

adulto é valor 1(um) caso contrário 0(zero). A Tabela 5 de frequência absoluta mostra a quantidade de título adultos e não adultos por ano.

TABELA 5 - QUANTIDADE DE FILMES ADULTOS PRODUZIDOS

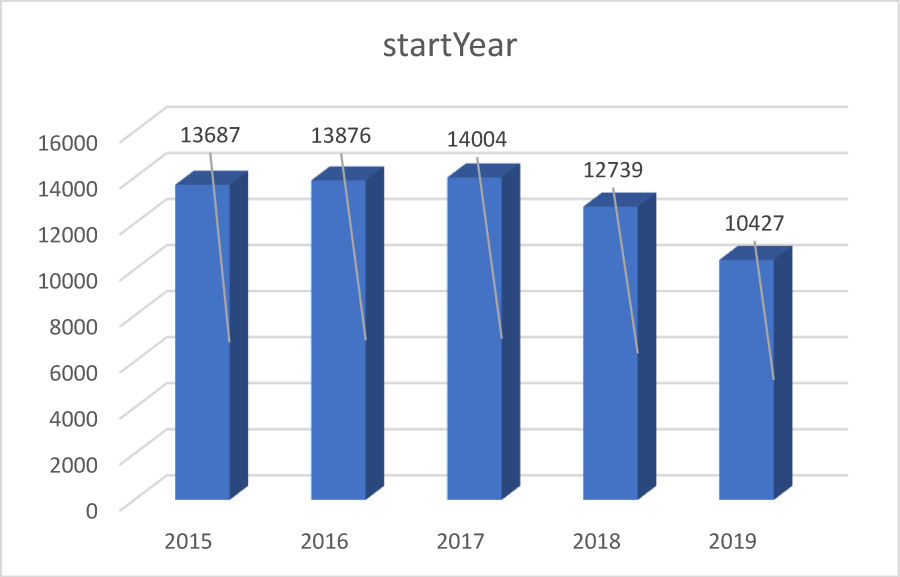
isAdult	2015	2016	2017	2018	2019	Total Geral
Não é adulto (0)	13661	13849	13976	12733	10424	64643
É adulto (1)	26	27	28	6	3	90
Total Geral	**13687**	**13876**	**14004**	**12739**	**10427**	**64733**

FONTE: O AUTOR (2020)

Na Tabela 58, percebe-se que filmes não adultos possuem uma quantidade relativamente grande comparado aos filmes adultos.

A terceira análise realizada é do atributo “startYear”, classificada como uma variável qualitativa ordinal, essa variável mostra o ano de lançamento dos filmes, séries e minisséries. No Gráfico 2 mostra a quantidade de filmes, séries e minisséries feitos por ano.

GRÁFICO 2 - QUANTIDADE DE TÍTULOS PRODUZIDOS POR ANO



FONTE: O AUTOR (2020)

Observando o Gráfico 2, (2017) foi o ano em que mais foram produzidos filmes, séries e minisséries enquanto (2019) foi o que teve menos produção. A Tabela 9 mostra de forma detalhada o que o Gráfico 2 está ilustrando.

TABELA 6 - CONTAGEM DOS ANOS

Rótulos de Linha	startYear
2015	13687
2016	13876
2017	14004
2018	12739
2019	10427
Total Geral	**64733**

FONTE: O AUTOR (2020)

Através desta análise da Tabela 6 percebe-se também que o ano de 2017 conteve mais títulos que os demais.

A quarta análise é o atributo “Genres”, classificada como uma variável qualitativa nominal. A Tabela 7 contém a quantidade de títulos em relação a gêneros por anos. Para melhor análise foram considerados apenas os 10 principais gêneros que mais tiveram produções as demais informações está presente no apêndice B.

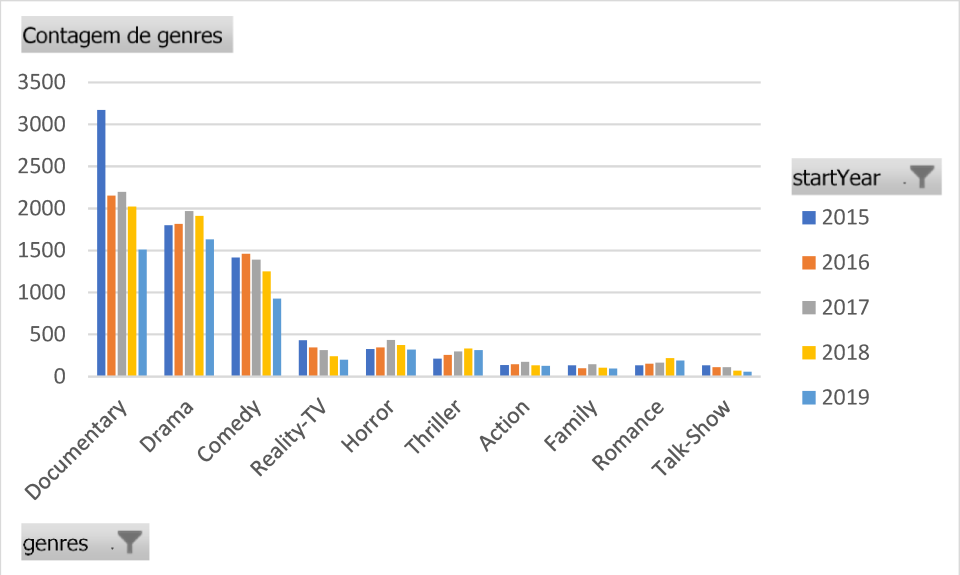
TABELA 7 - QUANTIDADE DE GÊNEROS POR ANO SEM A NORMATIZAÇÃO

Gênero	2015	2016	2017	2018	2019	Total Geral
Drama	4707	4809	4878	4594	4030	23018
Documentary	3164	3092	3091	2699	1959	14005
Comedy	2504	2601	2446	2197	1672	11420
Horror	663	709	778	682	555	3387
Thriller	452	498	549	567	511	2577
Action	353	355	422	337	279	1746
Reality-TV	469	389	366	281	231	1736
Romance	164	202	197	268	220	1051
Animation	163	219	187	198	156	923
Family	223	144	217	145	123	852
Total Geral	**12862**	**13018**	**13131**	**11968**	**9736**	**60715**

FONTE: O AUTOR (2020)

O gênero que mais contêm títulos é Drama com um total de (23018) títulos, seguindo pelo gênero Documentary com (14005) títulos. Já o gênero Family obteve o menor resultado com um total de (852) títulos neste ranking de 10 principais gêneros com mais títulos. Para melhor visualização do Gráfico 3, contém a representação da Tabela 7.

