UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Avaliação de sistemas de recomendação para plataformas de streaming com animes

Marden Nilton Rodrigues da Silva

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Marden Nilton Rodrigues da Silva

Avaliação de sistemas de recomendação para plataformas de streaming com animes

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Rafael Geraldeli Rossi

Versão original

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Silva, Marden
S856m Avaliação de sistemas de recomendação para plataformas de streaming com animes / Marden Nilton Rodrigues da Silva ; orientador Rafael Geraldeli Rossi. – São Carlos, 2024.
81 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Sistemas de recomendação. 2. Streaming de animes. 3. Classe USPSC. 4. Tese. 5. Documentos eletrônicos. I. ROSSI, R. G., orient. II. Título.

Marden Nilton Rodrigues da Silva

Avaliação de sistemas de recomendação para plataformas de streaming com animes

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Nome Orientador

Original version

São Carlos

2024

Dedico este trabalho à minha família, que sempre valorizou o estudo como pedra angular da prosperidade, e à minha companheira de vida Vanessa, que me oferece apoio incondicional para perseguir meus sonhos e estudar aquilo que faz meus olhos brilharem.

AGRADECIMENTOS

Agradeço à minha noiva, Vanessa Coelho, por todo o suporte emocional e incentivo constante, mesmo nos momentos mais difíceis.

À minha família, por ser a base da minha vida, e por quem eu luto por um futuro melhor.

Agradeço especialmente ao meu orientador, Prof. Dr. Rafael Rossi, por sua sabedoria e presença durante todo o desenvolvimento deste trabalho. Honestamente, gostaria que todo estudante pudesse ter um orientador como ele.

Agradeço à Universidade de São Paulo e às coordenadoras Solange Rezende e Roseli Romero pela dedicação em garantir que nenhum estudante fique para trás.

Agradeço à Barkus e minha sócia Maria Beatriz Santos pela parceria e investimento no meu desenvolvimento profissional.

"A única coisa que podemos fazer é acreditar que não vamos nos arrepender da escolha que fizemos."

Levi Ackerman, Shingeki no Kyojin

RESUMO

SILVA, Marden R. **Avaliação de sistemas de recomendação para plataformas de streaming com animes**. 2024. 81p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

O mercado de streaming passou por um crescimento vertiginoso nos últimos anos. Plataformas como Netflix, Amazon, Disney e HBO continuam investindo em tecnologia para facilitar o acesso a conteúdos on demand, incluindo filmes, séries e animações japonesas (animes). Com a quantidade cada vez maior de conteúdos disponíveis, sistemas de recomendação assumem papel fundamental para aumentar a retenção e satisfação dos usuários. Os animes, uma categoria crescente em popularidade, têm recebido destaque em plataformas para o público geral e, anualmente, registra-se uma quantidade cada vez maior de lançamentos. No entanto, a literatura carece de estudos sobre recomendações para esse tipo de categoria. Por esse motivo, este trabalho visou avaliar técnicas de recomendação de conteúdo para plataformas de streaming, com foco específico em animes. Buscou-se determinar quais técnicas são mais eficazes em prever as notas que os usuários dariam a animes. Foram exploradas diversas técnicas de recomendação, incluindo sistemas baseados em popularidade, baseados em conteúdo (utilizando metadados como gênero e sinopse), filtragem colaborativa (baseada em memória e em modelo) e sistemas híbridos. A avaliação foi feita com as métricas RMSE e MAE utilizando um conjunto de dados originado da plataforma MyAnimeList. Os resultados demonstraram que a filtragem colaborativa baseada em modelo, especificamente usando SVD (Singular Value Decomposition), apresentou a melhor performance de recomendação, seguida por um sistema híbrido que combinou filtragem colaborativa baseada em item com a abordagem baseada em conteúdo. Porém, destaca-se também a importância de utilizar diferentes métodos de recomendação a depender da etapa da jornada do usuário, garantindo uma experiência personalizada e eficiente na escolha de conteúdos.

Palavras-chave: Sistemas de Recomendação. Plataformas de Streaming. Animes. Filtragem Colaborativa. Filtragem Baseada em Conteúdo. Filtragem Baseada em Popularidade. Sistema de Recomendação Híbrido. CRISP-DM.

ABSTRACT

SILVA, Marden R. **Recommender systems evaluation for streaming platforms with animes**. 2024. 81p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

The streaming market has experienced tremendous growth in recent years. Platforms like Netflix, Amazon, Disney, and HBO invest continuously in technology and content, including movies, series, and Japanese animations (animes). With the increasing amount of available content, recommendation systems play a crucial role in enhancing user retention and satisfaction. Animes, a category growing in popularity, have gained prominence on platforms for the general audience, with an increasing number of releases recorded annually. However, the literature lacks studies about recommendations on this kind of category. For this reason, this work aimed to evaluate content recommendation techniques for streaming platforms, with a specific focus on animes. The goal was to determine which techniques are most effective in predicting the ratings users would give to animes. Various recommendation techniques were explored, including popularity-based systems, content-based systems (using metadata such as genre and synopsis), collaborative filtering (both memory-based and model-based), and hybrid systems. The evaluation was conducted using RMSE and MAE metrics on a dataset sourced from the MyAnimeList platform. The results showed that model-based collaborative filtering, specifically using SVD (Singular Value Decomposition), demonstrated the best recommendation performance, followed by a hybrid system that combined item-based collaborative filtering with the content-based approach. However, it is also important to highlight the significance of using different recommendation methods depending on the stage of the user's journey, ensuring a personalized and efficient content selection experience.

Keywords: Recommendation Systems. Streaming Platforms. Animes. Collaborative Filtering. Content-Based Filtering. Popularity-Based Filtering. Hybrid Recommendation System. CRISP-DM.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O mercado de streaming passou por um crescimento vertiginoso nos últimos anos. A prova disso é que a gigante do setor, Netflix, alcançou em 2023 a marca de 238 milhões de assinantes pelo mundo todo (Estadão Conteúdo, 2023), o dobro do que tinha cinco anos antes (Folha de São Paulo, 2018). Acompanhando esse movimento, muitas outras plataformas de grande alcance como a Amazon, Disney e HBO investiram montantes volumosos de dinheiro na tecnologia capaz de facilitar o acesso on demand de conteúdos de diversos tipos como filmes, séries e animações. Só no Brasil, 70% da população afirma ter a assinatura de alguma delas (Ligia Mello, 2023), indicando que se trata de um mercado consolidado com algum espaço para crescer. Globalmente, espera-se que até 2027 o valor desse mercado alcance 1 trilhão de dólares (Paula Filizola, 2021).

As plataformas de streaming apresentam diversos tipos de conteúdos, como filmes e séries de diferentes gêneros, transmissões esportivas, reality shows e animações japonesas (animes). Essas últimas vêm se destacando nos últimos anos, deixando de habitar apenas plataformas de nicho (como a Crunchyroll), e se consolidando até mesmo com produções originais em plataformas de propósito geral (como Netflix, Max e Amazon Prime) (Patrick Macias, 2021). Líder no mercado neste segmento, A Crunchyroll chegou ao Brasil em 2012 e foi comprada pela Sony em 2021 por mais de um bilhão de dólares (Felipe Vinha, 2021), o que demonstra uma alta valorização por esse tipo de conteúdo. Enxergando o Brasil como uma peça-chave para o crescimento da empresa, a Sony pretende enriquecer cada vez mais o catálogo e torná-lo mais acessível para os consumidores (CARBONE, 2024).

