UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS RUSSAS

CURSO DE GRADUAÇÃO EM ENGENHARIA DE SOFTWARE

JOSÉ ULISSES SILVA MACEDO OLIVEIRA

APLICAÇÃO DE UM ALGORITMO DE CLUSTERIZAÇÃO EM BASES DE DADOS DE PLATAFORMAS DE *STREAMING*

RUSSAS

2022

JOSÉ ULISSES SILVA MACEDO OLIVEIRA

APLICAÇÃO DE UM ALGORITMO DE CLUSTERIZAÇÃO EM BASES DE DADOS DE PLATAFORMAS DE *STREAMING*

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Orientadora: Prof. Dra. Tatiane Fernandes Figueiredo

RUSSAS

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O47a Oliveira, José Ulisses Silva Macedo.
Aplicação de algoritmos de clusterização em bases de dados de plataformas de streaming / José Ulisses Silva Macedo Oliveira. – 2022.
35 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Russas, Curso de Engenharia de Software, Russas, 2022.
Orientação: Prof. Dr. Tatiane Fernandes Figueiredo.

1. Streaming. 2. Clusterização. 3. K-Means. 4. Mineração de Dados. I. Título.

CDD 005.1

JOSÉ ULISSES SILVA MACEDO OLIVEIRA

APLICAÇÃO DE UM ALGORITMO DE CLUSTERIZAÇÃO EM BASES DE DADOS DE PLATAFORMAS DE *STREAMING*

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Aprovada em:

BANCA EXAMINADORA

Prof. Dra. Tatiane Fernandes Figueiredo (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Bonfim Amaro Júnior
Universidade Federal do Ceará (UFC)

Prof. Dr. Pablo Luiz Braga Soares
Universidade Federal do Ceará (UFC)

À minha família, minha Irmã, minha Mãe e meu Pai, aos meus Avós e todas as pessoas que passaram pela minha vida e acreditaram em mim.

AGRADECIMENTOS

Deixo aqui os meus sinceros agradecimentos a minha família, principalmente minha mãe Maria de Lourdes, a minha irmã Maria Luiza que já me irritou e ainda vive me irritando, mas que amo e as minhas tias Elismar dos Santos e Francisca da Silva que cuidaram de mim e até hoje cuidam. Não posso esquecer do meu pai Eulicio Macedo que também sempre esteve ali para mim.

Agradeço a todos os meus professores do ensino fundamental ao ensino médio, que foram responsáveis pela minha formação.

Agradeço a todas as pessoas que participaram da minha vida, acrescentando ou não, já que foram com essas experiências que me fizeram ser quem eu sou agora.

Agradeço a Barbara e Tereza, amigas a tanto tempo que não tenho palavras para agradecer. Deixo o agradecimento ao Gustavo Girão, André Vinicius, Jeferson Ribeiro, Francisco Eudes, Vicente Augusto, Lucas Emanuel, Isaac Freitas, Vitoria Helen, Josué Lamec e Fábio Fiuzza por terem me acompanhado durante a jornada que foi a faculdade. Adiciono também ao meus queridos Mateus Araújo, Levi Miquéias e Renan Arthur por serem meus colegas de trabalho e meus amigos. Agradeço também a todos os meus amigos e amigas que não foram citados aqui.

Agradeço também a orientação da professora Tatiane Fernandes que esteve comigo desde do início da pesquisa, me incentivando e auxiliando nas minhas decisões. Foi uma inspiração e companheira em todos os projetos que participei em conjunto, desde o LED, Tilapia e agora no meu projeto de pesquisa.

E por fim agradeço a UFC por existir e por proporcionar o investimento necessário a minha formação acadêmica e social. Em especial agradeço a todos integrantes LED, principalmente aos professores pelas experiências e conhecimentos transmitidos aos alunos.

"O sonho é que leva a gente para frente. Se a gente for seguir a razão, fica aquietado, acomodado."

(Ariano Suassuna)

RESUMO

Com a popularização da internet e consequentemente o uso das plataformas *streaming*, o número de empresas que prestam esse serviço tem crescido a cada dia. Para manter seus clientes ativos, essas empresas têm investido em funcionalidades que podem apresentar opções de filmes de acordo com as preferências e o histórico do usuário. Buscando entender como esses algoritmos funcionam, esta monografia propõe um estudo de caso que utiliza bancos de dados de plataformas *streaming*. Este estudo usará o algoritmo *K-Means* para criar agrupamentos a fim de detectar padrões e gerar *insights* sobre seus usuários e filmes.

Palavras-chave: *streaming*; clusterização; *k-means*; mineração de dados;

ABSTRACT

With the internet popularization and consequently the use of *streaming* platforms, the number of companies that provide this service has grown every day. In order to keep their customers active, these companies have invested in functionalities that can present movie options according to the user's preferences and history. Seeking to understand how these algorithms work, this monograph proposes a case study that uses databases from *streaming* platforms. This study will use the *K-Means* algorithm to create clusters in order to detect patterns and generate *insights* about its users and movies.

Keywords: streaming; clustering algorithm; k-means; data mining

LISTA DE FIGURAS

Figura 1 – Etapas do Processo de KDD 16
Figura 2 – Tipos de Aprendizados 18
Figura 3 – Classificação e Regressão 19
Figura 4 – *Clustering* 20
Figura 5 – Execução do *K-Means* 21
Figura 6 – Representação da metodologia proposta 25
Figura 7 – Quantidade de shows/filmes por *clusters* 27
Figura 8 – *Cluster* 0: Gêneros 28
Figura 9 – *Cluster* 1: Gêneros 28
Figura 10 – *Cluster* 2: Gêneros 29
Figura 11 – *Cluster* 3: Gêneros 30
Figura 12 – Anos de lançamentos de produções nos *Streaming* e seus agrupamentos por *clusters* 31
Figura 13 – Pontuação IMDB X Anos de lançamentos por *clusters* 31
Figura 14 – Pontuação IMDB X Tempo de Execuçãos por *clusters* 32

LISTA DE QUADROS

Quadro 1 – Formato dos Dados 26
Quadro 2 – Quantidade de registros por *streaming* no *cluster* 0 e seus tipos. 27
Quadro 3 – Quantidade de registros por *streaming* no *cluster* 3 e seus tipos. 29

LISTA DE ABREVIATURAS E SIGLAS

DBI	*Davies Bouldin Index*
EM	*Elbow Method*
IA	Inteligência Artificial
IMDB	*Internet Movie Database*
KDD	*Knowledge Discovery in Databases*
KDT	*Knowledge Discovery from Text*
LGS	*Latent Genre Space*

SUMÁRIO

1 INTRODUÇÃO 13
2 OBJETIVOS 15
2.1 Objetivos Gerais 15
2.2 Objetivos específicos 15
3 FUNDAMENTAÇÃO TEÓRICA 16
3.1 ***Descoberta de conhecimento em bases de dados*** 16
3.1.1 ***Identificação do problema*** 17
3.1.2 ***Preparação dos Dados*** 17
3.1.3 ***Mineração dos Dados*** 17
3.1.4 ***Pós-processamento*** 18
3.2 Aprendizado de máquina 18
3.2.1 ***Aprendizado Supervisionado*** 19
3.2.2 ***Aprendizado Não Supervisionado*** 19
3.3 Algoritmos de clusterização 20
3.3.1 ***K-Means*** 21
4 TRABALHOS RELACIONADOS 22
4.1 ***Prevendo a preferência do usuário para filmes usando o banco de dados Netflix*** 22
4.2 Análise de dados no *Internet Movie Database* (IMDb) 23
4.3 Aplicação de mineração de dados com o método de agrupamento *K-Means* e índice *Davies Bouldin* para agrupamento de filmes IMDB . . . 23
5 METODOLOGIA 25
5.1 Tratamento e Análise da base de dados 25
5.2 Geração de Modelos Inteligentes 26
6 RESULTADOS 27
6.1 ***Cluster* 0** 27
6.2 ***Cluster* 1** 28
6.3 ***Cluster* 2** 28
6.4 ***Cluster* 3** 29
6.5 Análises gráficas gerais 30

7 CONCLUSÕES E TRABALHOS FUTUROS 33
7.1 Considerações gerais 33
7.2 Trabalhos futuros 33
REFERÊNCIAS 34

1 INTRODUÇÃO

Com a popularização da internet e consequentemente o uso de plataformas de *streaming*, o número de empresas que disponibilizam este serviço tem crescido a cada dia mais. Desde de o lançamento do serviço de *streaming* da *Netflix* em 2007, houve uma "revolução"na maneira do consumo de produções audiovisuais, onde diversas empresas construíram suas plataformas no mesmos moldes da veterana *Netflix*, como é o caso da plataforma *HBO Max* criada em 2021.