GRÁFICO 3 - QUANTIDADE DE GÊNEROS POR ANO SEM A NORMATIZAÇÃO

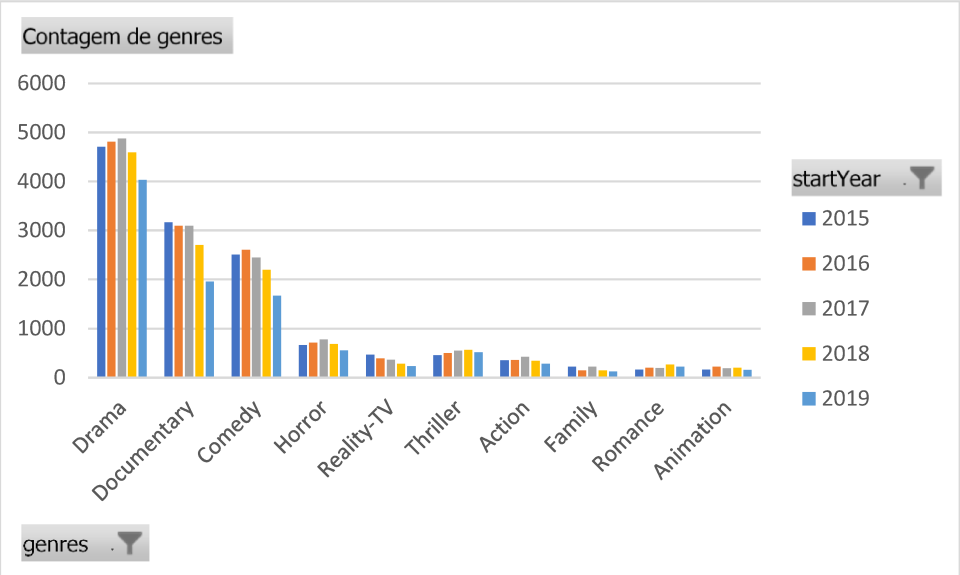


FONTE:O AUTOR (2020)

Através do Gráfico 3 observa-se de forma decrescente a evolução do total de títulos por gêneros. Outro ponto importante de se notar foi que drama, horror e ação tiveram alta até 2017 e nos anos seguintes foram decaindo.

Aplicando o método descrito na seção de metodologia podemos ver na GRÁFICO 4 que, dos 10 gêneros mais produzidos, muitos dos gêneros obtiveram um crescimento expressivo na quantidade produzida por ano e os filmes e séries de animação superou o talk-show.

GRÁFICO 4 - QUANTIDADE DE GÊNEROS POR ANO COM A NORMATIZAÇÃO



FONTE:O AUTOR (2020)

A quarta análise realizada é do atributo “averageRating” relacionada com os gêneros e total de classificação, classificada como uma variável qualitativa ordinal apresentado na Tabela 8.

TABELA 8 - PORCENTAGEM DE NOTAS RECEBIDAS POR CADA GÊNERO

Rótulos de Linha	A	B	C	D	E
Action	1,6%	32,2%	43,9%	17,7%	4,6%
Adult	2,0%	26,0%	66,0%	6,0%	0,0%
Adventure	9,3%	37,3%	36,9%	13,8%	2,6%
Animation	3,0%	41,1%	40,7%	12,6%	2,6%
Biography	3,6%	52,0%	37,3%	5,6%	1,6%
Comedy	2,4%	31,5%	46,0%	17,4%	2,8%
Crime	1,5%	40,1%	48,2%	8,9%	1,3%
Documentary	5,3%	61,0%	30,4%	2,9%	0,4%
Drama	1,6%	35,1%	51,8%	10,3%	1,2%
Family	3,3%	28,5%	51,9%	14,9%	1,4%
Fantasy	1,1%	43,1%	36,7%	12,8%	6,4%
Game-Show	3,7%	26,5%	40,2%	26,9%	2,7%
History	3,5%	59,0%	31,0%	5,7%	0,9%
Horror	0,3%	11,9%	37,2%	40,3%	10,3%
Music	6,7%	49,0%	35,2%	6,7%	2,4%
Musical	1,2%	47,8%	38,5%	11,2%	1,2%
Mystery	1,7%	38,4%	49,0%	10,3%	0,7%
News	8,2%	29,9%	30,9%	26,8%	4,1%
Reality-TV	4,3%	41,0%	39,1%	13,3%	2,3%
Romance	1,4%	30,1%	57,4%	10,0%	1,1%
Sci-Fi	1,9%	35,5%	32,8%	21,4%	8,4%
Short	30,4%	39,1%	17,4%	13,0%	0,0%
Sport	5,8%	53,6%	32,4%	7,2%	1,0%
Talk-Show	10,3%	37,5%	33,4%	13,3%	5,5%
Thriller	0,5%	21,2%	49,3%	26,5%	2,4%
War	5,8%	40,4%	28,8%	23,1%	1,9%
Western	2,9%	21,7%	36,2%	23,2%	15,9%
Total Geral	**2,7%**	**38,7%**	**43,9%**	**12,7%**	**2,1%**

FONTE:O AUTOR (2020)

Analisando a Tabela 8 podemos ver que no geral 43,9% dos filmes, séries e minisséries tiveram a nota C enquanto somente 2,1% tiveram nota E, das notas C o gênero que teve a maioria delas foi o gênero adulto, e para a nota E foi o gênero

horror, para melhor análise no apêndice C contém a tabela de total de classificação por quantidade.

Através da análise do atributo “groupvotes”, classificada como uma variável qualitativa ordinal, o atributo mostra a quantidade de votos que cada título recebeu. A Tabela 9 mostra a quantidade de votos que cada gênero recebeu por número de votos.

TABELA 9 - QUANTIDADE DE VOTOS POR GÊNERO

Rótulos de Linha	G1	G2	G3	G4	Total Geral
Drama	3905	5259	6174	7680	23018
Documentary	5157	4383	2913	1552	14005
Comedy	2496	2749	2933	3242	11420
Horror	542	816	932	1097	3387
Thriller	394	495	797	891	2577
Action	344	342	419	641	1746
Reality-TV	572	577	426	161	1736
Romance	209	280	287	275	1051
Animation	265	315	264	79	923
Family	223	206	292	131	852
Crime	178	259	240	113	790
Talk-Show	261	159	97	25	542
Sci-Fi	132	101	86	50	369
Mystery	99	92	58	43	292
Adventure	104	77	62	25	268
Biography	79	84	56	33	252
History	91	63	56	19	229
Game-Show	94	66	42	17	219
Music	93	55	32	30	210
Sport	104	54	35	14	207
Fantasy	58	60	47	23	188
Musical	61	39	41	20	161
News	32	40	16	9	97
Western	23	14	18	14	69
War	16	17	14	5	52
Adult	35	12	3	0	50
Short	14	7	2	0	23
Total Geral	**15581**	**16621**	**16342**	**16189**	**64733**

FONTE:O AUTOR (2020)

Analisando a Tabela 9 pode-se notar que os gêneros drama, documentary e comedy são as que mais recebem notas, o que faz sentido, levando em conta que o

número de produções para esses gêneros é grande, já os gêneros adulto e curtas são os que tiveram menos notas.

Na próxima análise, foi utilizado os tipos de títulos para verificar a quantidade de votos de cada grupo de número de votos de cada tipo que recebeu, obteve-se os seguintes resultados como é possível observar na Tabela 10.

TABELA 10 - QUANTIDADE DE VOTOS NO GRUPOS POR TIPO DE TÍTULO

Rótulos de Linha	G1	G2	G3	G4	Total Geral
movie	8783	10232	11167	11765	41947
tvMiniSeries	938	794	633	536	2901
tvMovie	1631	1544	1199	1061	5435
tvSeries	4229	4051	3343	2827	14450
Total Geral	**15581**	**16621**	**16342**	**16189**	**64733**

FONTE: O AUTOR (2020)

Diferente da Tabela 4 que também avalia o titleType por startyear, a Tabela 10 analisa o titleType por grupos e observa-se uma leve mudança nos valores dentro de cada tipo. Por exemplo, o titleType “movie” na Tabela 4 no ano de 2015 e contagem de anos obteve-se (8628) títulos, já a análise da Tabela 10 com o agrupamento de contagem no grupo G1 obteve-se no titleType “movie” (8783) registros.

4.3 MINERAÇÃO DE DADOS

Após a realização do pré-processamento e da análise estatística dos atributos, este capítulo de mineração de dados tem como objetivo aplicar os algoritmos com a finalidade de identificar padrões na base do IMDB com a aplicação do algoritmo de árvore de decisão e o algoritmo Naïve Bayes e com os dados selecionados entre os anos de 2015 e 2019.