Atualmente, a quantidade de conteúdos já é bastante volumosa. Somente na Netflix, são mais de 3.300 filmes e 1.850 séries (JúNIOR DANIEL MATTOS, 2022). Contudo, a distribuição de mais conteúdos carrega consigo um risco relacionado à forma como a mente dos consumidores funciona. Mediante tantas opções disponíveis, é comum se sentir paralisado em relação ao que assistir, fenômeno associado ao chamado Paradoxo da Escolha, conceituado por Barry Schwartz. Na prática, o Paradoxo da Escolha se refere à condição em que, mediante inúmeras possibilidades, uma pessoa se vê impedida de tomar uma decisão, o que invalida a hipótese de que mais opções implicam em mais liberdade, a origem do paradoxo (SCHWARTZ, 2005). A partir dos experimentos de Amos Tversky e Daniel Kahneman, confirmou-se que o ser humano em geral tem uma forte aversão à perda (KAHNEMAN; TVERSKY, 1979). Por isso, escolher uma opção de anime para assistir entre tantas disponíveis pode gerar uma experiência negativa na medida em que desperta a sensação de estar renunciando às demais.

Diante disso, fez-se necessária não só a tecnologia que possibilitasse a manutenção das plataformas de streaming no ar, mas também mecanismos que apoiassem os usuários

em seus processos decisórios. Afinal, se os usuários têm dificuldade em encontrar conteúdo que seja significativo para eles, vão passar menos tempo lá, o que aumenta as chances de cancelamento da assinatura (KNAUER, 2019). É estratégico que a tecnologia não forneça apenas uma quantidade virtualmente limitada de conteúdos para facilitar a escolha, mas indique aquelas opções que fazem mais sentido para cada indivíduo, de acordo com suas preferências. É nesse contexto em que sistemas de recomendação tiveram sua importância acentuada nos últimos anos.

Um sistema de recomendação é um algoritmo de inteligência artificial (IA), geralmente associado à aprendizagem de máquina, que utiliza Big Data para sugerir ou recomendar produtos adicionais aos consumidores. Essas sugestões podem ser baseadas em vários critérios, incluindo compras anteriores, histórico de busca, informações demográficas e outros fatores (ARYOSETO; MARDIANTO; ARIWIBOWO, 2023). No caso de plataformas como a Netflix e a Crunchyroll, sabemos que a recomendação é destinada às melhores indicações de filmes, séries ou animes de acordo com o comportamento e feedback prévio do usuário. No entanto, sistemas de recomendação também estão por trás das redes sociais e buscadores, o que mostra o quanto fazem parte da vida cotidiana atualmente.

É interessante notar como a ascensão dessas tecnologias afetou a forma como os consumidores acessam conteúdo. O Brasil tem um longo histórico com a pirataria, ocupando a 5ª posição dos países que mais consomem conteúdos piratas no mundo (PODER360, 2022). Estima-se, ainda, que mais de 280 bilhões de reais sejam perdidos pelo país para o mercado ilegal (DALL'ARA, 2022). Ainda assim, entre 2013 e 2019, a frequência de consumo de conteúdos piratas reduziu fortemente segundo a pesquisa "Retratos da Sociedade" realizada pela Confederação Nacional da Indústria. A pesquisa aponta que um dos fatores para essa tendência foi justamente a popularização de alternativas de consumo de produtos originais a preços menores, como as plataformas de streaming de músicas e filmes, que reduziram a demanda por CDs e DVDs pirateados que eram muito populares no Brasil (CNI, 2022).

1.1 Objetivo e Questões de Pesquisa

Portanto, dado i) o consumo cada vez maior de plataformas de streaming; ii) o benefício destas plataformas para a diminuição da pirataria; iii) a necessidade do uso de sistemas de recomendação para mitigar os efeitos do paradoxo da escolha e a retenção de clientes na plataforma; e iv) o fato de não haver uma única técnica de recomendação capaz de prover os melhores resultados para diferentes tipos de recomendação e para diferentes tipos de dados, o objetivo deste trabalho de conclusão de curso é avaliar técnicas de recomendação de conteúdo para plataformas de streaming.

Mais especificamente, o objeto a ser recomendado serão animes, categoria que têm apresentado relevância crescente nos últimos anos e que recebeu relativamente pouca atenção acadêmica até o presente momento. Em suma, busca-se responder às questões:

- Q1: Qual técnica é mais assertiva (com menor erro) ao prever a nota que um usuário dará a um anime?
- Q2: Um sistema de recomendação de animes se sobrepõe aos demais em todas as ocasiões ou combinar diferentes recomendadores pode ser mais promissor?

Com isso, espera-se contribuir para um mercado em ascensão onde cada vez mais consumidores acessem conteúdos com o devido resguardo dos direitos autorais, tendo boas experiências e beneficiando não apenas as plataformas de streaming, mas também criadores, estúdios de animação e a sociedade.

1.2 Organização do Texto

Este trabalho está estruturado em cinco capítulos:

- No Capítulo 1 foi apresentada a introdução e a motivação para o estudo.
- No Capítulo 2 são apresentados os fundamentos teóricos.
- No Capítulo 3 são apresentados diferentes trabalhos relacionados ao tema deste documento.
- No Capítulo 4 é descrita a metodologia e as técnicas propostas pelo trabalho, bem como o processo de desenvolvimento do experimento.
- No Capítulo 5 são apresentados os resultados das técnicas implementadas e a análise das métricas de avaliação estabelecidas.
- Por fim, no Capítulo 6 são apresentadas as conclusões, discutindo quais foram os avanços e limitações do trabalho e também as sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Como consequência da superabundância de dados que precisam ser analisados para a tomada de decisão, a área de mineração de dados tem se tornado cada vez mais importante (CASTRO; FERRARI, 2017). Porém, o sucesso de um projeto de mineração de dados não depende apenas de boas ferramentas e analistas habilidosos. É necessária uma metodologia para o efetivo gerenciamento do projeto, seja qual for o nível de complexidade (WIRTH; HIPP, 2000). Nesse ponto, o Cross Industry Standard Process for Data Mining (CRISP-DM) emerge como um framework adaptável a múltiplas indústrias, com histórico comprovado de aplicações bem sucedidas e com um passo a passo definido (CHATTERJEE, 2022).

Fruto do trabalho de vários pesquisadores na década de 1990 (CHAPMAN *et al.*, 2000), o CRISP-DM se tornou o *framework* mais utilizado para gerenciar projetos de mineração de dados (HOTZ, 2023). Este *framework* consiste em seis etapas, a saber: i) Entendimento do Negócio; ii) Compreensão dos Dados; iii) Preparação dos Dados; iv) Modelagem; e v) Avaliação e Implantação. Assim como demonstrado na Figura 1, existem iterações entre etapas e, caso uma dessas etapas apresente resultado insatisfatório, é permitido o retorno para uma etapa anterior.

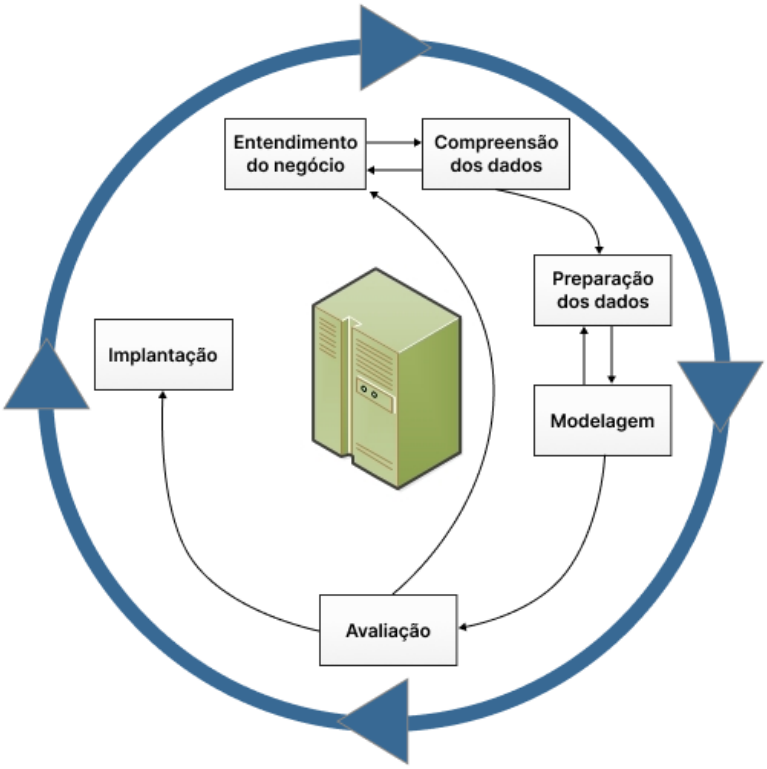


Figura 1 – Ciclo de vida da mineração de dados. Fonte: adaptado de (IBM, 2021).