Para conquistar mercado e manter seus clientes ativos nestas plataformas, estas empresas tem apostado em algoritmos inteligentes que possam apresentar recomendações de filmes para os usuários de acordo com suas preferências, avaliações e histórico de uso na plataforma. Essas recomendações e recursos dependem diretamente da análise de uma grande quantidade de dados, que em sua maioria são obtidas através de técnicas de mineração de dados, algoritmos de clusterização e outros algoritmos de aprendizado de máquina.

É um fato que os dados, algoritmos e análises realizadas pelas empresas de *streaming* mencionadas não são compartilhadas de forma pública e portanto, bases de dados públicas de plataformas como a *Netflix* e *HBO Max* ainda são raras. Por este motivo, a grande maioria dos trabalho relacionados à este tema, ainda estão concentrados em analisar preferências de usuários apenas utilizando a base de dados do site IMDB (GOEL; BATRA, 2009). Como exemplo, pode-se citar o trabalho de Cardoso (2021) que apresenta a análise de uma base dados pública IMDB. Neste trabalho, os autores buscam por padrões de produção de conteúdo, enquanto o trabalho de Ashari *et al.* (2022) apresenta a geração de agrupamentos, assim como o uso de técnicas como o *Davies Bouldin Index* também aplicada a bases de dados pública do IMDB.

Buscando remover essa lacuna, esta monografia apresenta a aplicação do algoritmo de clusterização *K-Means*, em dois conjuntos de dados relacionados as plataformas de *streaming*: *Netflix* e *HBO Max*, respectivamente. Para tal, a metodologia utilizada foi dividida em 3 etapas, sendo a primeira: o tratamento e análise de bases de dados, onde os dados utilizados foram padronizados para se adequar as etapas posteriores. Na segunda etapa, foi aplicado o algoritmo de clusterização *K-Means*, de modo a definir agrupamentos com o objetivo de gerar *insights* sobre as bases de dados estudadas. Por fim, foi realizada uma análise dos resultados, investigando se os agrupamentos realizados e *insights* obtidos foram satisfatórios para o objetivo proposto.

A estrutura deste trabalho encontra-se da seguinte forma. No Capítulo 2 é apresentado o objetivo geral e os objetivos específicos deste trabalho; no Capítulo 3 expõe os pontos

chaves para a teoria da pesquisa; enquanto o Capítulo 4 apresenta os trabalhos que se possuem algum aspecto semelhante com este trabalho. No Capítulo 5 descreve os procedimentos utilizados para realizar a pesquisa e assim como seu desenvolvimento, o Capítulo 6 apresenta os resultados obtidos e por fim o Capítulo 7 sintetiza a conclusão e os possíveis trabalhos futuros.

2 OBJETIVOS

Nesta seção é apresentado o objetivo principal deste trabalho que será atingido com o desenvolvimento das etapas descritas nos objetivos específicos.

2.1 Objetivos Gerais

Utilizar um algoritmo de clusterização para geração de *insigths* em base de dados pública de plataformas de *streaming Netflix* e *HBO Max*.

2.2 Objetivos específicos

- Estudar e compreender as bases de dados públicas disponíveis de *streaming* das empresas *Netflix* e *HBO Max*;
- Normalizar e padronizar as bases de dados;
- Utilizar um algoritmo clusterização para geração de agrupamentos de dados em cada uma das bases estudadas;
- Analisar os resultados obtidos;

3 FUNDAMENTAÇÃO TEÓRICA

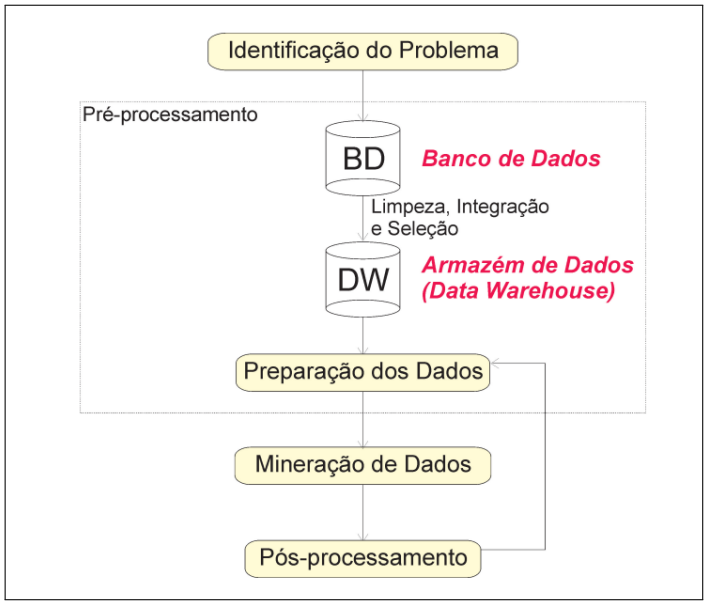
Para a compreensão e execução desta monografia, neste capítulo são apresentados os principais conceitos utilizados durante toda a pesquisa científica realizada. Na Seção 3.1 é apresentada as principais técnicas para descoberta de novos conhecimento em bases de dados. Na Seção 3.2 é definido de forma geral o funcionamento de algoritmos de Aprendizado de Máquina Não Supervisionados. Na Seção 3.3 é apresentado o algoritmo de clusterização que foi utilizado neste trabalho.

3.1 *Descoberta de conhecimento em bases de dados*

De acordo com Castanheira (2008), o processo de descoberta de conhecimento em bases de dados, conhecido como KDD (do inglês *Knowledge Discovery in Databases*) tem como objetivo principal extrair conhecimento a partir de grandes bases de dados. Para atingir esse objetivo é necessário o envolvimento de diversas áreas do conhecimento, como técnicas matemáticas e estatísticas, uso de banco de dados, técnicas para visualização dos dados e entre outras.

Ao se executar a metodologia KDD é necessário ter uma contextualização clara do domínio do problema, assim como qual objetivo deseja-se conquistar. De acordo com Morais e Ambrósio (2007) o processo possui quatro principais etapas como podem ser vistas na Figura 1 a seguir:

Figura 1 – Etapas do Processo de KDD



Fonte: Morais e Ambrósio (2007).

3.1.1 Identificação do problema

Para Morais e Ambrósio (2007), nesta fase deve ser realizado um estudo do domínio da aplicação, e a definição dos objetivos e metas a serem alcançados. Além disso, é também nesta etapa que deve ser obtido os dados que serão utilizados nas etapas de descoberta de informações.

3.1.2 Preparação dos Dados

A preparação dos dados é uma etapa complexa do ponto de vista de atividades que precisam ser realizadas. É nesta etapa que deve ser aplicada técnicas para limpeza, normalização, integração e/ou seleção dos dados mais relevantes para o objetivo pretendido. É comum que os dados obtidos inicialmente não estejam padronizados para a extração de conhecimento (MORAIS; AMBRÓSIO, 2007). A partir disso é preciso analisar qual processo de preparação deve ser utilizado. Para a etapa de mineração de dados, algumas vezes se faz necessária a transformação dos dados, de modo a se adequarem as limitações existentes do algoritmo utilizado.

3.1.3 Mineração dos Dados

O termo Mineração de Dados é muitas vezes utilizada como sinônimo para o KDD, mas para Fayyad *et al.* (1996), KDD é a metodologia completa para descoberta de conhecimento, enquanto a Mineração de Dados refere-se a uma etapa específica desta metodologia. Desta forma, a Mineração de Dados dentro da metodologia KDD é o processo de aplicação de técnicas estatísticas e algoritmos inteligentes que serão utilizados para encontrar padrões e novas informações relevantes para o problema em estudo. É importante mencionar que utilização cega de métodos de Mineração de Dados é uma atividade perigosa, já que facilmente leva a descoberta de informações sem sentido ou padrões inválidos (FAYYAD *et al.*, 1996).