O software utilizado para a mineração de dados é o RStudio (Versão 1.3.1093), conforme descrito no capítulo de metodologia.

Como visto no capítulo 4.1 a base foi encontrada e dividida entre sete arquivos em formato. TSV (Valores Separados por Tabulações), sendo necessário convertê-los para .CSV. Após a realização do pré-processamento com a limpeza de dados, tratamento e a análise de estatísticas, foi realizada a carga dos dados para a aplicação

dos algoritmos de mineração de dados assim como a instalação dos pacotes, como mostra os scripts na Figura 15 no capítulo da Árvore de Decisão.

4.3.1 Árvore de Decisão

Antes da aplicação do algoritmo de mineração de dados é realizado a carga de dados no RStudio através da Figura 15.

FIGURA 15 - SCRIPT DA CARGA DE DADOS

```
1  #Instalando os pacotes necessários
2  install.packages("rpart")
3  install.packages("prp")
4  install.packages("rpart.plot")
5  install.packages("data.table")
6
7  #Executando os pacotes necessários
8  library(rpart)
9  library(prp)
10 library(rpart.plot)
11 library(data.table)
12
13 #Carregando os dados
14 dados <- read.csv2('IMDB_AGRUPADO.csv')
15
16 #Executando os dados carregados na função a qual possibilita mais funções que o padrão do R
17 dados2 <- data.table(dados)
```

FONTE:O AUTOR (2020)

Como ilustra a Figura 15, o código começa com a instalação de 4 pacotes, “rpart”, “prp”, “rpart.plot”, “data.table”. Na linha 14 o código vai executar o arquivo IMDB_AGRUPADO.csv e colocará na variável dados, após os dados serem carregados, foi executado a base na função data.table conforme a linha 17 e esta função permite disponibilizar mais funções de execução no R renomeado para a variável dados2.

O segundo passo realizado após inserção dos dados foi a aplicação dos scripts da árvore de decisão, como mostra a Figura 16:

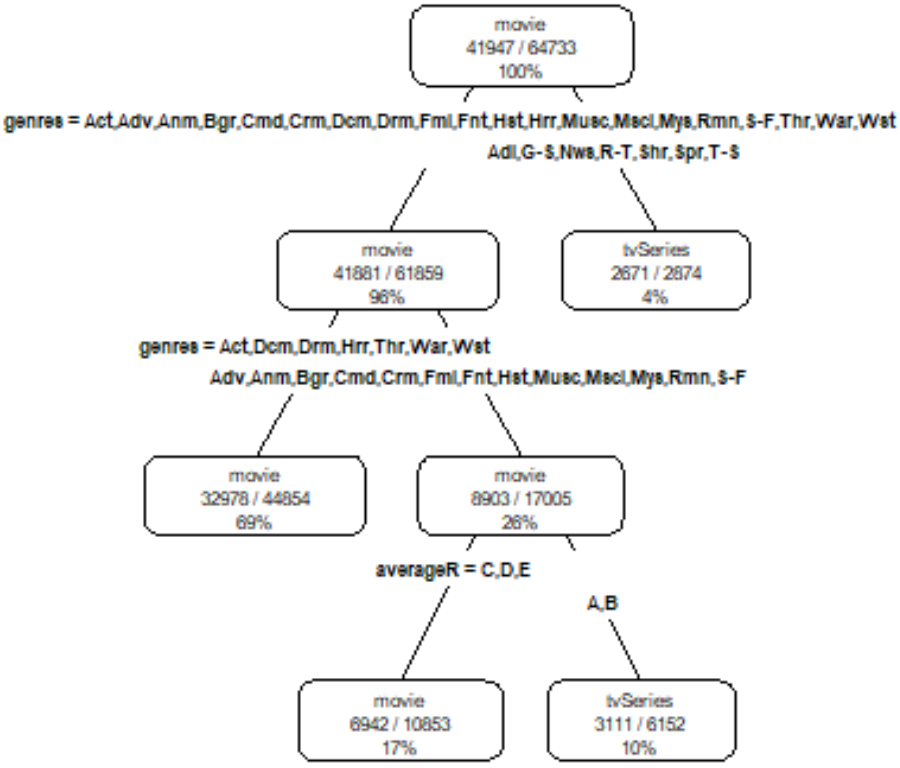
FIGURA 16 - SCRIPT DA ÁRVORE DE DECISÃO

```
19 #Arvore de decisão
20 arvore <- rpart(titleType ~ isAdult + startYear + groupvotes + averageRating + genres, data = dados2, method = 'class')
21
22 #Visualização da Arvore
23 prp(arvore, type = 4, extra = 102) #árvore de decisão
24 rpart.plot(arvore, type = 4, extra = 102) #árvore de decisão simplificada
```

FONTE: O AUTOR (2020)

Através dos scripts apresentados na Figura 16, foi aplicado o algoritmo da Árvore de Decisão, em seguida, na linha 20 foi realizada a árvore de decisão e levou menos de um minuto de processamento, onde foi utilizada a função “rpart” e selecionado o atributo meta como “titleType” e os atributos “isAdult”, “startYear”, “groupvotes”, “averageRating”, “genres” foram utilizados para a análise. Na linha 23, foi executada a função “prp” que é a visualização da árvore, e dentro dessa função é carregado o resultado da variável árvore com “type = 4” que por sua vez rotula todos os nós além das folhas e o “extra = 102” exibe as informações extras do nó, sendo o número 10 as probabilidades e o número 2 as taxas de classificação no nó. Na Figura 17 é demonstrada como é a visualização da árvore de decisão quando executado no RStudio:

FIGURA 17 - ÁRVORE DE DECISÃO



FONTE: O AUTOR (2020)

Como visualizado na Figura 17, a árvore mostra que os tipos de título que tiveram mais relevância foi “movie” e “tvseries”, também é possível notar que a árvore se repartiu em duas ramificações a partir da nó raiz pelo atributo gênero sendo que a primeira ramificação consiste nos gêneros: action, adventure, animation, biography, comedy, crime, documentary, family, fantasy, history, horror, music, musical, mystery, romance, Sci-Fi, thriller, war, western e desses gêneros. Dessa forma, 96% pertencem ao tipo de título “movie” e no segundo galho consiste nos gênero: adult,game-show, News, reality-tv, short, sport, talk-show e 4% pertencem ao tipo de título “tvSeries”. Já na próxima divisão de nós, foi visto que os gêneros que obtiveram mais relevância dentro de filmes, foi action, documentary, horror, thriller, war e western. A outra ramificação foi a que obteve menos relevância, e desses com menos relevância, 17% obtiveram notas baixas em relação aos 10% de “tvseries” que tiveram notas altas.

Após ser realizada a visualização da árvore, foi realizada a matriz confusão de acordo com os scripts da Figura 18:

FIGURA 18 - SCRIPT DA MATRIZ CONFUSÃO DA ÁRVORE DE DECISÃO

```
26  #Matriz confusão
27  matriz <- predict(arvore, dados2, type = 'class')
28  table(dados2$titleType, matriz)
```

FONTE:O AUTOR (2020)

Na linha 27 do script apresentado na Figura 18 foi feita a matriz confusão, onde foi executada a função predict utilizando o resultado da variável árvore. Utilizando a tabela dados2 e colocando o tipo para classe. Na linha 28 com a função *table* foi relacionada ao resultado da matriz com o atributo meta gerando a matriz confusão conforme visto na Figura 19:

FIGURA 19 - MATRIZ CONFUSÃO DA ÁRVORE DE DECISÃO

```
             matriz
              movie tvMiniSeries tvMovie tvSeries
movie         39920            0       0     2027
tvMiniSeries  2183             0       0      718
tvMovie       4936             0       0      499
tvSeries      8668             0       0     5782
```

FONTE:O AUTOR (2020)

A matriz confusão é analisada na diagonal, e através da Figura 19 percebe-se que tvminiseries e tvmovie não tiveram acertos pelo algoritmo, já os outros ele acertou 39.920 para movie e 5.782 para tvSeries. Portanto, considerando o resultado da matriz confusão, foi calculado a taxa de acerto, onde o cálculo é feito somando todos os valores da diagonal dividindo pelo total de linhas da base, sendo (39.920 + 0 + 0 + 5.782) /64.733 = 0,706007 o que vai dar aproximadamente 70% de acerto.