Além disso, o ciclo externo demonstra a natureza cíclica de um processo de mineração de dados, em que os passos mencionados anteriormente podem ser repetidos conforme a chegada de novos dados, novos *insights*, ou ainda a utilização de novos algoritmos à procura de melhores resultados.

A seguir, cada etapa será apresentada detalhadamente em conjunto com as técnicas necessárias para o domínio de aplicação deste trabalho.

2.1 Entendimento do Negócio

A fase de Entendimento do Negócio é crucial para estabelecer os objetivos do projeto. Assim como realizar vendas começa com o entendimento de uma necessidade de mercado, a exploração de dados também começa com a identificação de uma necessidade que possa ser sanada (PYLE, 1999). Algumas perguntas a serem respondidas nesta etapa são: Qual é o contexto em que o projeto está inserido? Quais são os requisitos sob o ponto de vista de negócio para que o projeto seja considerado bem sucedido? Responder essas e outras perguntas similares possibilita que um problema sob o ponto de vista de negócio possa ser convertido em um trabalho de mineração de dados (WIRTH; HIPP, 2000).

2.2 Compreensão dos Dados

A etapa de Compreensão dos Dados, permite um aprofundamento no contexto onde os dados estão inseridos, o que permite identificar problemas com relação à qualidade dos dados, a forma como se distribuem, se contêm muitos valores nulos, entre outras informações pertinentes às etapas posteriores (CHAPMAN *et al.*, 2000). Esta etapa consiste em 4 sub-etapas: i) coleta inicial; ii) descrição dos dados, iii) exploração dos dados; e iv) verificação da qualidade dos dados (IBM, 2021).

Como ponto de partida, a coleta de dados inicial consiste em identificar fontes de dados e coletá-los, o que pode ser realizado por meio da utilização de dados previamente obtidos e sob domínio próprio, adquirindo dados de terceiros ou por meio de pesquisas adicionais (IBM, 2021). É um momento importante para identificar se os dados obtidos são suficientes para atender às necessidades levantadas na etapa de Entendimento do Negócio identificando aspectos como: quais atributos parecem mais promissores, quais são os meios de importação e como acessá-los (PYLE, 1999).

Em seguida, na etapa de descrição dos dados, a quantidade e a forma dos dados é analisada de maneira mais atenciosa, isto é, qual o volume de dados disponível e em que condição estão (IBM, 2021). Sabe-se que conjuntos grandes de dados podem produzir modelos mais precisos, mas que isso pode acarretar em um custo maior em termos de tempo de processamento, por exemplo. Por isso, esta etapa serve para identificar possíveis subconjuntos de dados de interesse. Além disso, os dados podem assumir diferentes

formatos, como numérico, categórico e booleano. Atentar-se a isso e definir o tipo de tratamento que será utilizado para os diferentes tipos de dados pode evitar problemas de modelagem posteriormente (IBM, 2021).

Ao adentrar a etapa de exploração de dados, diversas técnicas de visualização, como gráficos de dispersão e histogramas, e estatísticas descritivas contendo métricas como média, mediana e desvio padrão são utilizadas para compreender mais profundamente a base de dados coletada. Isso facilita a identificação de relações entre atributos e fortalece a formulação de hipóteses (TUKEY *et al.*, 1977).

Por fim, a etapa de verificação da qualidade dos dados consiste em diagnosticar características do conjunto de dados que possam interferir no desenvolvimento de modelos mais precisos, como a presença de valores ausentes, inconsistências nos dados, erros de codificação, presença de outliers, erros de medição, metadados ruins, entre outras (IBM, 2021). Neste momento, são consolidadas estratégias para contornar tais características, podendo haver imputação de dados faltantes seguindo medidas personalizadas ou até mesmo o filtro ou remoção de um subconjunto de exemplos (PIPINO; LEE; WANG, 2002).

Ao fim da fase de Compreensão dos Dados, espera-se ter todas as fontes de dados claramente identificadas e acessadas, com suas devidas restrições e problemas identificados para que a fase seguinte (de Preparação dos Dados) possa executar tratamentos. Também é esperado neste ponto que os dados tenham sido devidamente coletados para as etapas seguintes, assim como a identificação dos atributos-chave que serão utilizados. Isso inclui a análise sobre utilização de todo o conjunto de dados ou subconjuntos do mesmo, que deriva do *trade-off* entre precisão e tempo de processamento citado anteriormente. Por meio de explorações detalhadas, é natural que os responsáveis pelo projeto tenham, a essa altura, um domínio muito mais acentuado sobre possíveis padrões e tenham formulado hipóteses iniciais cujas validações os levarão para mais perto da solução ideal.

2.3 Preparação dos Dados

Os dados raramente chegam perfeitos, pois carregam consigo a ambiguidade do mundo real. Por isso, um aforismo que permanece popular no universo da mineração de dados é o chamado GIGO ("garbage in, garbage out", a qual a tradução para o português é: "lixo entra, lixo sai") (PYLE, 1999). A ideia associada a ele é a de que dados frequentemente carregam distorções e características que interferem na produção de modelos assertivos e que, por esse motivo, é imprescindível um tratamento correspondente, o que associamos à etapa de Preparação dos Dados.

A preparação de dados é uma das fases mais importantes e possivelmente mais demoradas da mineração de dados, podendo chegar a até 70% do tempo e esforço de um projeto (IBM, 2021). A realização das fases de Entendimento do Negócio e Compreensão dos

Dados de forma bem feita pode reduzir esse custo, mas não eliminá-lo, pois é possível que os dados sejam coletados de fontes diferentes e precisam ser mesclados, ou seja necessária uma seleção de subconjuntos de amostra segundo critérios específicos. Também é natural que ocorra a criação de novos atributos derivados de outros, a classificação de dados para a modelagem, remoção ou substituição de valores em branco ou ausentes.

Segundo o guia de uso do CRISP-DM utilizado pela IBM, recomenda-se a divisão da fase de Preparação de dados nas seguintes etapas:

- Seleção: tomada de decisão sobre quais atributos e conjuntos de exemplos serão utilizados
- Limpeza: tratamentos dados aos dados faltantes, com erros ou inconsistentes
- Construção: derivação de atributos e geração de novos exemplos, caso necessário
- Integração: junção de conjuntos de dados com exemplos semelhantes mas com diferentes atributos
- Formatação: conversões de formato dos dados necessários aos modelos que serão aplicados

No que tange à formatação, vale destacar que, mesmo que os dados estejam estruturados (isto é, no formato de atributo-valor), é comum utilizar técnicas de codificação, como Label encoding e One-hot-Encoding, e normalização, como normalização min-max, para que os modelos processem corretamente, respectivamente, dados categóricos e numéricos. Em caso de dados não estruturados como textos extraídos de sinopses, por exemplo, outras técnicas ganham protagonismo, como a bag-of-words, em que esquemas de pesos como TF-IDF se tornam úteis. Com isso, se torna possível formatar esses dados de forma que modelos de aprendizado de máquina possam entender e processar. Abaixo, cada uma dessas técnicas será abordada detalhadamente.

2.3.1 Codificação

Nas atividades de aprendizado de máquina ou ciência de dados, é comum que o conjunto de dados contenha valores de texto ou categóricos (basicamente, valores não numéricos). Alguns algoritmos podem lidar muito bem com valores categóricos em dados como de gênero de filmes, mas a maioria deles espera valores numéricos para alcançar resultados de ponta (YADAV, 2019).

É sabido que existem diversas maneiras de converter valores categóricos em numéricos, cada uma com suas vantagens e desvantagens. Contudo, para o domínio deste trabalho, serão utilizados dois: One-hot encoding e Label encoder.