Para Castanheira (2008), a Mineração de Dados difere de uma simples aplicação de técnicas estatísticas porque ao invés de verificar padrões hipotéticos, utiliza os próprios dados para descobrir tais padrões. A obtenção destes possíveis padrões estão diretamente relacionadas a execução dos algoritmos inteligentes, mas é preciso executa-lós de maneira interativa buscando ajusta-lós e muitas vezes utilizando mais de um único algoritmo, já que pode não ser trivial a escolha de qual método será a melhor opção a ser utilizada.

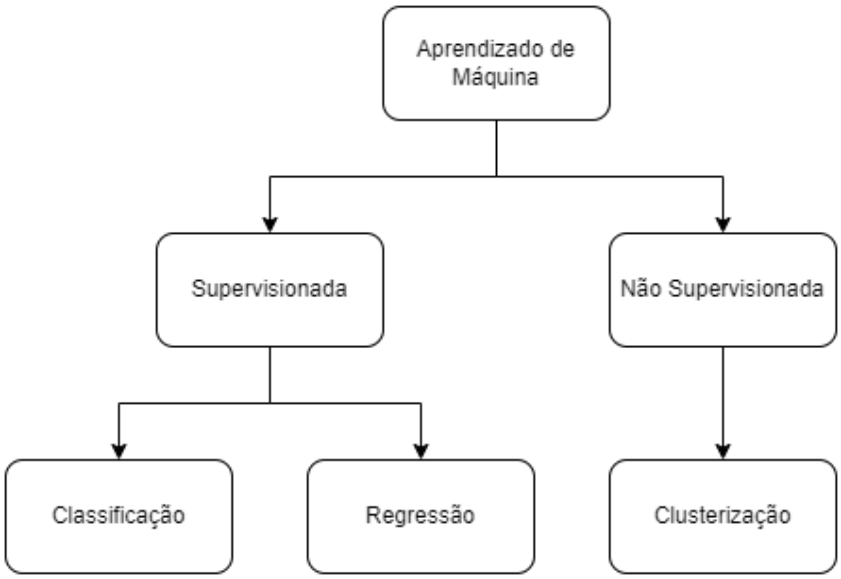
3.1.4 Pós-processamento

Os resultados obtidos pelo KDD podem ser visualizados de diferentes formas, já que a execução dos processos provavelmente irão gerar uma grande quantidade de padrões, dos quais muitos poderão não ser importantes ou interessantes para o propósito definido. Para lidar com essa fase é necessário a execução de uma análise criteriosa do conjunto de resultado com medidas desempenho, qualidade ou medidas subjetivas. No caso do resultado final não ser adequado para o objetivo, todo o processo poderá ser repetido indefinidas vezes até que o retorno se torne satisfatório.

3.2 Aprendizado de máquina

De acordo com Monard e Baranauskas (2003), o Aprendizado de Máquina é uma área de Inteligência Artificial (IA) cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática. Desta forma, o Aprendizado de Máquina pode ser divido em duas categorias principais, a Aprendizagem Supervisionada e a Não Supervisionada, as duas categorias serão expandidas nas seções subsequentes.

Figura 2 – Tipos de Aprendizados



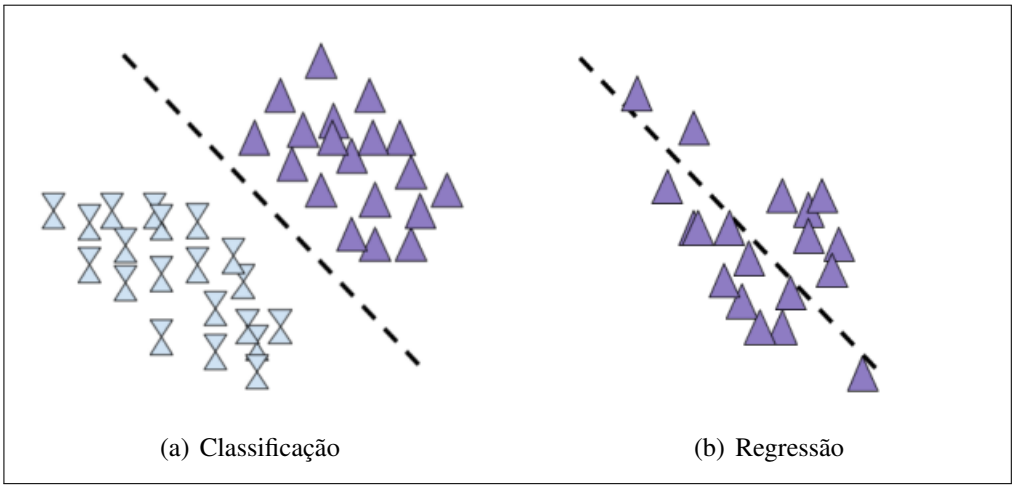
Fonte: elaborado pelo autor (2022).

3.2.1 Aprendizado Supervisionado

Sendo um subconjunto do Aprendizado de Máquina, o Aprendizado Supervisionado precisa de uma grande quantidade de dados para trabalhar, e como indicado no próprio nome precisa de alguma supervisão. A supervisão vem do conceito de indicar ao algoritmo quais são as respostas aceitáveis para o problema tradado. Já que ao utilizarmos essa técnica, a cada exemplo executado é preciso indicar qual é saída correta (LUDERMIR, 2021).

Essa técnica é utilizada para resolver dois tipos de problemas: classificação e regressão. A Figura 3 apresenta um resultado hipotético para um problema de classificação e para um problema de regressão. A classificação é o processo onde é necessário categorizar um conjunto de dados em classes pré-determinadas. Como exemplo, pode-se citar um sistema de e-mails, onde poderia haver um algoritmo que classificasse se um e-mail recebido é spam ou não.

Figura 3 – Classificação e Regressão



(a) Classificação (b) Regressão

Fonte: **??**).

Algoritmos de regressão são utilizados quando é necessário prever algum dado. Como exemplo, pode-se citar a analise de preços das casas em um bairro, com base em seus atributos como número de quartos ou de banheiros, área construída, área total do terreno ou a proximidade ao centro da cidade. Todas essas informações podem fornecer uma função que apresente uma relação entre os dados de entrada e o dado de saída (o preço da propriedade).

3.2.2 Aprendizado Não Supervisionado

Diferente dos algoritmos de Aprendizado de Máquina Supervisionados, os algoritmos Não Supervisionados trabalham sem vigilância, ou seja, não há necessidade do rótulo de resposta

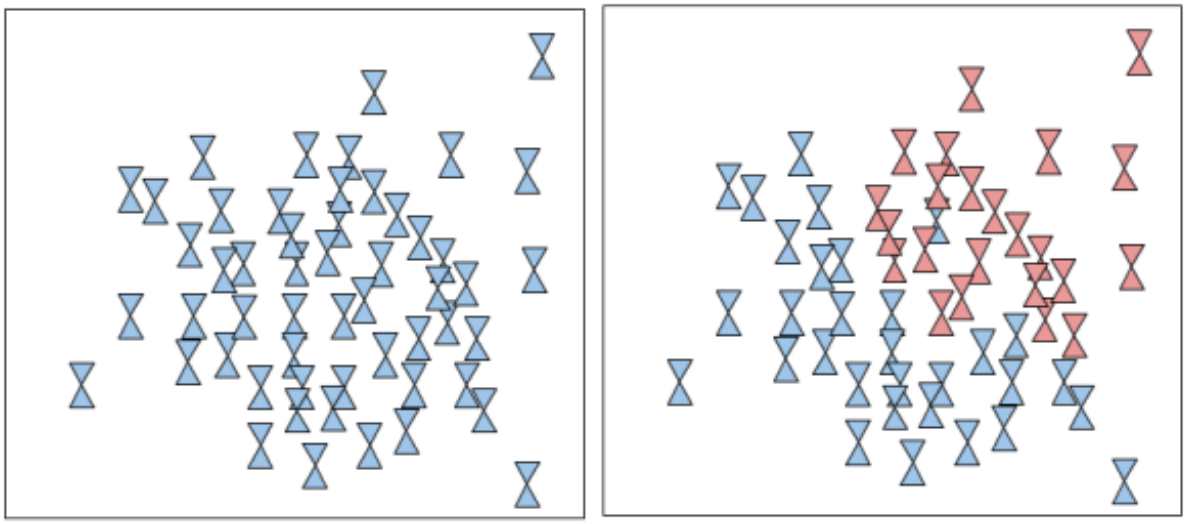
final. Desta forma, estes algoritmos buscam aprender as relações entre os dados através de similaridades. Como dito por **??**), o objetivo principal aqui é justamente encontrar padrões, regularidades ou estruturas existentes no conjunto inicial de dados.