4.3.2 Naïve Bayes

O segundo algoritmo utilizado foi o Naïve Bayes, e de acordo com a Figura 20, segue os mesmos passos realizados em relação ao script da Árvore de Decisão, com uma diferença nos carregamentos de pacotes (“psych”, “dplyr”, “caret”, “stringr”, “naivebayes”, “e1071” e “data.table”), como mostra a Figura 20:

FIGURA 20 - INSTALAÇÃO DE PACOTES DO NAIVE BAYES

```
#Naive Bayes
#Instalando os pacotes necessários
install.packages("psych")
install.packages("dplyr")
install.packages("caret")
install.packages("stringr")
install.packages("naivebayes")
install.packages("e1071")
install.packages("data.table")

#Executando os pacotes necessários
library("caret")
library("psych")
library("dplyr")
library("stringr")
library("naivebayes")
library("e1071")
library("data.table")
```

FONTE:O AUTOR (2020)

Na Figura 20, foi realizada a instalação e carregamento dos pacotes necessários para a aplicação do algoritmo Naïve Bayes.

Já a Figura 21 contém os scripts necessários para realização da carga de dados dos indicadores.

FIGURA 21 - CARREGAMENTO DOS DADOS

```
20  #Carregando os dados
21  dados <- read.csv2('IMDB_AGRUPADO.csv')
22
23  #Executando os dados carregados na função a qual possibilita mais funções que o padrão do R
24  dados2 <- data.table(dados)
```

FONTE:O AUTOR (2020)

O carregamento da planilha em CSV do IMDB para a variável “dados” na linha 20 foi executado, em sequência foi executado a função data.table na variável “dados” e renomeado para “dados2”. Essa função é um pacote R que fornece uma versão aprimorada de data.frame, que são a estrutura de dados padrão para armazenar dados na base R, de acordo com a Figura 21.

O script apresentado na Figura 22 executa o algoritmo Naive Bayes e utiliza assim como o algoritmo da árvore de decisão o atributo “titleType” como atributo meta.

FIGURA 22 - EXECUÇÃO DO ALGORITMO

```
26  #Naive Bayes
27  bayes <- naiveBayes(titleType ~ isAdult + startYear + groupvotes + averageRating + genres, data = dados2)
```

FONTE:O autor (2020)

De acordo com a Figura 22, foram considerados os mesmos atributos da árvore de decisão descritos conforme a Figura 16 sendo eles: “isAdult”, “startYear”, “groupvotes”, “averageRating”, “genres”.

A execução do algoritmo Naive Bayes levou menos de 1 minuto de processamento na base, para a visualização dos resultados foi executado o comando “Bayes”, e apresentados os resultados. O primeiro resultado apresentado é a probabilidade referente ao atributo titleType de acordo com a Figura 23.

FIGURA 23 – PREDIÇÃO

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
      movie tvMiniSeries      tvMovie     tvSeries
0.64800025   0.04481485   0.08396027   0.22322463
```

FONTE:O AUTOR (2020)

Após o processamento do algoritmo Naive Bayes o resultado mostra que das quatros instâncias do atributo titleType (“movie”, “tvMiniseries”, “tvMovie” e “tvSeries”) percebe-se que o “movie” é o que tem a maior probabilidade em relação as demais instâncias do atributo está com (0.64) de acordo com a Figura 23.

Através da aplicação do algoritmo o atributo “genres” nota-se que a concentração das melhores probabilidades no “tvMovie” conforme a Figura 24.

FIGURA 24 - PROBABILIDADES CONDICIONAL DOS GÊNEROS