2.3.1.1 Label encoding

Essa abordagem é muito simples e envolve a conversão de cada valor em uma coluna para um número. Para exemplificar, vamos supor a seguinte lista de classificações etárias, que são as indicações de idade a partir da qual um conteúdo pode ser consumido: "Todas as idades", "Crianças", "Adolescentes", "Maiores de idade". Com a utilização do Label encoder, cada valor único nesta lista seria mapeado para um número diferente, por exemplo:

- Todas as idades - 0
- Crianças - 1
- Adolescentes - 2
- Maiores de idade - 3

Dessa forma, transformamos os valores categóricos em valores numéricos, permitindo que os algoritmos de aprendizado de máquina os utilizem de forma mais eficaz.

2.3.1.2 One-hot encoding

Ainda que o Label encoding funcione de forma simples, essa técnica tem a desvantagem de gerar valores que podem ser interpretados erroneamente pelos algoritmos. Isto acontece porque, ao assumir valores inteiros como 1, 2, 3, isso pode induzir a ideia de que representam algum tipo de ordem, grandeza ou hierarquia. Sabe-se que, uma vez que são valores categóricos, isso não faria sentido, portanto uma outra técnica se torna mais interessante: o One-hot encoding. Nesta técnica, cada categoria se torna uma coluna (atributo) cujos valores são binários, assumindo 0 quando o exemplo ou objeto tem aquele atributo ou característica, e 1 caso contrário. Utilizando como exemplo outro dado categórico presente no contexto deste trabalho, que são os gêneros dos animes, poderíamos obter algo como a Tabela 1:

Animes	Aventura	Fantasia	Drama	Comédia	Romance
Anime 1	1	0	0	1	1
Anime 2	0	1	1	0	0
Anime 3	1	0	0	0	1

Tabela 1 – Gêneros após One-hot encoding.

É interessante notar que essa técnica também contempla casos de uso onde um mesmo objeto contém vários atributos, que é exatamente o caso dos animes, uma vez que podem ter diferentes gêneros. A desvantagem nesse caso é o aumento da dimensionalidade dos dados, uma vez que, a depender do número de categorias, o número de colunas na

matriz atributo-valor pode crescer muito, o que pode demandar mais processamento na etapa de Modelagem.

2.3.2 Normalização

Em muitas ocasiões, os atributos estão em escalas diferentes e, portanto, podem não ser comparáveis entre si. Por exemplo, um atributo como ano de lançamento de um anime está em uma escala totalmente diferente da nota que esse anime recebeu de um usuário. Consequentemente, um modelo que receba essas duas informações em suas escalas originais provavelmente será fortemente influenciado pelo atributo de maior magnitude, principalmente se esse modelo utilizar medidas de proximidade para o aprendizado (AGGARWAL *et al.*, 2015).

Por isso, faz-se importante que os atributos passem por uma normalização, isto é, um processo de transformação dos dados com o objetivo de torná-los mais adequados para a aplicação de algum algoritmo de mineração (CASTRO; FERRARI, 2017). Existem vários tipos de normalização, de forma que a utilização de um ou de outro depende do seu contexto de aplicação. Neste trabalho, daremos foco à normalização Min-max.

A normalização Min-max realiza uma transformação linear nos dados originais de forma que os valores do atributo fiquem no intervalo [0, 1] (CASTRO; FERRARI, 2017). Suponha que $\max_a$ e $\min_a$ sejam, respectivamente, os valores máximo e mínimo de um determinado atributo a. A normalização Min-max procederá da seguinte forma com cada valor do atributo:

$$a_{\text{norm}} = \frac{a - \min_a}{\max_a - \min_a} \tag{2.1}$$

onde:

- a é o valor original do atributo
- $\min_a$ é o valor mínimo do atributo
- $\max_a$ é o valor máximo do atributo

2.3.3 Vetorização

Como dito anteriormente, a maioria dos algoritmos espera lidar com atributos numéricos. Porém, há atributos textuais não categóricos que podem ser de extrema importância e que não podem ser tratados com uma simples codificação. Citando como exemplo as sinopses de animes, não é de se surpreender que milhares de palavras sejam utilizadas, carregando semanticamente uma infinidade de significados. Nesse contexto, uma das formas de representação mais comuns é a do espaço-vetorial (AGGARWAL, 2014).

Neste modelo, cada instância (no exemplo mencionado, cada sinopse de anime) é constituída por um vetor, e cada dimensão desse vetor corresponde a um atributo do conjunto de dados (ROSSI, 2016). Usualmente, essas dimensões que são geradas com base nas palavras de textos como as sinopses recebem o nome de "termo", gerando uma matriz documento-termo denominada bag-of-words (ROSSI, 2016).

2.3.3.1 Bag-of-words

A bag-of-words é um modelo que se encaixa em múltiplos propósitos, como na seleção de atributos, classificação de documentos e imagens. Na classificação de documentos, é conhecido como um vetor do número de ocorrências de termos (QADER; AMEEN; AHMED, 2019). Os valores na matriz documento-termo são numéricos e representam o peso de um termo ou atributo em um documento (ROSSI, 2016).

Vale mencionar que uma matriz bag-of-words apresenta alta dimensionalidade e alta esparsidade, uma vez que há muitas palavras diferentes em uma coleção de textos, mas é natural que várias delas não se repitam em todos os documentos (ROSSI, 2016).

Para lidar com essa condição, técnicas de pré-processamento são fundamentais. Stemming e lematização, por exemplo, são importantes processos que têm o mesmo princípio básico: agrupar palavras semelhantes que possuem a mesma raiz ou a mesma forma canônica (PRAMANA *et al.*, 2022).

Enquanto o stemming corta o prefixo ou sufixo da palavra, a lematização considera a morfologia da palavra, o que pode adicionar significado à própria palavra. Por exemplo, no stemming, a palavra "amigos"(com o sufixo -os") ficaria cortada, resultando em "amig", o que não é uma palavra real em português. Por outro lado, a lematização identificaria que "amigos"é a forma plural da forma base "amigo"e converteria para a forma base ou dicionário, que neste caso é "amigo".

Outra técnica de pré-processamento importante é a remoção de *stopwords*, que são palavras que não adicionam valor nos padrões aprendidos por algoritmos de aprendizado de máquina, como preposições, pronomes e interjeições (ROSSI, 2016). Além delas, a depender do contexto, os documentos podem conter palavras específicas que se repetem muito, mas que não necessariamente adicionam valor para os resultados, como por exemplo a palavra "filme"em uma coleção de textos sobre cinema. Essas palavras podem ser consideradas *stopwords* de domínio (ROSSI, 2016) e podem causar um efeito negativo no desempenho.

Para atenuar esse tipo de efeito, é introduzida a frequência de documentos com o termo (df), que se diferencia da frequência do termo em uma coleção de documentos (cf). Vamos para um novo exemplo por meio da Tabela 2, construída a partir de um conjunto de sinopses de filmes hipotético:

A razão para preferir df a cf é ilustrada na tabela acima, onde um exemplo simples

Palavra	cf	df
pessoas	10422	8760
mistério	10440	3997

Tabela 2 – Frequência do termo na coleção de sinopses (*cf*) e frequência de sinopses com o termo (*df*). Adaptado de (SCHÜTZE; MANNING; RAGHAVAN, 2008).

mostra que a frequência do termo na coleção de filmes (*cf*) e a frequência do número de documentos com o termo (*df*) podem se comportar de maneira bastante diferente. Em particular, os valores de *cf* tanto para 'pessoas' quanto para 'mistério' são aproximadamente iguais, o que mostra que essas palavras são relativamente comuns neste domínio de aplicação. Contudo, seus valores de *df* diferem significativamente. Intuitivamente, é desejável que 'mistério' receba um peso diferente da palavra 'pessoas', por se tratar de um termo que qualifica melhor um conjunto específico de filmes, diferentemente de 'pessoas', que poderia ser considerada uma *stopword* de domínio.