3.3 Algoritmos de clusterização

Os algoritmos de clusterização fazem parte do conjunto de algoritmos de Aprendizado de Máquina Não Supervisionados, ou seja, estes algoritmos buscam agrupar dados durante a sua execução. De acordo com Maia (2020), os algoritmos de clusterização tem como tarefa identificar e agrupar automaticamente dados pelo seu grau de similaridade. Portanto, este tipo de algoritmo irá utilizar o conjunto de dados de tal forma que os dados contidos sejam unidos em grupos que representem fatias de informações com muitas semelhanças entre si.

Como definido por **??**), um algoritmo de clusterização agrupa dados que aparentemente algo em comum, essa tarefa consiste em reunir cada subconjunto de dados em um *cluster*. Para garantir a existência desta similaridade entre elementos de um mesmo *cluster* é necessário considerar uma medida que indique correspondência entre estes dados(DIAS *et al.*, 2004). A Figura 4 apresenta um exemplo hipotético de aplicação de um algoritmo de clusterização com a geração de dois *clusters*.

Figura 4 – *Clustering*

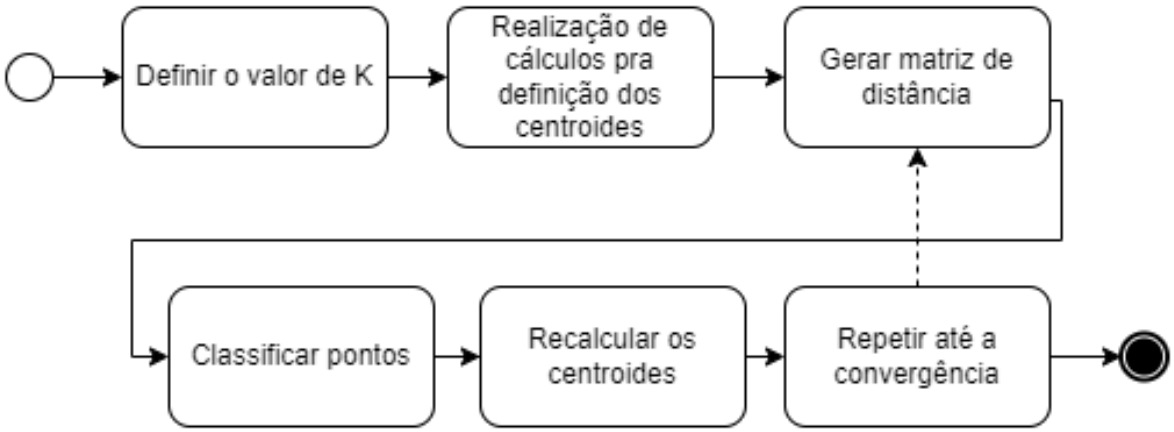


Fonte: **??**).

3.3.1 K-Means

De acordo com Pimentel *et al.* (2003), objetivo deste algoritmo é encontrar a melhor divisão de *P* dados em *K* grupos utilizando uma função de distância para calcular a similaridade dos dados de cada grupo. Esta distância é calculada utilizando o conceito de centroíde, sendo somada todas as distâncias de cada elemento de cada *cluster* para seu centroide. O objetivo principal deste algoritmo é minimizar as distâncias entre os elementos de cada *cluster*. A Figura 6 apresenta as 7 etapas principais utilizadas pelo *K-Means*.

Figura 5 – Execução do *K-Means*



Fonte: elaborado pelo autor (2022).

1. **Definição do valor** *K*: representa os *clusters* que o algoritmo vai separar os dados. Esse valor pode ser definido pelo domínio ou pela realização de testes que podem auxiliar a determinação do número de grupos que retornarão uma solução mais condizente com o problema em estudo.
2. **Realização de cálculos iniciais para definição dos centroides**: definição de centroides iniciais para cada *cluster* durante a primeira execução do algoritmo.
3. **Geração da matriz de distância**: realização de cálculos para definir a distância de cada elemento de cada *cluster* para cada um dos centroides definidos.
4. **Classificar dos elementos**: atualização para redefinir, se necessário, qual *cluster* cada elemento pertence, de acordo com a distância calculada para o centroide mais próximo.
5. **Recalcular os centroides**: nesta etapa os centroides poderão ser reajustados.
6. **Repetir até a convergência**: há o retorno do algoritmo para etapa 3 caso o algoritmo ainda não tenha atingido a taxa convergência.

4 TRABALHOS RELACIONADOS

Este capítulo descreve trabalhos da literatura mais relevantes para a contextualização do problema proposto nesta monografia.

4.1 *Prevendo a preferência do usuário para filmes usando o banco de dados Netflix*

Com a crescente popularidade de serviços de *streaming* nos últimos anos e consequentemente o aumento do número empresas que disponibilizam este tipo de serviço, o uso de funções de recomendações de conteúdos personalizados tem sido utilizada como uma ferramenta crucial para manter o usuário o maior tempo possível envolvido nestas plataformas. Procurando solucionar questões a respeito desta problemática, o trabalho de Goel e Batra (2009), busca prever a preferência do usuários por filmes utilizando dados da plataforma de *streaming Netflix*.

No trabalho em questão, os autores propuseram a aplicação do algoritmo *Latent Genre Space* (LGS), que tem como foco apreender traços de personalidades dos usuários. Para melhoria do banco de dados, além dos dados extraídos da plataforma *Netflix*, também foram utilizados dados do *Internet Movie Database* (IMDB), site de classificação de séries e filmes, para compor os gêneros que seriam utilizados. O algoritmo LGS foi executado em um espaço de baixa dimensão, considerando 23 gêneros de filmes, sendo cada usuário representado por um vetor nesse espaço. Desta forma, para cada filme não visto pelo usuário é definida uma pontuação correspondente, ponderada pela classificação de gênero com base em notas atribuídas por pessoas com gostos semelhantes, onde o filme pode também ser classificado como um filme multigênero. Neste último caso, todos os gêneros recebem o mesmo peso, de modo que, são tratados de formas iguais para a ponderação.

Por fim, também é apresentado agrupamentos de usuários no espaço gerado pelo algoritmo LGS a partir da aplicação do algoritmo *K-Means*. Este agrupamentos foram utilizados com o objetivo de encontrar usuários parecidos para preencher notas ausentes de filmes ainda não vistos por um indivíduo, com base nas classificações de usuários semelhantes.

Em contraste com o trabalho de Goel e Batra (2009), que apresentou uma metodologia para prever possíveis notas que um usuário daria a um filme com base em seus "vizinhos". Este trabalho busca clusterizar os dados das séries disponíveis em diferentes *streaming (Netflix, Prime Video, Disney+)*, além de buscar prever possíveis classificações para novas entradas de dados.

4.2 Análise de dados no *Internet Movie Database* (IMDb)

O site IMDB, também conhecido como *Internet Movie Database*, organiza e disponibiliza dados sobre produtos da cultura pop como filmes, séries e jogos, ao público. O trabalho de Cardoso (2021), busca analisar uma base de dados do IMDB para determinar quais dados podem ser de valia para uma analise quantitativa apontando preferências e padrões para os dados. Para tal, os autores selecionaram variáveis e arquivos de interesse para a pesquisa, normalizando estes dados para a aplicação de algoritmos como Árvore de Decisão e Naive Bayes na tentativa de encontrar padrões relevantes.