```
Conditional probabilities:
              genres
Y                Action        Adult    Adventure    Animation    Biography       Comedy
  movie       2.870289e-02 2.383961e-04 3.480583e-03 9.321286e-03 3.337545e-03 1.484016e-01
  tvMiniSeries 2.516374e-02 3.447087e-04 6.894174e-03 1.964840e-02 4.825922e-03 1.916580e-01
  tvMovie     1.159154e-02 3.679853e-04 2.023919e-03 7.911684e-03 8.831647e-03 1.617295e-01
  tvSeries    2.809689e-02 2.560554e-03 6.297578e-03 2.989619e-02 3.460208e-03 2.602076e-01
              genres
Y                 Crime  Documentary        Drama       Family      Fantasy    Game-Show
  movie       6.198298e-03 2.341526e-01 4.000524e-01 1.220588e-02 2.360121e-03 0.000000e+00
  tvMiniSeries 2.102723e-02 2.430196e-01 3.085143e-01 4.136505e-03 4.825922e-03 1.034126e-03
  tvMovie     2.907084e-02 3.396504e-01 2.842686e-01 1.655934e-02 2.207912e-03 0.000000e+00
  tvSeries    2.152249e-02 1.129412e-01 2.627682e-01 1.647059e-02 4.359862e-03 1.494810e-02
              genres
Y                History       Horror        Music      Musical      Mystery         News
  movie       1.835650e-03 7.328295e-02 1.931008e-03 1.954848e-03 3.027630e-03 4.767921e-05
  tvMiniSeries 1.137539e-02 2.068252e-02 1.172010e-02 3.447087e-03 1.895898e-02 2.068252e-03
  tvMovie     6.071757e-03 8.831647e-03 8.095676e-03 7.911684e-03 8.279669e-03 0.000000e+00
  tvSeries    5.951557e-03 1.418685e-02 3.529412e-03 1.799308e-03 4.498270e-03 6.159170e-03
              genres
Y              Reality-TV      Romance       Sci-Fi        Short        Sport    Talk-Show
  movie       9.535843e-05 1.294491e-02 5.244714e-03 0.000000e+00 1.191980e-03 0.000000e+00
  tvMiniSeries 2.309548e-02 2.447432e-02 8.962427e-03 1.034126e-03 1.103068e-02 4.481213e-03
  tvMovie     5.519779e-04 4.084637e-02 3.127875e-03 0.000000e+00 1.287948e-03 0.000000e+00
  tvSeries    1.150173e-01 1.487889e-02 7.335640e-03 1.384083e-03 8.166090e-03 3.660900e-02
              genres
Y                Thriller          War      Western
  movie       4.779841e-02 8.343862e-04 1.358858e-03
  tvMiniSeries 2.481903e-02 2.412961e-03 3.447087e-04
  tvMovie     4.967801e-02 5.519779e-04 5.519779e-04
  tvSeries    1.591696e-02 4.844291e-04 5.536332e-04
```

FONTE:O AUTOR (2020)

No atributo “genres” como mostra a Figura 24, dos 27 Gêneros as melhores probabilidades se concentram no “tvMovie” com 10 gêneros, 6 gêneros no “movie”, 6 gêneros no “series” e 5 gêneros no “tvminiserries”.

Através da aplicação do algoritmo o atributo “isAdult” tem a concentração das melhores probabilidades na “tvMiniSeries” conforme a Figura 25.

FIGURA 25 - PROBABILIDADES CONDICIONAL "ISADULT"

Y	isAdult [,1]	[,2]
movie	0.0006436694	0.02536278
tvMiniSeries	0.0041365047	0.06419357
tvMovie	0.0007359706	0.02712129
tvSeries	0.0032525952	0.05694067

FONTE:O AUTOR (2020)

De acordo com a Figura 25 observa-se que o melhor resultado foi no tipo não.

Através da aplicação do algoritmo o atributo “startYear” tem a concentração das melhores probabilidades no “TvMovie” conforme a Figura 26.

FIGURA 26 - PROBABILIDADES CONDICIONAL DO "STARTYEAR"

Y	startYear A1	A2	A3	A4	A5
movie	0.2056881	0.2120295	0.2163683	0.2022075	0.1637066
tvMiniSeries	0.2019993	0.2085488	0.2412961	0.1999311	0.1482248
tvMovie	0.2299908	0.2242870	0.2079117	0.1854646	0.1523459
tvSeries	0.2230450	0.2185467	0.2143945	0.1847059	0.1593080

FONTE: O AUTOR (2020)

No atributo “startYear”, como mostra a Figura 26, as melhores probabilidades se concentram no tvMovie e tvMiniSeries, e o melhor resultado das probabilidades foi no ano de 2017(A3).

Através da aplicação do algoritmo o atributo “averageRating” observa-se a concentração das melhores probabilidades no tvMiniSeries conforme a Figura 27.

FIGURA 27 - PROBABILIDADE CONDICIONAL DOS "AVERAGERATING"

Y	averageRating A	B	C	D	E
movie	0.020573581	0.329439531	0.469091949	0.155744153	0.025150786
tvMiniSeries	0.062736987	0.572905895	0.319544984	0.035849707	0.008962427
tvMovie	0.021711132	0.343698252	0.535418583	0.089788408	0.009383625
tvSeries	0.040553633	0.532249135	0.337923875	0.073840830	0.015432526

FONTE: O AUTOR (2020)

No atributo “averageRating”, como mostra a Figura 27, as melhores probabilidades se concentram no tvMiniSeries, e o tipo de classificação que teve a melhor probabilidade foi no Grupo B que varia de (5 a 6,9).

Através da aplicação do algoritmo o atributo “groupvotes” tem a concentração das melhores probabilidades no “movie” conforme a Figura 28.

FIGURA 28 - PROBABILIDADE CONDICIONAL DO "GROUPVOTE"

Y	groupvotes G1	G2	G3	G4
movie	0.2093833	0.2439269	0.2662169	0.2804730
tvMiniSeries	0.3233368	0.2736987	0.2182006	0.1847639
tvMovie	0.3000920	0.2840846	0.2206072	0.1952162
tvSeries	0.2926644	0.2803460	0.2313495	0.1956401

FONTE: O AUTOR (2020)

No atributo “groupvotes” como mostra a Figura 28 as melhores probabilidades se concentram no “movie”, os grupos G1, G2 são os que concentram a maior quantidade de votos.

Após a visualização dos resultados do algoritmo foram executados os scripts da Figura 29 os quais geraram a matriz confusão.

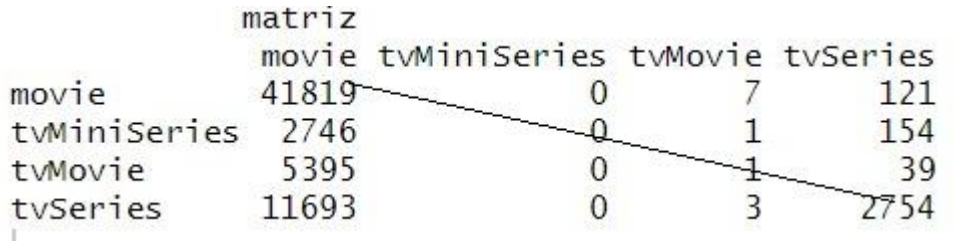
FIGURA 29 – SCRIPT DA MATRIZ CONFUSÃO DO NAIVE BAYES