Como demonstrado por meio do exemplo acima, os pesos que os termos recebem refletem quantitativamente a frequência de um termo em cada documento, e podem ser estabelecidos utilizando diferentes métodos. A seguir, serão detalhados dois entre os mais comuns, que são: (i) frequência do termo (do inglês term frequency - TF) e (ii) frequência do termo ponderada pelo inverso da frequência de documento (do inglês *term frequency - inverse document frequency* - TF-IDF) (ROSSI, 2016).

2.3.3.2 TF e TF-IDF

Além de verificar se um termo está contido em um texto ou não, é importante identificar qual é a frequência em que esse termo aparece em cada documento (SCHÜTZE; MANNING; RAGHAVAN, 2008). Uma abordagem simples, que ficou conhecida como frequência do termo (TF), consiste em atribuir ao peso o valor correspondente ao número de ocorrências do termo em cada documento (SCHÜTZE; MANNING; RAGHAVAN, 2008).

Por outro lado, a frequência inversa do documento (IDF) de um termo t é definida como:

$$\text{idf}_t = \log\left(\frac{N}{\text{df}_t}\right) \tag{2.2}$$

onde N é o número total de documentos na coleção e df_t é a frequência do documento do termo t.

E o esquema de ponderação TF-IDF atribui ao termo t um peso no documento d dado por:

$$\text{TF-IDF}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t \tag{2.3}$$

onde:

1. O peso é mais alto quando o termo t ocorre muitas vezes dentro de um pequeno número de documentos (assim, fornecendo um alto poder discriminatório para esses documentos);

2. O peso é menor quando o termo ocorre menos vezes em um documento, ou ocorre em muitos documentos (assim, oferecendo um sinal de relevância menos pronunciado);

3. O peso é o mais baixo quando o termo ocorre em virtualmente todos os documentos.

Esta métrica é útil porque palavras que aparecem frequentemente em um documento, mas não em muitos documentos da coleção, receberão um peso TF-IDF alto. Isso destaca palavras que são importantes ou distintas para um documento específico (RAMOS *et al.*, 2003).

O TF-IDF é amplamente aplicado em tarefas de processamento de linguagem natural. Apesar de sua eficácia em muitas tarefas, o TF-IDF tem suas limitações. Ele não considera a ordem ou a estrutura semântica das palavras em um documento, o que pode ser crucial para algumas análises (TURNEY; PANTEL, 2010).

2.4 Modelagem

Neste momento, os dados que passaram por um processo de preparação frequentemente extenso finalmente podem ser utilizados para treinar e testar os modelos escolhidos para o experimento. Também é um momento propício para, a partir dos primeiros testes, verificar se os objetivos estabelecidos na fase de Entendimento do Negócio são factíveis (IBM, 2021).

É importante que sejam utilizados parâmetros padronizados para que os testes sejam comparáveis entre si, mas é raro que um único modelo ou execução traga respostas satisfatórias, o que torna essa fase repleta de iterações, assim como a de preparação de dados. Dessa forma, ajustes são constantemente feitos e novos testes são realizados para identificar oportunidades de melhoria. Dado que o objetivo deste trabalho é o estudo e avaliação de técnicas para a recomendação de animes, o referencial teórico da Modelagem focará em modelos para serem aplicados em sistemas de recomendação. Mais detalhes sobre os sistemas de recomendação e modelos serão apresentados a seguir.

2.4.1 Sistemas de recomendação

Atualmente, as plataformas de streaming oferecem uma vasta gama de conteúdos. Na Netflix, por exemplo, há mais de 3.300 filmes e 1.850 séries disponíveis (JúNIOR DANIEL MATTOS, 2022). Entretanto, essa grande oferta de conteúdos traz um risco relacionado ao funcionamento da mente dos consumidores. Com tantas opções à disposição, é comum sentir-se indeciso sobre o que consumir, o que pode estimular o usuário a desistir

e sair da plataforma. O que a ciência mostra é que isso se relaciona com um fenômeno conhecido como Paradoxo da Escolha, descrito por Schwartz (2005).

Esse fenômeno descreve como a quantidade excessiva de alternativas, em vez de gerar uma sensação de liberdade nos consumidores, os sobrecarrega cognitivamente à medida que tentam considerar cada alternativa disponível. Isso leva a uma paralisia na tomada de decisão, que ocorre quando a decisão sobre uma alternativa se torna tão angustiante que um indivíduo pode decidir não escolher (SCHWARTZ, 2005).

Além disso, mesmo após tomar uma decisão, a pessoa pode sentir-se menos satisfeita, atormentada pela dúvida de que outra opção poderia ter sido melhor. Afinal, considerando o tempo como um recurso limitado, cada escolha de um anime para ser assistido representa uma renúncia de investir o mesmo tempo em outras opções igualmente atraentes. Essa renúncia, por sua vez, gera uma experiência negativa uma vez que a mente humana é fortemente avessa à perda (KAHNEMAN; TVERSKY, 1979).

Para contornar essa situação, faz-se necessário o uso de sistemas de filtragem de informação (conhecidos como sistemas de recomendação) que exibam de forma inteligente as opções de conteúdo que as pessoas têm acesso. Uma vez que uma boa experiência de consumo e compra é fornecida, adaptando-se às necessidades de cada indivíduo, os consumidores se beneficiam por ter um acesso personalizado e um processo decisório menos angustiante. Consequentemente, as empresas observam a retenção de clientes crescer, bem como outros indicadores-chave favoráveis ao negócio (KNAUER, 2019). É nesse contexto em que sistemas de recomendação tiveram sua importância acentuada nos últimos anos.

O objetivo mais básico dos sistemas de recomendação é utilizar várias fontes de dados para inferir os interesses dos clientes. Dessa forma, conseguem prever preferências e até avaliações que um indivíduo dará a um produto ou conteúdo, o que facilita a filtragem de quais opções devem ser expostas e em qual ordem (AHUJA; SOLANKI; NAYYAR, 2019).

O sujeito que recebe a recomendação é referido como usuário, enquanto o conteúdo ou produto sendo recomendado é chamado de item (AGGARWAL *et al.*, 2016). Em suma, modelos de recomendação frequentemente se baseiam nas interações entre essas duas entidades para gerar as recomendações, uma vez que inclinações e preferências demonstradas no passado podem ser bons indicadores de escolhas no futuro (AGGARWAL *et al.*, 2016). É possível agrupar os tipos de dados que alimentam esses modelos em dois grupos, a saber:

1. Interações entre usuário e item, que podem tomar a forma de avaliações, e;

2. Informações sobre usuários, como características pessoais, e itens, como atributos do produto ou conteúdo.

Nesse contexto, a capacidade de reter feedbacks de usuários tem sido um catalisador para que sistemas de recomendação sejam desenvolvidos. Existem feedbacks explícitos, como por exemplo avaliações formais que um cliente fornece e curtidas, e implícitos, que estão mais relacionados ao comportamento, como o tempo gasto em uma tela em específico (AGGARWAL *et al.*, 2016). No que tange ao feedback explícito, existem diversas abordagens. A Netflix, por exemplo, possibilita que os usuários deem suas avaliações para cada conteúdo na forma de três alternativas: "amei", "gostei"e "não é para mim". Já a Crunchyroll utiliza uma forma de avaliação mais tradicional, pautada em cinco estrelas, como na Figura 2. Existem, ainda, plataformas que aceitam avaliações como notas de 1 a 10, como é o caso da MyAnimeList.

Figura 2 – Exemplo de avaliação baseada em 5 estrelas. Fonte: Adaptado de (AGGARWAL *et al.*, 2016)

Existem métodos que usam essencialmente o primeiro grupo para gerar as inferências, como a filtragem colaborativa, e outros que se especializam no segundo, como a filtragem baseada em conteúdo (AGGARWAL *et al.*, 2016). Também existem abordagens híbridas, que podem mesclar características de métodos diferentes (BURKE, 2003), e abordagens sem personalização, também conhecidas como baseadas em popularidade (JI *et al.*, 2020). Cada um desses métodos é explicado mais detalhadamente a seguir.