Os autores trabalharam na base de dados, inicialmente, para determinação de informações válidas para o problema, a partir disso a base de dados foi construída e normalizada utilizando dois arquivos que apresentavam dados como o título da produção, ano de lançamento, gêneros, a média das avaliações, o numero de votos, entre outros dados. Para atingir o objetivo de analisar padrões de gêneros, de acordo com cada tipo de produção (filmes, séries, filmes para TV ou minisséries) os dados foram analisados com o intuito de responder questões como: "quantidade de títulos produzidos por ano?" ou "a quantidade de votos por gêneros".

Ao analisar a tabela normalizada, os autores foram capazes de apresentar algumas métricas gerais a respeito dos questionamentos. Após as análises das *features* "A quantidade de filmes adultos produzidos por ano" ou até mesmo a "Porcentagem de notas recebidas por cada gênero", o resultado obtido pelo algoritmo "Naive Bayes" ele demonstrou relações destas *features*, como: a probabilidade de um tipo de gênero ser utilizado em maioria por um determinado tipo de mídia ou a probabilidade de um tipo de produção produzir um conteúdo adulto ou não.

Esta pesquisa difere do trabalho de Cardoso (2021), pois ao invés de apresentar métricas e agrupamentos relacionados aos dados do , este trabalho busca apresentar métricas e agrupamentos os dados de diversas séries disponíveis em diferentes *streaming* (*Netflix*, *Prime Video* e *Disney+*).

4.3 Aplicação de mineração de dados com o método de agrupamento *K-Means* e índice *Davies Bouldin* para agrupamento de filmes IMDB

Ashari *et al.* (2022) aplicaram mineração de dados conjuntamente com o algoritmo de clusterização K-Means, em uma base de dados de séries e filmes disponíveis no IMDB, na tentativa de compreender quais dados são mais relacionáveis entre si e como essa relação é

transpassada para os dados do conjunto.

Para trabalhar com a base de dados dados, inicialmente, foi aplicado o *Knowledge Discovery in Databases* (KDD), um método que busca novos conhecimentos a partir dos dados coletados e processados. Após a aplicação do KDD, a base de dados foi padronizada, sendo determinado quais dados seriam adequados para a modelagem, com intuito de obter resultados mais precisos. Das 16 variáveis disponíveis para o conjunto dados, foram selecionadas cinco (*runtime, imdb_rating, meta_score, no_of_votes, e gross*). A partir desta etapa, realizou-se o pré-processamento dos dados, limpando os dados considerados desnecessários como o *poster_links, series_title, released_year, certificate, genre, director, star1, star2, star3 e star4*. Nas etapas posteriores, os autores aplicaram uma transformação nos dados para se tornarem mais precisos e adequados, como o *IMDB rating* são ajustado para uma pontuação de 1 a 10, enquanto o *Meta score* foi ajustado para valores entre 10 a 100.

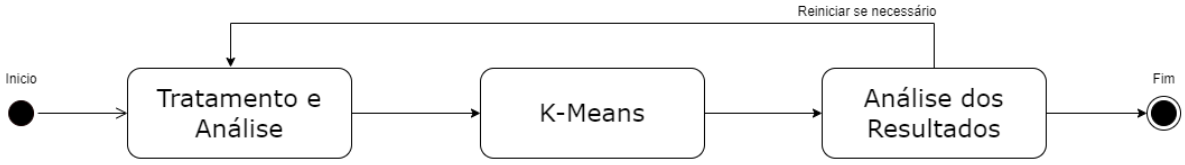
Por fim, foi utilizando o algoritmo *K-Means* para encontrar padrões por agrupamento, para tal foram testadas duas técnicas: o *Elbow Method* (EM) and *Davies Bouldin Index* (DBI) com o objetivo de encontrar qual dos dois encontraria o número de *clusters* mais adequados para o conjunto. Como resultado final, os dados foram agrupados em 4 grupos (valor encontrado pelo DBI) para análise geral.

Semelhante a pesquisa de Ashari *et al.* (2022), este trabalho buscar agrupar os dados para uma análise através do uso de algoritmos de clusterização. Mas diferente do trabalho de Ashari *et al.* (2022) que lidou com uma base de dados dos Top 1000 itens do IMDB, este projeto vai utilizar um conjunto de base de dados públicas da *Netflix*, *Disney* + e *Amazon Prime Video*

5 METODOLOGIA

Neste capítulo é apresentada a metodologia utilizada nesta monografia para obtenção do objetivos pretendidos. Cada uma das seções subsequentes descreve as atividades realizadas em cada uma das três etapas da metodologia definida. A Figura 6 demonstra o fluxo das etapas realizadas.

Figura 6 – Representação da metodologia proposta



Fonte: elaborado pelo autor (2022).

5.1 Tratamento e Análise da base de dados

As bases de dados utilizadas para a realização desta pesquisa foram retiradas do site *Kaggle*, que disponibiliza de forma pública base de dados e competições relacionadas a análise de dados, sendo selecionadas um total 2 bases de dados "*Netflix TV Shows and Movies*"e "*HBO Max TV Shows and Movies*". A Tabela 1 apresenta os tipos de dados disponibilizados nestas bases.

Para manipulação dos dados, foi utiliza a linguagem de programação *Python* 3.0 e a biblioteca *Pandas*. Inicialmente, todos os dados foram agrupados em uma única base de dados, sendo utilizada a estrutura de dados *dataframe* para seu armazenamento, uma nova coluna denominada "streaming", do tipo "object" foi criada para definir qual a base de dados original a qual o conjunto de dados (linha do *dataframe*) pertence.

As seguintes *features* foram removidas: "*description*", "*age_certification*", "*seasons*", "imdb_id". A decisão pela remoção destas *features* foi tomada após a realização de uma análise gráfica dos dados que demonstrou que tais informações não auxiliaram o aprendizado do algoritmo *K-means*. Também foi realizada a remoção de dados nulos e algumas *features* que possuíam valores no formato de listas ("*production_countries*"e "*genres*"), assim como algumas *features* que possuíam valores no formato texto, precisaram serem convertidas para valores numéricos. Como por exemplo, pode-se citar a *feature* "*streaming*", onde se criou duas novas *features*, uma para cada respectivo serviço, assim se o registro era "netflix" a *feature*

Quadro 1 – Formato dos Dados

Feature	Tipo	Exemplo
id	object	"tm84618"
title	object	"Taxi Driver"
type	object	"MOVIE"
description	object	"A mentally unstable Vietnam War veteran..."
release_year	int64	1976
age_certification	object	R
runtime	int64	114
genres	object	['drama', 'crime']
production_countries	object	['US']
seasons	float64	1.0
imdb_id	object	"tt0075314"
imdb_score	float64	8.2
imdb_votes	float64	808582.0
tmdb_popularity	float64	40.965
tmdb_score	float64	8.179
Streaming	object	netflix

Fonte: elaborado pelo autor (2022).

"netflix"recebe o valor 1, já se um registro fosse da "hbo", *feature* "hbo"recebe o valor 1.

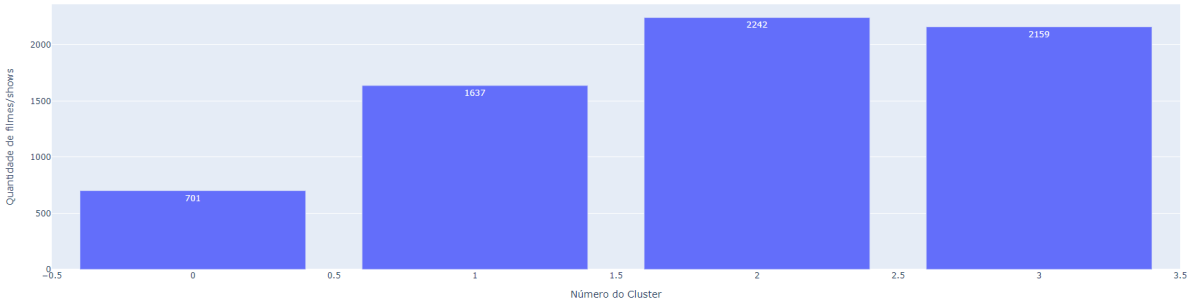
5.2 Geração de Modelos Inteligentes

Para geração do agrupamento foi utilizado o algoritmo *K-Means* da biblioteca *sklearn* da linguagem *Python* 3.0, sendo definido um total inicial de doze *clusters*. A partir deste valor foram realizados testes empíricos decrementando, em cada teste, uma unidade do número total de *clusters* inicial com o intuito de obter o número de *clusters* mais viável para a base de dados estudada. Por fim, foi estabelecido um número total de 4 *clusters* como o valor ótimo.