```
#Matriz confusão do algoritimo Naive Bayes
matriz <- predict(Bayes, dados2)
table( dados2$titleType, matriz)
```

FONTE: O AUTOR (2020)

Através da Figura 29 foram executados os scripts apresentados da matriz confusão, onde executou-se a função *predict* utilizando o resultado da variável bayes, utilizando a tabela dados2. E na função *table* foi relacionado o resultado da matriz com o atributo meta gerando a matriz confusão conforme visto na Figura 30:

FIGURA 30 - MATRIZ CONFUSÃO DO NAIVE BAYES



matriz

	movie	tvMiniSeries	tvMovie	tvSeries
movie	41819	0	7	121
tvMiniSeries	2746	0	1	154
tvMovie	5395	0	1	39
tvSeries	11693	0	3	2754

FONTE:O autor (2020)

A matriz confusão como mostra a Figura 30, apresenta que “tvMiniSeries” e “tvMovie” teve 0 e 1 acertos respectivamente enquanto “movie” com 41.817 acertos e tvSeries com 2.754 acertos, realizando os cálculos de acerto: ((41.819) + (0) + (1) + (2.754)) / 64734 = 0,6885 = obteve-se 68% de taxa de acerto.

4.4 PÓS PROCESSAMENTO

Esta seção de pós processamento abordará a análise dos resultados da base de dados do IMDB com base no que foi realizado no pré-processamento, mineração de dados e estatística.

No capítulo de pré-processamento foram realizadas a verificação e a junção dos arquivos a fim de gerar a base de dados e a partir disso utilizando as quatro validações como valores ruidosos, inconsistentes, ausentes e redundantes. Dos 10 atributos presentes na base de dados, foram empregados 8 para o presente trabalho os quais são: tconst, titleType, originalTitle, isAdult, startYear, genres, averageRating e groupvotes. A base de dados ao todo continha 6.629.737 linhas e 10 variáveis, na limpeza de dados foi realizado o corte no atributo “startYear” onde utilizaria dados com anos iguais ou acima de 2015 resultando em 1.755.703 linhas e eliminando 6.459.876 títulos. No atributo “titleType” foram utilizados somente 4 tipos de tipos de título, “movie”, “tvMovie”, “tvSeries” e “tvminiSeries”, os demais foram cortados da base ficando a mesma com 169.861 linhas. Foi inconsistência no atributo “startYear” em uma base extraída com período fechado de 2019 continha registros futuros com dados de 2020 até 2115 da base, com isso o número de linhas foi para 156.818. Por fim, a última análise e corte foram realizados nos dados ausentes 92.085 títulos que não continham informações e foram eliminados da pesquisa, desta forma a base ficou com 64.733 dados na base final.

Logo após a realização dos devidos cortes na base com a limpeza dos dados realizou-se o tratamento dos mesmos. O agrupamento necessário tanto para a análise estatística da base quanto para a mineração, os atributos que necessitaram de agrupamento foram: “averageRating” e “numVotes”, com exceção do “startYear” que não foi agrupado, mas sim feito uma alteração para facilitar a leitura do algoritmo.

No capítulo da estatística foi possível observar que a instância “filmes” foi o tipo de título que foi mais produzido entre os anos de 2015 e 2019. E boa parte dos títulos produzidos correspondem à filmes não adultos, e o maior ano de produção de títulos foi em 2017 enquanto 2019 obteve a menor produção. Dos gêneros produzidos, drama foi o mais escolhido nas produções. Já em relação as notas atribuídas aos títulos 43,9% das produções obtiveram nota C (Nota entre 5 a 6,9) e dessas produções que tiraram nota C, 66% eram do gênero “adult”. E em relação a quantidade de votos que cada gênero recebeu, drama foi o que obteve maior quantidade de votos e o tipo de título com a maior quantidade de votos foi filmes.

No capítulo de mineração de dados foram aplicados dois algoritmos, a Árvore de Decisão utilizando o método rpart e o algoritmo Naive Bayes, aplicando o algoritmo da Árvore de Decisão, foi visto na matriz confusão na Figura 20 que a taxa de acerto foi de 70%, (39.920 + 0 + 0 + 5.782) /64.733 = 0, 706007. Já aplicando o algoritmo Naive Bayes, foi visto na matriz confusão na Figura 21 que a taxa de acerto foi de 68%, ((41.819) + (0) + (1) + (2.754)) / 64734 = 0,6885.

Logo percebe-se que apesar dos resultados próximos a árvore de decisão obteve um melhor resultado com 70%. Porém nota-se que a maior concentração dos títulos está presente no tipo “movies”.

5 CONSIDERAÇÕES FINAIS

Após os resultados obtidos com esse estudo, pode-se concluir que é desafiador para um Gestor da Informação trabalhar com bases grandes pois, para se realizar uma análise, necessitam que sejam agrupadas, realizado cortes, identificar e remover *outliers*, e realizar todos os tipos de limpeza na base para conseguir obter um bom resultado. Analisar os dados no setor de entretenimento de filmes é algo que está começando, como é o exemplo da Netflix que analisa os usuários e isso é muito motivador, poder ver a prática de algumas empresas analisando dados além de poder aplicar meus conhecimentos na área de dados. Pensando nestes cenários o presente trabalho surgiu com interesse de poder trabalhar com dados de algo sempre gostei de avaliar filmes e poder aplicar meus conhecimentos na área de dados.

Em relação aos objetivos, o objetivo geral de aplicar os métodos de análise de dados em uma base de dados da IMDB foi atingido como consequência dos objetivos derivados

O primeiro objetivo consistiu em realizar o levantamento bibliográfico e estudos do tema, o que foi realizado pois o levantamento nas bases do EBCOHOST e na CAPES foi o modo de provar que esse tema utilizado não tinha sido muito explorado na área acadêmica e como podemos observar no capítulo de justificativa, existem poucas pesquisas que tratam sobre o tema.

O segundo objetivo proposto pelo trabalho foi construir uma base de dados composta por um conjunto de arquivos do IMDB. Tal objetivo foi alcançado pois a base disponibilizada está dividida em sete arquivos sendo que dos sete arquivos foram utilizados somente dois e para fazer a junção dessa base. Para isso foi utilizado o programa em RStudio e no código foi utilizado a função “merge” que une todos os arquivos pelo atributo comum entre eles e sendo assim foi construída uma base com 6.629.737 linhas e 10 atributos.

O terceiro objetivo foi analisar a base de dados e selecionar os principais atributos e instâncias a serem analisados, esse objetivo foi concluído pois um dos critérios e delimitação da pesquisa consistiu na análise das notas dos filmes, os tipos de títulos como filmes, séries, minisséries e os gêneros daquele título, já os outros dados como diretores, elenco, atores e informações adicionais sobre o título não foram considerados para essa pesquisa. Feita essa delimitação, cinco arquivos não foram

utilizados para a criação da base pois não atendiam ao objetivo da pesquisa em analisar votos e classificação.

O quarto objetivo fundamentou-se na aplicação de estatísticas descritivas para a análise da base, esse objetivo foi alcançado pois de início do estudo foi feita a identificação da classificação dos atributos, “titleType” como variável qualitativa nominal, “isAdult” como variável booleana, “startYear” como uma variável qualitativa ordinal, “Genres” como variável qualitativa nominal, “averageRating” como uma variável qualitativa ordinal e “groupvotes” como variável qualitativa ordinal. Dessa forma, verificando esses atributos, percebeu-se que muito dos títulos possuíam mais de um gênero. E tiveram que ser unificados para ser possível fazer uma análise, muitos títulos possuíam anos de lançamentos futuros, os números de votos dos títulos possuíam uma variação de votos que se concentravam num só grupo e teve que ser feito o agrupamento por código no RStudio.

O quinto objetivo foi definir os dois algoritmos a partir da compreensão dos dados e as possibilidades que os diferentes algoritmos poderiam representar, tanto em sua visualização quanto na heurísticas escolhidas. Para esta pesquisa foram utilizados a Árvore de Decisão e o Naive Bayes.