2.4.1.1 Sistemas baseados em popularidade

Provavelmente um dos métodos de recomendação mais simples, a abordagem baseada em popularidade recomenda itens que são muito consumidos e/ou muito bem avaliados. Como um método de recomendação fácil de implementar e não personalizado, é amplamente utilizado como *baseline* para fornecer um desempenho de referência para um sistema de recomendação (JI *et al.*, 2020).

Para que seja bem compreendido o que diferencia um sistema de recomendação baseado em popularidade dos demais é importante, em primeiro lugar, determinar o que pode ser considerado popularidade. Muitos estudos definem e avaliam "popularidade"como

uma medida baseada no número de interações de usuários com os itens de um conjunto de dados. Em outras palavras, os itens recomendados aos usuários neste tipo de sistema de recomendação são aqueles com o maior número de interações nos dados de treinamento (JI *et al.*, 2020). Como exemplo, podemos citar a plataforma MyAnimeList, onde há uma extensa base de animes que os usuários podem avaliar e guardar em suas listas pessoais. Nela, o ranking de popularidade (chamado de 'Most popular') é baseado no número de usuários que adicionaram um determinado anime em sua lista, o que é consonante com a definição apresentada.

Contudo, existem outros parâmetros que podem ser utilizados além da quantidade de usuários que interagiram, como a nota que esses usuários deram ao item. Nesses casos, no entanto, um questionamento emerge: seria razoável um anime com uma nota média de 9,8 com 100 avaliações estar ranqueado em uma posição superior a um que tenha a nota média 9,2 com 500 mil avaliações? Para responder a esse tipo de questionamento, rankings de popularidade utilizam um estimador Bayesiano para ponderar as avaliações de forma mais sofisticada (IMDB, 2023).

Antes de aprofundar a explicação, é importante estabelecer algumas premissas estatísticas:

1. Suponha que a nota média real de um anime $\overline{W}$ é desconhecida e, portanto, precisa ser inferida;

2. Um anime que tem muitas avaliações provavelmente tem uma nota média W muito similar à nota média real $\overline{W}$;

3. Um anime que tem poucas avaliações não tem uma nota média confiável, pois é mais provável que uma amostra pequena de avaliações seja fortemente viesada.

Com isso, se faz necessário inferir qual seria a nota média que um anime pouco avaliado obteria caso tivesse muitas avaliações. Esta seria uma forma razoável de comparar essas produções em pé de igualdade, levando em consideração não apenas a nota média para ranquear, mas também uma ponderação que leva em consideração o número de avaliações que o anime recebeu. Para tanto, rankings dos melhores filmes como o famoso Internet Movie Database (IMDb) e dos melhores animes como o da MyAnimeList se utilizam de um Estimador Bayesiano (IMDB, 2023), que é uma técnica de inferência estatística para estimar parâmetros desconhecidos de um modelo (COLES; JR, 2023). A fórmula que fornece uma estimativa Bayesiana das notas médias reais de cada produção informada pelas próprias plataformas é a que segue:

$$W = \frac{(v \times R) + (C \times m)}{v + m}$$

Onde:

W é a estimativa da nota média real, ponderada pelo número de avaliações,
R é a nota média atual da produção como um número de 1 a 10,
v é o número de avaliações que a produção recebeu até o momento,
m o mínimo de avaliações que uma produção deve ter para entrar no ranking,
C é a média dos votos em todo o conjunto de filmes ou animes.

Note que W é apenas a média ponderada de R e C com o vetor de peso (v, m). À medida que o número de avaliações ultrapassa m, a confiança da nota média ultrapassa a confiança da nota média de todas as produções (C), e a estimativa Bayesiana da nota ponderada (W) se aproxima de uma média simples (R). Quanto mais próximo v (o número de avaliações para o filme) estiver de zero, mais próximo W estará de C, isto é, da nota média de todos os filmes.

Portanto, animes com pouquíssimas avaliações terão uma nota ponderada em direção à média de todas as notas, enquanto animes com muitas avaliações terão uma nota ponderada em direção à sua nota média de fato. Essa abordagem garante que um anime com apenas algumas avaliações, mesmo que todas com nota 10, não seja classificado acima de "Fullmetal Alchemist: Brotherhood", por exemplo, com uma média de 9.09 obtida a partir de mais de 3 milhões de avaliações.

Tanto rankings gerados pela definição mais tradicional de popularidade (baseada apenas em número de interações) quanto aqueles que destacam produções bem avaliadas por meio de uma nota média ponderada podem ser utilizados como referência para sistemas de recomendação baseados em popularidade, mas as plataformas têm liberdade para inserir outros fatores também.

Isso é demonstrado por meio da Figura 3, onde é apresentado um conjunto de recomendações baseadas em popularidade nas plataformas Crunchyroll e Netflix. Vale notar que aspectos temporais podem ser levados em consideração, como é o caso demonstrado na Netflix sob o título "top 10 em séries hoje". Ainda que de forma mais implícita, todos os animes sugeridos nas recomendações populares da Crunchyroll receberam episódios recentemente, o que demonstra que o fator temporal de recência também influenciou o tratamento de popularidade nesse caso.

Porém, como citado anteriormente, a limitação desse tipo de sistema é a falta de personalização, uma vez que sugere a nível individual os mesmos itens para todos os usuários. Ainda assim, uma vez que as métricas que formam esses rankings são um forte indicativo do que usuários em geral gostam, este tipo de sistema continua sendo utilizado até hoje em plataformas de streaming e serve como uma excelente linha de base para comparação com os demais métodos que serão avaliados.

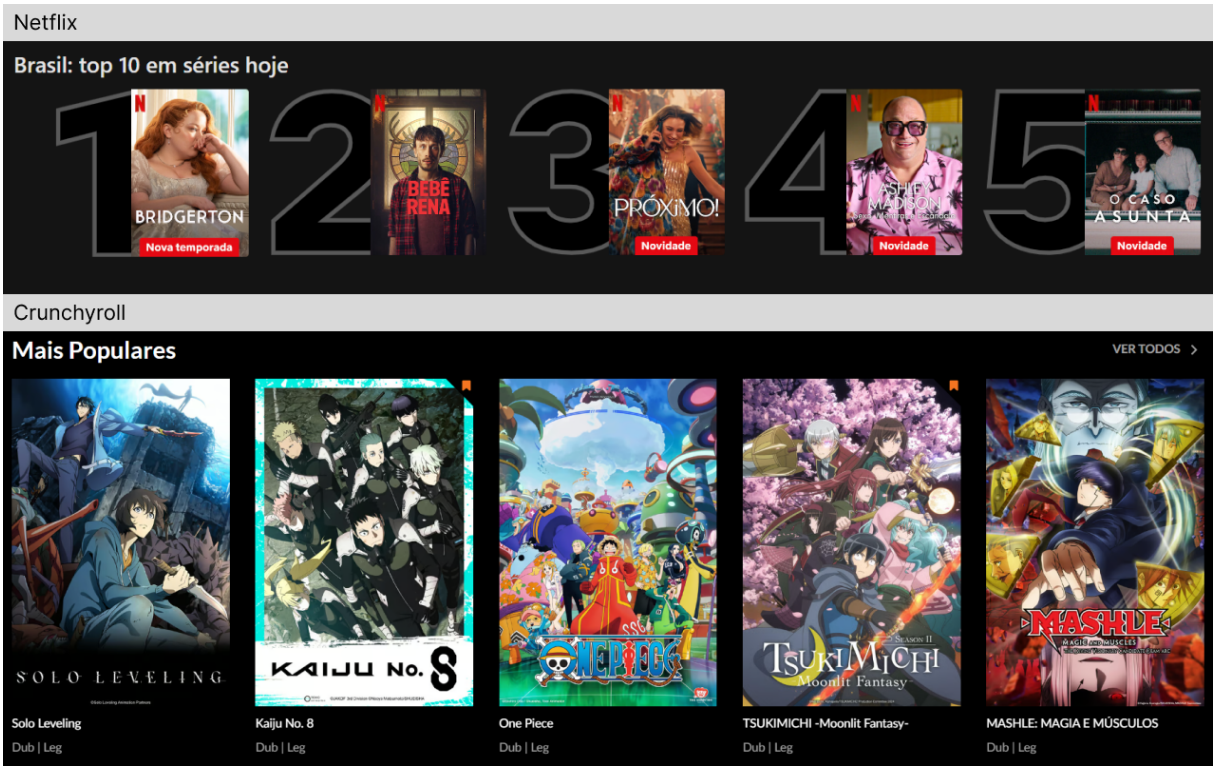


Figura 3 – Exemplos de recomendações baseadas em popularidade. Fontes: Netflix e Crunchyroll.