6 RESULTADOS

Como mencionado na capítulo anterior, a aplicação do algoritmo *K-Means* retornou um total de 4 *clusters*. O Gráfico 7 apresenta o número de registros agrupados em cada *cluster* e a seguir é apresentado uma análise geral de cada um destes *clusters*.

Figura 7 – Quantidade de shows/filmes por *clusters*



Fonte: elaborado pelo autor (2022).

6.1 *Cluster* 0

Neste *cluster* foram agrupados 701 registros, desses, 516 são produções disponíveis na *Netflix*, que se dividem em 331 filmes e 185 séries, já para *HBO Max* foram agrupados 185 produções, sendo 27 séries e 158 filmes. O Quadro 2 apresenta um resumo dos dados mencionados.

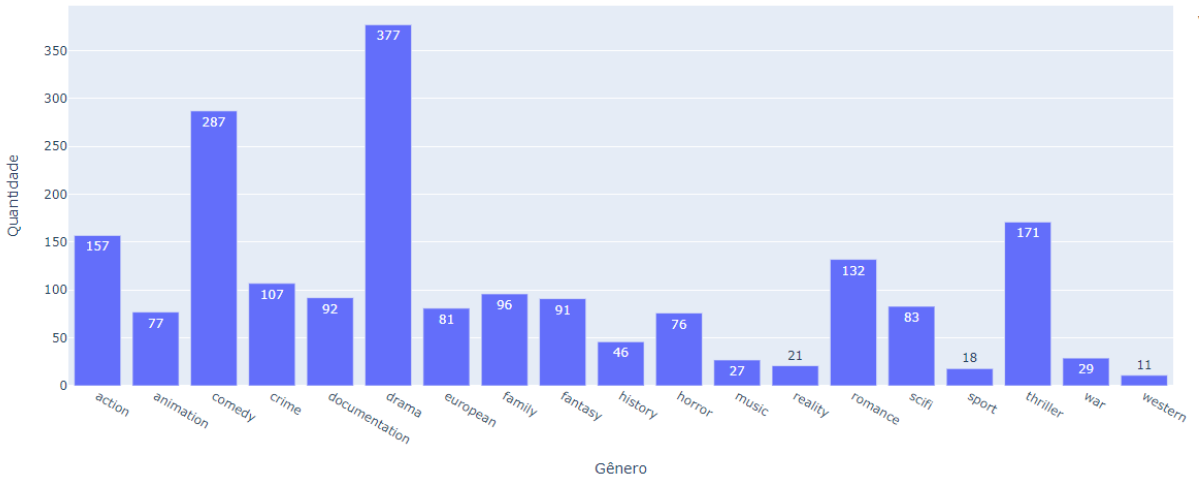
Quadro 2 – Quantidade de registros por *streaming* no *cluster* 0 e seus tipos.

Streaming	Tipo	Quantidade	Total
Netflix	*MOVIE*	331	516
	SHOW	185	
HBO Max	*MOVIE*	158	185
	SHOW	27	
			701

Fonte: elaborado pelo autor (2022).

As produções deste *cluster* possuem uma variedade de gêneros, alguns destes categorizados com mais de um gênero. Os gêneros os quais os usuários mais demostraram interesse foram: "*drama*"e "*comedy*", com 377 e 287 das produções agrupadas, respectivamente. Acima de 100 produções também se encontram os gêneros: "*thriller*"(171 produções), em seguida "*action*" com 157, "*romance*" com 132 e "*crime*" com 107, como pode ser visto no Gráfico 8.

Figura 8 – *Cluster* 0: Gêneros

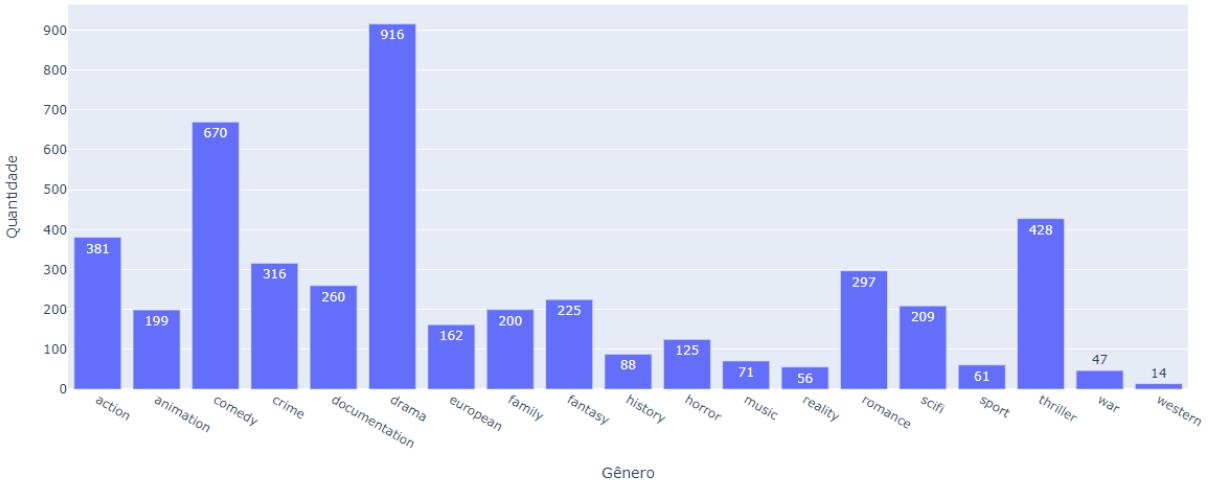


Fonte: elaborado pelo autor (2022).

6.2 *Cluster* 1

No *cluster 1* foram associados apenas produções disponíveis na *Netflix* com uma composição de 1021 filmes e 616 séries. Ao analisarmos os gêneros, 916 dos itens foram categorizados como "*drama*", enquanto 670 produções como "*comedy*", 428 registros como "*thriller*", em seguida "*action*" com 381 e "romance" com 297, como pode ser visto no Gráfico 9.

Figura 9 – *Cluster* 1: Gêneros



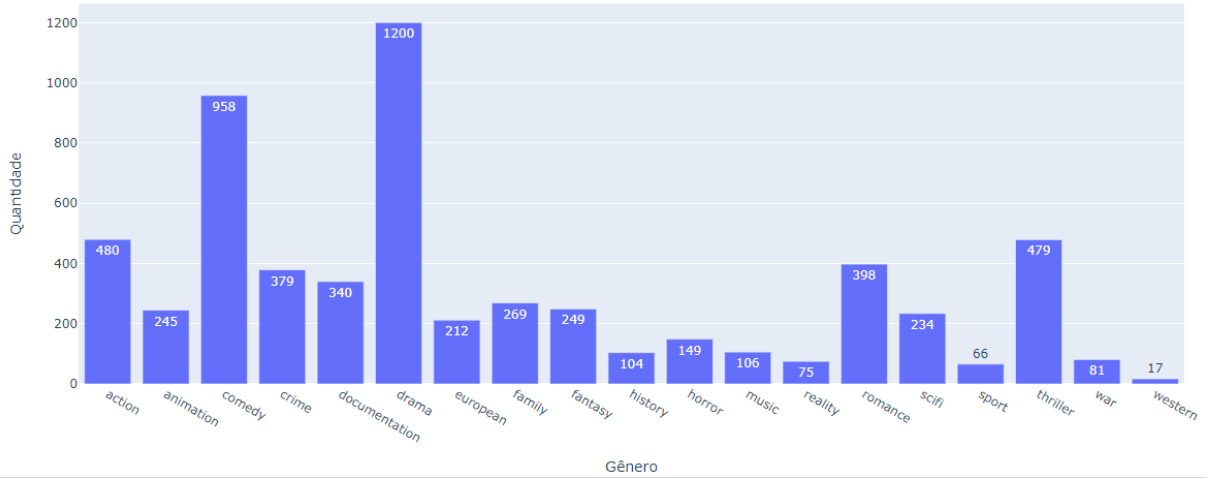
Fonte: elaborado pelo autor (2022).