Por fim, o sexto objetivo consistiu na aplicação dos algoritmos de mineração de dados na base de dados.

Este estudo respondeu a problematização apresentada no início do trabalho, pois os algoritmos conseguiram prever padrões e tendências das avaliações, porém foi necessário utilizar o agrupamento, pois ficaria muito disperso o resultado e com isso, os dois algoritmos utilizados obtiveram um acerto de aproximadamente 70% o que nos leva a perceber que a taxa de acerto relativamente alta.

Ao realizar esse trabalho, foi observado que apesar da mineração de dados estar em alta atualmente, muito pouco se estuda sobre isso na parte de entretenimento de filmes. Fato esse apontado na justificativa, já que foram poucos os estudos sobre esse tema. Dessa forma, este trabalho contribui para o uso da mineração de dados na parte do entretenimento e foi observado que realizar análises de dados na área é algo em crescimento exponencial, e que muitas empresas veem a influência da análise de dados na tomada de decisão, e que conseguem obter uma boa vantagem competitiva como nos casos da Netflix, Marvel Studios, Amazon, Disney+ entre outros.

Por fim sugere-se que a utilização de outros algoritmos, ferramentas, métodos e comparações de resultados como para possíveis trabalhos futuros.

REFERÊNCIAS

AMARAL, Fernando. **Aprenda mineração de dados.** Rio de Janeiro: Alta Books, 2016. 225 p.

ASSIS, Wilson Martins de. **Gestão da informação nas organizações:** como analisar e transformar em conhecimento informações captadas no ambiente de negócios. Autêntica, 2008. 140 p.

AVANCHA, Sasikanth; KALLURKAR, Srikanth; KAMDAR, Tapan. **Design of ontology for the Internet Movie Database (IMDb).** 2001. 9 p.

BEAL, A. Gestão estratégica da informação. São Paulo: Atlas, 2004.

BERTONCINI, Cristine; BRITO, Asriana; SILVA, Ismael. **Processo decisório: a tomada de decisão.** São Paulo: Faeg/aceg,. 12 p.

BOENTE, Alfredo Nazareno Pereira; OLIVEIRA, Fabiano Saldanha Gomes de; ROSA, José Luiz dos Anjos. **Utilização de ferramentas de kdd para integração de aprendizagem e tecnologia em busca da gestão estratégica do conhecimento na empresa.** 12 f. Monografia (Especialização) - Centro Universitário Estadual da Zona Oeste, Rio de Janeiro.

CARAVANTES, Geraldo; PANNO, Cláudia; KLOECKNER, Mônica. **Administração:** teorias e processo. São Paulo: Pearson, 2005. 572 p.

CARDOSO, Olinda Nogueira Paes. **Recuperação da informação.** 6 f. Monografia (Especialização) - Curso de Ciência da Computação, Ciência da Computação, Universidade Federal de Lavras, Lavras.

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. **introdução à mineração de dados:** Conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva Educação, 2016. 559 p.

CHIAVENATO, Idalberto. **Introdução à Teoria Geral da Administração.** 7. ed. Rio de Janeiro: Campus, 2003. 630 p.

CHMIELEWSKI, Dawn C.; TIMES, Los Angeles. **Col Needham created IMDb.**2013. Disponível em: <https://www.latimes.com/business/la-xpm-2013-jan-19-la-fi-himi-needham-20130120-story.html>. Acesso em: 22 abr. 2019.

CHOO, Chun Wei. **A organização do conhecimento:** como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. Tradução Eliana Rocha. - São Paulo: Editora Senac São Paulo, 2003.

DAVENPORT, Thomas H. **Ecologia da informação:** por que só a tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 1998. 312 p. (ISBN 85-86082-72-4) Tradução Bernadette Siqueira Abrão

DUARTE, Emeide Nóbrega; PAIVA, Simone Bastos; SILVA, Alzira Karla Araújo da. **Múltiplas Abordagens da Gestão da Informação e do Conhecimento no Contexto Acadêmico da Ciência da Informação.** João Pessoa: Editora da Ufpb, 2014. 168 p.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas S.a, 2008. 220 p.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining:** Um Guia Prático. 4. ed. Rio de Janeiro: Elsevier, 2005.

MASCARELLO, Fernando. **História do cinema mundial.** São Paulo: Papirus Editora, 206. 430 p.

MAXIMIANO, A. C. A. **Introdução à administração**. 5. ed. rev. e ampl. – São Paulo:Atlas, 2000.

MORAES, G. D. A.; FILHO, E. E. A gestão da informação diante das especificidades das pequenas empresas. **Ci. Inf**., Brasília, v. 35, n. 3, p. 124-132, set./dez. 2006. Disponível em: http://www.scielo.br/pdf/ci/v35n3/v35n3a12.pdf. Acesso em: 11 mai. 2019.

MOREIRA, Cleverson Bayer. **Gestão da informação.** Guarapuava: Gráfica Unicentro, 2014.

NEVES, Rita de Cássia David das. **Pré-Processamento no Processo de Descoberta de Conhecimento em Banco de Dados.** 2003. 137 f. Tese (Mestrado) - Curso de Ciência da Computação, Instituto de InformÁtica, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003.

OLIVEIRA, Marcus. **Cultura data driven:** Entenda como a Marvel e Netflix produzem conteúdos orientados a dados. 2018. Disponível em: <https://inteligencia.rockcontent.com/cultura-data-driven/>. Acesso em: 21 jun. 2019.

PACHECO, Liliana. **Marketing, recepção e crítica cinematográfica na era digital**. Lisboa: Instituto Universitário de Lisboa, 2012, p. 351 - 365.

PONTES JUNIOR, João de; CARVALHO, Rodrigo Aquino de; AZEVEDO, Alexander William. **Da recuperação da informação à recuperação do conhecimento:** Reflexões e propostas. 2013. 16 f. Tese (Doutorado) - Curso de Ciência da Informação, Universidade Católica de Campinas, Campinas, 2010.

REZENDE, Y. **Informação para negócios:** os novos agentes do conhecimento e a gestão do capital intelectual. Ciência da Informação, Brasília, v. 31, n. 2, p. 120-128, maio/ago. 2002.

RODRIGUES, Charles; BLATTMANN, Ursula. **Gestão da informação e a importância do uso de fontes de informação para geração de conhecimento.** 2014. 29 f. Tese (Doutorado) - Curso de Engenharia de Produção, Ciência da Informação, Universidade Federal de Santa Catarina, Belo Horizonte, 2013.

SANTOS, Adam Marcel de Oliveira. **Marketing e Cinema:** Estratégias de Lançamento de Filmes. 2016. 48 f. TCC (Graduação) - Curso de AdministraÇÃo, Departamento de Administração, PontifÍcia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2016.

SILVA, Edna Lucia da; MENEZES, Estera Muszkat. **Metodologia da Pesquisa e Elaboração de Dissertação.** 4. ed. Florianópolis: Ufsc, 2005. 139 p.

TOPAL, Kamil; OZSOYOGLU, Gultekin. Movie Review Analysis: Emotion Analysis of IMDb Movie Reviews. **2016 ieee/acm Asonam**, Sam Francisco, p.1-7, 21 2016.

SHOPPING, Redação Jardim Pamplona. **Tipos de filmes**: quais são os melhores de cada gênero?.2020. Disponível em: https://jardimpamplonashopping.com.br/tipos-de-filmes/. Acesso em: 03 nov. 2020.

LUDIASBH. **OS GÊNEROS DO CINEMA**. 2020. Disponível em: https://virusdaarte.net/os-generos-do-cinema/. Acesso em: 03 nov. 2020.

CINEMA10. **Os Melhores Filmes de Família**. 2020. Disponível em: https://cinema10.com.br/generos/filmes-de-familia. Acesso em: 03 nov. 2020.

CINEMA10. **Os Melhores Filmes de Mistério**. 2020. Disponível em: https://cinema10.com.br/generos/filmes-de-misterio. Acesso em: 03 nov. 2020.

CINEMA10. **Os Melhores Filmes sobre Esportes**. 2020. Disponível em: https://cinema10.com.br/generos/filmes-sobre-esporte. Acesso em: 03 nov. 2020.

CINEMA10. **Os Melhores Filmes de Suspense/Thriller**. 2020. Disponível em: https://cinema10.com.br/generos/filmes-de-suspense. Acesso em: 03 nov. 2020.

CINEMA10. **Os Melhores Filmes de Guerra**. 2020. Disponível em: https://cinema10.com.br/generos/filmes-de-guerra. Acesso em: 03 nov. 2020.

SANTOS, Agenor Soares dos. **TALK SHOW: conheça significado, pronúncia e tradução de talk show**!. 2020. Disponível em: https://www.teclasap.com.br/talk-show/. Acesso em: 10 nov. 2020.

SANTOS, Agenor Soares dos. **REALITY SHOW: o que significa esse anglicismo?**. 2020. Disponível em: https://www.teclasap.com.br/reality-show/. Acesso em: 10 nov. 2020.

LTDA, Travelbr Turismo. **Telejornalismo**. 2013. Disponível em: https://www.jornalista.com.br/telejornalismo.html. Acesso em: 10 nov. 2020.