2.4.1.2 Sistemas de filtragem colaborativa

Uma das formas de recomendar um novo item a um usuário é prever o valor da avaliação que esse usuário daria a um item. Nesse contexto, "filtragem colaborativa"emerge como um termo que se refere ao uso de avaliações de vários usuários de forma colaborativa para prever avaliações ausentes (AGGARWAL *et al.*, 2016).

Na prática, uma premissa básica desse tipo de abordagem é que, se uma pessoa A tem a mesma opinião que uma pessoa B sobre um determinado item, A é mais propensa a compartilhar a opinião de B sobre outro item do que a opinião de uma pessoa escolhida aleatoriamente (MASE; OHWADA, 2012).

Dessa forma, algoritmos de filtragem colaborativa geralmente requerem (i) a participação ativa dos usuários, (ii) uma maneira fácil de representar os interesses dos usuários e (iii) algoritmos capazes de agrupar pessoas com interesses semelhantes (MADHUKAR, 2014).

Assumindo que há dados prévios sobre as avaliações que usuários deram a diferentes itens, é possível construir uma matriz de dimensões m x n onde há m usuários nas linhas e n itens nas colunas. Isto é, os valores dessa matriz são as avaliações que forneceram, constituindo uma matriz usuário-item.

No entanto, uma vez que nem todos os usuários terão dado avaliações para um anime, assim como nem todos os os animes terão avaliações de todos os usuários, trata-se de uma matriz incompleta, onde há valores a preencher desconhecidos, conforme ilustrado no exemplo da Tabela 3.

Tabela 3 – Avaliação de Animes pelos Usuários. Fonte: adaptado de (JENA *et al.*, 2022)

Usuários	Anime X	Anime Y	Anime Z
A	10	10	2
B	10		2
C		9	
D	3	1	9
E	4	1	

Na Tabela 3, encontram-se três animes (X, Y e Z) nas colunas e cinco usuários (A, B, C, D e E) nas linhas. Cada par usuário-item representa uma avaliação que pode variar de 1 a 10 ou estar vazia caso o usuário não tenha avaliado o anime, possivelmente por não ter assistido (JENA *et al.*, 2022).

Por meio das avaliações, é perceptível que os usuários A e B da Tabela 3 parecem ter gostos por animes mais semelhantes do que B e D. Sendo assim, ao inferir qual nota o usuário B daria para o anime Y, é mais razoável esperar uma avaliação mais semelhante à de A do que de D. Portanto, avaliações não conhecidas podem ser inferidas utilizando tanto avaliações quanto o grau de semelhança entre usuários e itens (MUTTEPPAGOL, 2021).

Os métodos de filtragem colaborativa se ramificam em dois tipos: os baseados em memória e os baseados em modelo, que têm suas características particulares, conforme a seguir.

2.4.1.2.1 Baseados em memória

O termo "baseado em memória"se refere ao fato de que manifestações de preferências passadas podem ser bons indicadores de escolhas futuras. Neste caso, a similaridade entre os usuários ou itens é calculada utilizando dados de avaliações dadas anteriormente. Se o foco do sistema é utilizar a similaridade entre os usuários para recomendar novos itens para os mesmos, então o modelo de filtragem colaborativa é chamado de baseado em usuário ou usuário-usuário. Caso o foco do sistema seja a similaridade entre os itens para encontrar itens similares ao que o usuário gostou e recomendá-los, é referido como baseado em itens ou item-item (MUTTEPPAGOL, 2021).

Para que a similaridade seja calculada, pode-se utilizar diferentes medidas. Entre as mais utilizadas, a medida de Similaridade de Cosseno foi escolhida como a medida de similaridade comum a todos os modelos experimentados neste trabalho devido à sua

versatilidade especialmente no tratamento de textos e dados esparsos. Essa medida, que possibilita o cálculo da distância entre dois vetores a partir do valor do cosseno do ângulo entre eles, é definida da seguinte forma (AGGARWAL *et al.*, 2016):

$$\text{Cosine}(X, Y) = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2} \cdot \sqrt{\sum_{i=1}^{d} y_i^2}}$$

onde:

- X e Y são os dois vetores;
- d é a dimensão dos vetores;
- x_i e y_i são os elementos dos dois vetores respectivamente.

A facilidade de implementação e interpretabilidade de sistemas baseados em memória as tornam vantajosas. Por outro lado, é especialmente desafiador lidar com uma matriz usuário-item muito esparsa. A depender do volume do histórico de avaliações, pode ser difícil encontrar usuários suficientemente similares, o que dificulta prever de forma robusta a avaliação que daria para um anime específico (AGGARWAL *et al.*, 2016).

Além disso, quando o sistema ainda não coletou dados sobre a preferência de um usuário, há uma dificuldade em realizar recomendações assertivas, o que é conhecido na literatura como *cold start problem* ("problema de arranque frio", traduzido literalmente). Essa dificuldade também representa uma desvantagem para esse tipo de modelo (AGGARWAL *et al.*, 2016).

Na abordagem baseada em usuário (*user-based*), as avaliações fornecidas por usuários com preferências similares ao usuário alvo são usadas para as predições de nota de um item. Ou seja, uma das principais tarefas é determinar quais dos outros usuários são mais similares ao usuário alvo, e gerar uma predição da avaliação de um determinado item com base nas avaliações dos usuários mais similares (AGGARWAL *et al.*, 2016). Em outras palavras, as etapas típicas para esse método seriam:

1. Calcular similaridade entre os usuários a partir das avaliações dadas;
2. Localizar os k usuários mais similares ao usuário alvo que assistiram o anime alvo;
3. Verificar as avaliações dadas por esses usuários ao anime alvo;
4. Usar essas avaliações para prever a nota que o usuário alvo dará para o item.

Na Figura 4 é apresentada uma ilustração da predição de nota de um anime baseado nas notas dos usuários mais similares. Nesta ilustração, Alice e Vanessa são os usuários

mais similares a Jonathan. Utilizou-se então as notas atribuídas por ambos os usuários (10 e 8) para predizer a nota que Jonathan daria ao anime, que neste exemplo é 9 e corresponde à uma média simples das notas de Alice e Vanessa.

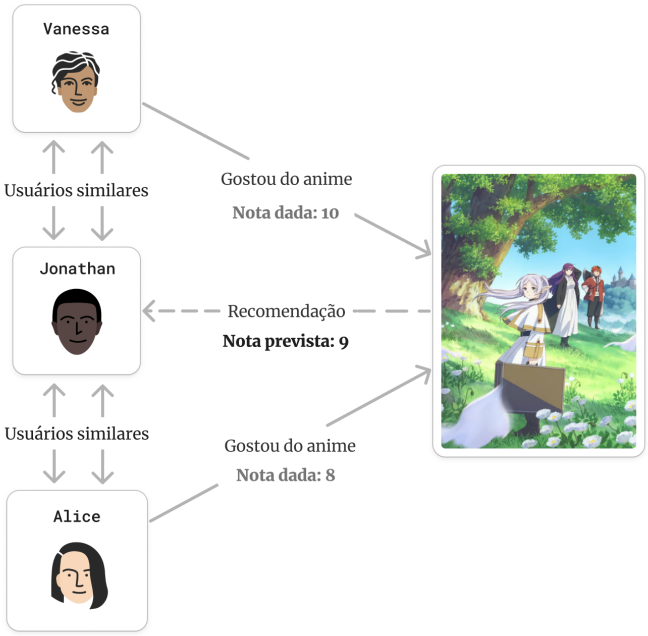


Figura 4 – Filtragem colaborativa baseada em usuários. Fonte: o autor.