6.3 *Cluster* 2

Np *cluster 2* foram agrupados apenas produções disponíveis na *Netflix* totalizando 1479 filmes e 763 séries. Em relação a análise dos gêneros, 1200 produções foram categori-

zados como "*drama*", já outras 958 como "*comedy*", 480 registros como "*action*", em seguida "*thriller*"com 479 e "*romance*"com 398, como pode ser visualizado no Gráfico 10.

Figura 10 – *Cluster* 2: Gêneros



Fonte: elaborado pelo autor (2022).

6.4 *Cluster* 3

No *cluster* 3, mais de 70% dos itens pertencem a *HBO Max* com 1333 filmes e 204 séries, enquanto 365 filmes e 257 séries estavam disponíveis na plataforma *Netflix* (veja Quadro 3) .

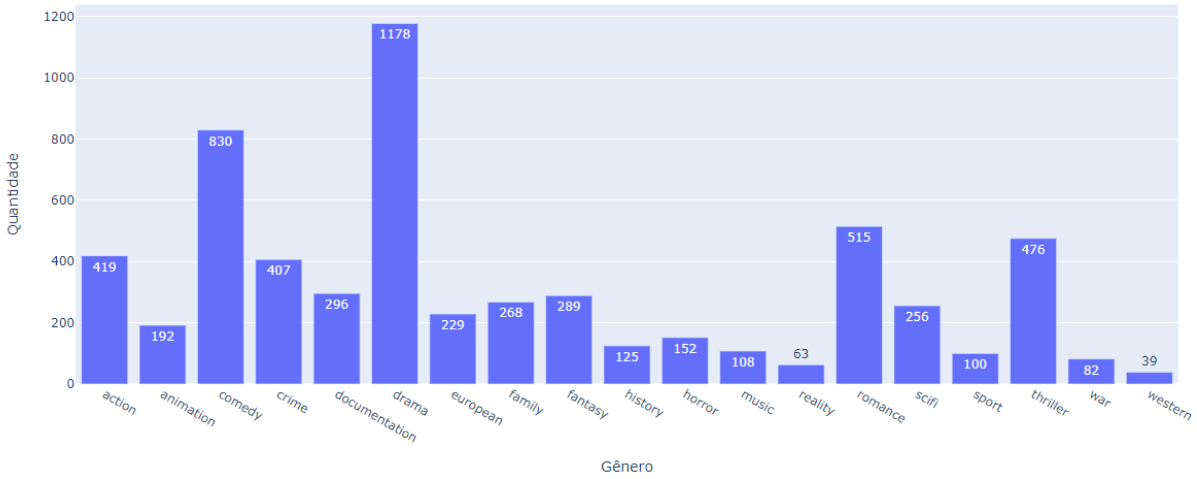
Quadro 3 – Quantidade de registros por *streaming* no *cluster* 3 e seus tipos.

Streaming	Tipo	Quantidade	Total
Netflix	*MOVIE*	365	622
	SHOW	257	
HBO Max	*MOVIE*	1333	1537
	SHOW	204	
			2159

Fonte: elaborado pelo autor (2022).

Analisando os gêneros do *cluster 3*, há 1178 produções categorizadas como "*drama*", outras 830 como "*comedy*", enquanto "*romance*" possui 515, "*thriller*"com 476 registros, em seguida "*action*"com 419, como pode ser visualizado no Gráfico 11.

Figura 11 – *Cluster* 3: Gêneros



Fonte: elaborado pelo autor (2022).

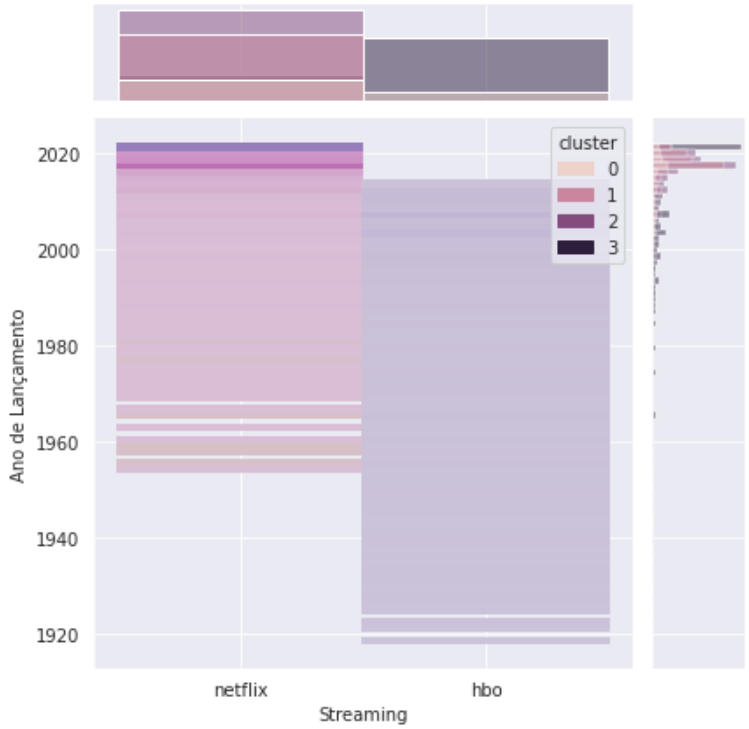
6.5 Análises gráficas gerais

Através da geração de alguns gráficos foi possível inferir alguns *insights* interessantes sobre a base de dados estudada. O Gráfico 12 representa os anos de lançamentos das produções existentes para cada *streaming* separados por seus respectivos *clusters*. É possível notar que há produções da *HBO Max* desde a década de 1920 e que as produções consideradas mais "antigas" foram agrupados na sua maior parte no *cluster* 3 e uma minoria no demais *clusters*. As produções da *Netflix*, independente do ano de produção, estão dividas principalmente entre os três primeiros *clusters*. Desta forma, podemos inferir que os usuários do *cluster* 3 podem ter uma tendência a ter interesse por produções mais antigas e por filmes da *HBO Max*.

Ao analisar as pontuações do IMDB disponíveis na base de dados, visualizando esta informação separa por cada *cluster* gerado (ver Figura 13), é possível notar os usuários pertencentes ao *cluster 1* e *cluster 2* apresentam um forte interesse em produções da *Netflix* relatando pontuações entre 4 a 8. Novamente, é possível fortalecer a hipótese que os usuários do *cluster 3* apresentam um um grande interesse em produções da *HBO Max* relatando pontuações entre 4 a 8. Pode-se notar também que os usuários do *cluster 0* não possuem preferência por plataforma de *streaming* específica, porém podemos visualizar uma alta preferência por filmes atuais neste *cluster*.

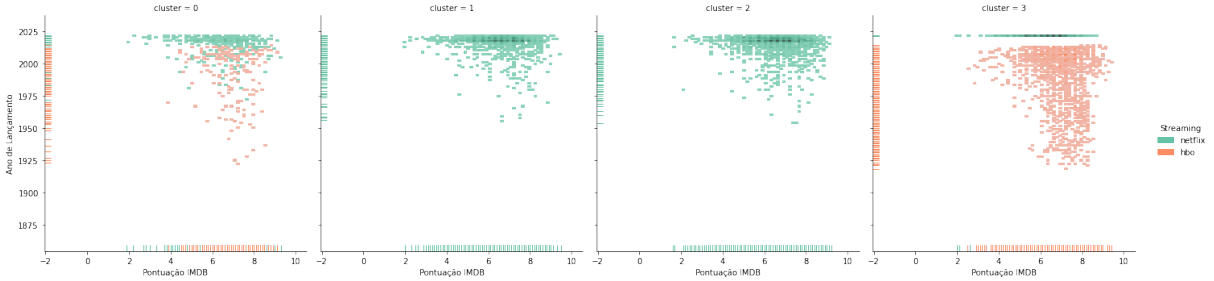
Ao realizar uma análise do tempo de excução das produções e sua relação com suas pontuações no IMDB (Figura 14), é possível observar que as produções da *HBO Max*, embora possuam produções com durações mais longas conseguem manter notas comparáveis a produções da *Netflix* com tempo de execução menores. Porém, os usuários do *cluster 1* 1

Figura 12 – Anos de lançamentos de produções nos *Streaming* e seus agrupamentos por *clusters*



Fonte: elaborado pelo autor (2022).