FILMEFEED. **O que é um filme curta-metragem**. 2019. Disponível em: https://filmefeed.telecineplay.com.br/o-que-e-um-filme-curta-metragem/p. Acesso em: 10 nov. 2020.

GUEDES, T.A.; ACORSI, C.R.L.; MARTINS, A.B.T.; JANEIRO, V. **Projeto de Ensino: Aprender Fazendo Estatística.** Disponível em: < http://www.each.usp.br/rvicente/Guedes_etal_Estatistica_Descritiva.pdf>. Acesso em: 08 de fevereiro. 2021.

BECKER, Lauro. **Algoritmo de Classificação Naive Bayes**. 2019. Disponível em: https://www.organicadigital.com/blog/algoritmo-de-classificacao-naive-bayes/. Acesso em: 18 jan. 2021.

CAMPOS, Raphael. **Árvores de Decisão**. 2017. Disponível em: https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 18 jan. 2021.

CANALTECH. **IMDB**. 2020. Disponível em: https://canaltech.com.br/empresa/imdb/. Acesso em: 21 jan. 2021.

APÊNDICE

APÊNDICE A - TIPOS DE GÊNEROS DE FILMES E SÉRIES

Gênero	Tradução	Descrição
Action	Ação	De acordo com LUDIASBH (2020), “Contam histórias envolvendo um protagonista contra um antagonista.”
Adult	Adulto	De acordo com LUDIASBH (2020), “Também conhecido como cinema pornográfico, apareceu rapidamente depois da criação da tecnologia de filmes, que fez com que esse tipo de filme fosse possível.”
Adventure	Aventura	De acordo com LUDIASBH (2020), “A história contada envolve o percurso de uma personagem ao longo do filme, passando por diversas situações e desafios para cumprir um objetivo.”
Animation	Animação	De acordo com LUDIASBH (2020), “É um filme ou série em que cada fotograma de um filme é produzido individualmente, podendo ser gerado quer por computação gráfica, ou fotografando uma imagem desenhada repetidamente”
Biograph	Biografia	De acordo com SHOPPING (2020), “Conta a história de pessoas reais.”
Comedy	Comédia	De acordo com LUDIASBH (2020), “Filmes que trazem o humor como característica principal do enredo.”
Crime	Crime	De acordo com LUDIASBH (2020), “Também conhecido como gênero policial, os argumentos quase sempre envolvem crimes e criminosos, policiais e detetives particulares, gangsteres e ladrões.”

Documentary	Documentário	De acordo com LUDIASBH (2020), “É um gênero cinematográfico que se caracteriza pelo compromisso com a exploração da realidade.”
Drama	Drama	De acordo com LUDIASBH (2020), “Gênero que possuem histórias que envolvem conflitos de sentimentos, mergulhar na vida das personagens e entender os sentimentos vividos por elas ao passar pelas situações apresentadas na história.”
Family	Família	De acordo com CINEMA10 (2020), “o gênero nada mais é do que uma história que pode ser vista por todas as idades, mas abrange assuntos mais amplos do que a temática infantil.”
Fantasy	Fantasia	De acordo com LUDIASBH (2020), “História onde elementos mágicos e sobrenaturais são uma das características principais.”
Game-Show	Game-Show	De acordo com a Wikipédia, é um gênero de programa de televisão onde pessoas comuns ou celebridades, em equipes ou não, participam numa prova que pode incluir testes de inteligência e/ou provas físicas com o objetivo de ganhar pontos ou prêmios.
History	Histórico	De acordo com SHOPPING (2020), “Esse tipo de filme aborda fatos históricos, trazendo representações, culturas e fatos sobre determinadas épocas.”
Horror	Terror	De acordo com SHOPPING (2020), “Apresenta histórias de ficção com a intuição de causar medo aos espectadores, utilizando efeitos especiais, situações e seres sobrenaturais como recurso.”
Music	Música	De acordo com SHOPPING (2020),” este gênero se encaixa muito mais por shows ao vivo e videoclipes.”

Musical	Musical	De acordo com LUDIASBH (2020), “Possuem uma narrativa leve e divertida, utilizando sequências de músicas e coreografias como apoio para contar uma história ao longo do filme.”
Mystery	Mistério	De acordo com CINEMA10 (2020), “o mistério foi criado pela junção de diversos outros gêneros e temáticas, se inspirando em parte da essência do suspense e na literatura instigante das histórias de agentes resolvendo grandes casos.”
News	Notícia	De acordo com LTDA (2013), “é também conhecido como telejornalismo as notícias podem ser relatadas sob vários formatos, como nota simples, nota coberta e reportagens, a forma mais completa de apresentar a notícia.”
Reality-Tv	Reality-TV	De acordo com Santos (2020),” é um programa de televisão em que se reúnem pessoas para mostrar cenas reais, seus diálogos e convivência.”
Romance	Romance	De acordo com LUDIASBH (2020), “Possui um enredo que gira em torno do relacionamento amoroso entre os protagonistas da história.”
Sci-Fi	Ficção cientifica	De acordo com LUDIASBH (2020), “Histórias que envolvem ciência, tecnologias e sociedades em um tempo futuro.”
Short	Curtas	De acordo com FILMEFEED(2019), “O gênero curtas ou curta-metragem tem uma abordagem simples e direta, e abusa da criatividade para passar uma mensagem da forma mais rápida possível.”
Sport	Esporte	De acordo com CINEMA10 (2020), “nesse gênero, o esporte se manifesta com belas histórias de superação e união geradas através do esporte.”

Talk-Show	Talk-Show	De acordo com Santos (2020), “é um programa de televisão, em que um apresentador-anfitrião conversa com pessoas de renome, ligadas a algum fato importante do momento, ou cuja atividade pode ter interesse para o público”.
Thriller	Suspense	De acordo com CINEMA10 (2020), “é conhecido como suspense no Brasil e thriller na língua inglesa, um filme desse gênero se caracteriza pela constante tensão na história.”
War	Guerra	De acordo com CINEMA10 (2020),” os filmes e séries de guerra são aqueles que abordam conflitos pontuais reais já ocorridos ou ficcionais. Com muita ação, explosões e mortes, sempre um cenário trágico.”
Western	Faroeste	De acordo com LUDIASBH (2020), “Filmes e séries que se referem à fronteira do Oeste norte-americano durante a colonização, também popularizado sob os termos “filmes de cowboys” ou “filmes de faroeste”.”

FONTE: IMDB (2020)

APÊNDICE B - RELAÇÃO DE GÊNEROS POR ANOS

Rótulos de Linha	A1	A2	A3	A4	A5	Total Geral
Drama	4707	4809	4878	4594	4030	23018
Documentary	3164	3092	3091	2699	1959	14005
Comedy	2504	2601	2446	2197	1672	11420
Horror	663	709	778	682	555	3387
Thriller	452	498	549	567	511	2577
Action	353	355	422	337	279	1746
Reality-TV	469	389	366	281	231	1736
Romance	164	202	197	268	220	1051
Animation	163	219	187	198	156	923
Family	223	144	217	145	123	852
Crime	105	133	152	197	203	790

Talk-Show	142	124	134	82	60	542
Sci-Fi	73	63	105	68	60	369
Mystery	56	80	60	56	40	292
Adventure	43	55	67	50	53	268
Biography	61	42	67	48	34	252
History	42	70	44	43	30	229
Game-Show	45	45	52	46	31	219
Music	72	39	30	23	46	210
Sport	43	57	35	32	40	207
Fantasy	35	45	37	39	32	188
Musical	49	32	20	30	30	161
News	13	27	27	22	8	97
Western	13	12	20	14	10	69
War	14	7	6	15	10	52
Adult	15	19	12	4	0	50
Short	4	8	5	2	4	23
Total Geral	**13687**	**13876**	**14004**	**12739**	**10427**	**64733**

FONTE: O AUTOR (2020)

APÊNDICE C- PORCENTAGEM DE NOTAS RECEBIDAS POR CADA GÊNERO

Rótulos de Linha	A	B	C	D	E	Total Geral
Action	28	563	766	309	80	1746
Adult	1	13	33	3	0	50
Adventure	25	100	99	37	7	268
Animation	28	379	376	116	24	923
Biography	9	131	94	14	4	252
Comedy	270	3602	5250	1983	315	11420
Crime	12	317	381	70	10	790
Documentary	746	8542	4260	400	57	14005
Drama	357	8087	11933	2367	274	23018
Family	28	243	442	127	12	852
Fantasy	2	81	69	24	12	188
Game-Show	8	58	88	59	6	219
History	8	135	71	13	2	229
Horror	9	403	1260	1366	349	3387
Music	14	103	74	14	5	210
Musical	2	77	62	18	2	161
Mystery	5	112	143	30	2	292
News	8	29	30	26	4	97
Reality-TV	74	712	679	231	40	1736

Romance	15	316	603	105	12	1051
Sci-Fi	7	131	121	79	31	369
Short	7	9	4	3	0	23
Sport	12	111	67	15	2	207
Talk-Show	56	203	181	72	30	542
Thriller	13	547	1271	683	63	2577
War	3	21	15	12	1	52
Western	2	15	25	16	11	69
Total Geral	**1749**	**25040**	**28397**	**8192**	**1355**	**64733**

FONTE: O AUTOR (2020)