Figura 13 – Pontuação IMDB X Anos de lançamentos por *clusters*

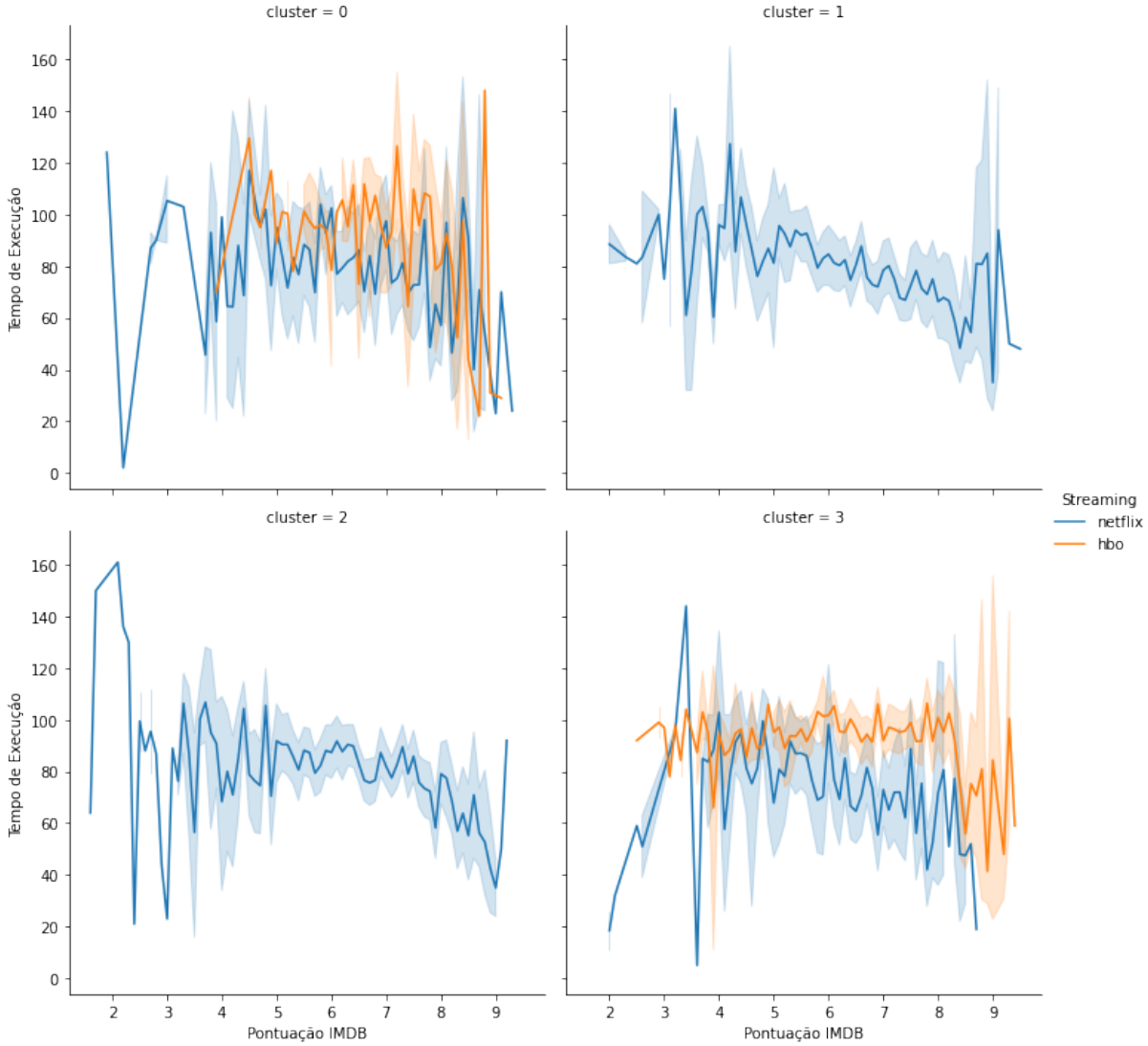


Fonte: elaborado pelo autor (2022).

e *cluster* 2 possuem uma notável preferência por produções mais curtas que por conseguinte conseguem pontuações IMDB melhores, havendo apenas alguns *outliers*.

Ao se depararmos com o quadro 3, é fácil notar que a *HBO Max* no espectro da duração dos itens entre 80 a 110 minutos, abrange o espectro de notas entre 2 a 8, com pontos mínimos fora da curva com nota 4 e tempo de execução perto dos 20 minutos, enquanto notas acima de 8 estão em produções com tempo de duração abaixo dos 80 minutos. Ao mesmo passo que a *Netflix* possui itens com durações abaixo de 40 minutos com notas indo de 2 a valores próximos de 4 e até notas próximas de 9.

Figura 14 – Pontuação IMDB X Tempo de Execuções por *clusters*



Fonte: elaborado pelo autor (2022).

7 CONCLUSÕES E TRABALHOS FUTUROS

Este capítulo discorre sobre as considerações e lições aprendidas a respeito da metodologia criada neste trabalho, para a aplicação de algoritmos de clusterização, assim como os resultados obtidos.

7.1 Considerações gerais

Neste trabalho buscou-se analisar conjuntos de dados sobre as produções disponíveis em plataformas de *streaming*, mais especificamente a *Netflix* e *HBO Max* e quais informações poderiam ser extraídas após a aplicação do algoritmo de agrupamento *K-Means*.

- Foi possível observar que a produções HBO Max, em geral possuem avaliações melhores que as da sua concorrente, mesmo para produções mais antigas;
- Há um *cluster* com usuários com preferências para filmes mais antigos, enquanto os demais *clusters* possuem usuários com preferências mescladas.
- Ao analisarmos os gêneros das produções podemos notar que em todos os quatros *clusters* definidos pelo algoritmo, os principais estilos são os mesmos, sendo preferencialmente "*comedy*"e "*drama*", e os demais, mesmos que em ordens ligeiramente diferentes: "*action*", "*thriller*"e "*romance*".

7.2 Trabalhos futuros

Em trabalhos futuros, é esperado a utilização de mais conjuntos de dados como o da *Prime Video*, *Disney+* e *Paramount+*, além do uso do outro arquivo disponibilizado nos conjuntos que continham informações sobre os créditos das obras. Tais dados novos deverão alterar consideravelmente os resultados obtidos, devendo fornecer novos *insights* possivelmente relevantes para uma análise mais ampla das plataformas de *streaming* e seus conteúdos.

Por fim, espera-se que em futuros trabalhos o conjunto de dados utilizado possa ser incrementando com outros *dataframes* complementares ao assunto, como o "The Movies Dataset", correlacionando os dados básicos e fornecendo novas informações ao algoritmo, trabalhando também com outras técnicas de análise de dados como *Knowledge Discovery from Text*.

REFERÊNCIAS

ASHARI, I. F.; BANJARNAHOR, R.; FARIDA, D. R.; AISYAH, S. P.; DEWI, A. P.; HUMAYA, N. *et al.* Application of data mining with the k-means clustering method and davies bouldin index for grouping imdb movies. **Journal of Applied Informatics and Computing**, v. 6, n. 1, p. 07–15, 2022.

CARDOSO, E. L. **Análise de dados no internet movie database**. 2021.

CASTANHEIRA, L. G. **Aplicação de técnicas de mineração de dados em problemas de classificação de padrões**. Universidade Federal de Minas Gerais, 2008.

DIAS, C. R. *et al.* **Algoritmos evolutivos para o problema de clusterização de grafos orientados: Desenvolvimento e análise experimental**. 2004.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996.

GOEL, D.; BATRA, D. Predicting user preference for movies using netflix database. **Department of Electrical and Computer Engineering, Carniege Mellon University**, p. 1–7, 2009.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. **Estudos Avançados**, SciELO Brasil, v. 35, p. 85–94, 2021.

MAIA, C. L. **Prevendo a força de conexão por meio da rede social online facebook.** 2020.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, Manole, v. 1, n. 1, p. 32, 2003.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007.

PIMENTEL, E. P.; FRANÇA, V. F. de; OMAR, N. A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2003. v. 1, n. 1, p. 495–504.

..