Para prever a avaliação $R_{U,I}$ de um usuário U para um item I, uma abordagem interessante é calcular a média das avaliações dadas pelos k usuários mais similares ponderada pelas suas respectivas similaridades (AJITSARIA, 2022):

$$R_{U,I} = \frac{\sum_{u=1}^{k} R_{u,I} * S_u}{\sum_{u=1}^{k} S_u}$$

onde:

- R_U, I é a nota prevista para o item i a ser dada pelo usuário U;
- u representa um usuário similar;
- k é a quantidade de usuários entre os mais similares que serão utilizados;
- $R_{u,I}$ é a avaliação de um usuário similar u deu para o item I;
- S_u é a similaridade entre u e U.

Por outro lado, na filtragem colaborativa baseada em itens (*item-based*), o que importa é a similaridade entre os itens a partir das avaliações que eles receberam anteriormente. Ou seja, quando deseja-se prever a nota para um determinado par usuário-item, as referências deixam de ser as notas dadas por usuários similares, que são substituídas pelas notas dadas pelo próprio usuário a outros itens similares (AGGARWAL *et al.*, 2016).

Na Figura 5 é apresentada uma ilustração da predição de nota de um anime baseado nas notas que o usuário deu a outros animes similares. Nesta ilustração, utilizou-se as notas atribuídas pelo usuário aos animes similares (10 e 8) para predizer a nota que ele daria ao anime alvo ainda não assistido, que neste exemplo é 9 e corresponde à uma média simples das notas dadas aos animes similares.

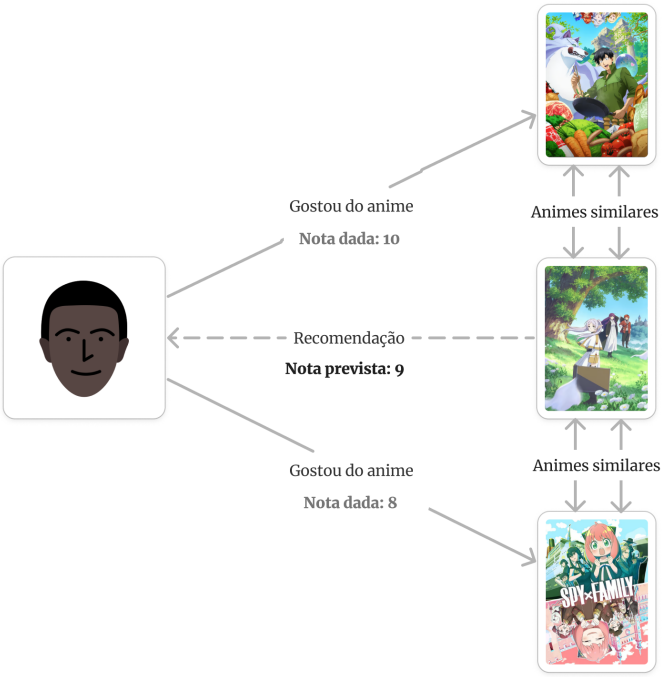


Figura 5 – Filtragem colaborativa baseada em itens. Fonte: o autor.

Na prática, para prever se um anime deve ser recomendado para um determinado usuário e qual seria a nota predita, o seguinte procedimento pode ser adotado:

1. Calcular similaridade entre os itens a partir das avaliações recebidas;
2. Localizar os k itens mais similares ao anime alvo que o usuário tenha assistido;
3. Verificar as avaliações que o usuário alvo deu a esses itens;

4. Usar essas avaliações para prever a nota que o usuário alvo dará para o item alvo.

Analogamente, a abordagem de notas preditas ponderadas pela similaridade pode ser aplicada aqui na seguinte forma (SARWAR *et al.*, 2001):

$$R_{U,I} = \frac{\sum_{i=1}^{k} R_{U,i} * S_i}{\sum_{i=1}^{k} S_i}$$

onde:

- $R_{U,I}$ é a nota prevista para o item I a ser dada pelo usuário U;
- i representa um item similar;
- k é a quantidade de itens entre os mais similares que serão utilizados;
- $R_{U,i}$ é a avaliação que U deu ao item similar i;
- S_i é a similaridade entre i e I.

2.4.1.2.2 Baseados em modelo

Por fim, o método de filtragem colaborativa baseado em modelo (*model-based*) se difere dos apresentados anteriormente, pois as avaliações aqui são previstas utilizando diferentes tipos de algoritmos de aprendizado de máquina. Nesse caso, em vez de se basear apenas na memorização das interações passadas dos usuários, constrói-se um modelo matemático para entender as relações entre usuários e itens (MUTTEPPAGOL, 2021).

Os sistemas de recomendação baseados em modelo possuem várias vantagens importantes em comparação com métodos baseados em memória. Em termos de eficiência no uso de espaço, geralmente o tamanho do modelo aprendido é muito menor que a matriz original de avaliações, o que reduz significativamente os requisitos de espaço. Outra vantagem é a velocidade de treinamento e predição: sistemas baseados em modelo geralmente são muito mais rápidos na fase de pré-processamento para construir o modelo treinado (AGGARWAL *et al.*, 2016).

Entre os estados-da-arte em filtragem colaborativa baseada em modelo, encontra-se a fatoração de matrizes (AGGARWAL *et al.*, 2016). A ideia-chave desses modelos é a de que é possível estimar matrizes com menor dimensionalidade completamente preenchidas e representativas das correlações anteriores a partir de matrizes incompletas como a matriz usuário-item mencionada anteriormente (AGGARWAL *et al.*, 2016).

Por meio do aproveitamento das correlações de linha e coluna da matriz original, essa matriz resultante é composta dos fatores latentes da mesma, isto é, características não observadas inferidas pelo modelo. Isso ajuda a lidar com um dos principais problemas

de outros métodos de filtragem colaborativa: a alta esparsidade dos dados em matrizes de vetores de alta dimensão (MUTTEPPAGOL, 2021).

Algumas das realizações mais bem-sucedidas de modelos de fatores latentes são baseadas na decomposição de matriz (KOREN; BELL; VOLINSKY, 2009). Em sua forma básica, a decomposição de matriz caracteriza tanto itens quanto usuários por vetores de fatores latentes (inferidos) a partir da matriz utilizada na filtragem colaborativa. Esses fatores latentes capturam características ocultas nos dados de avaliação e permitem uma representação mais compacta e informativa dos itens e usuários, facilitando a previsão de avaliações de itens não avaliados por usuários (KOREN; BELL; VOLINSKY, 2009).

Na Figura 6, demonstra-se como os dados de avaliação que são resultantes da interação entre usuários e animes (matriz R) podem ser representados como o produto da matriz que captura características latentes dos usuários (X^T) pela que captura fatores latentes de animes (Y).

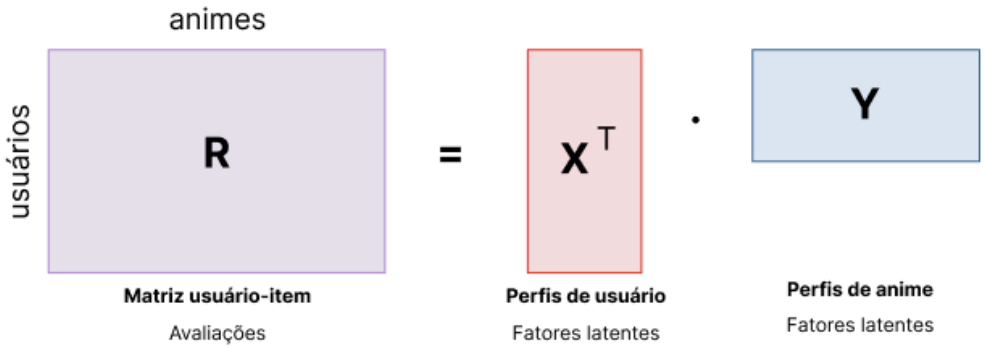


Figura 6 – Dados de avaliações de animes como um produto de fatores latentes. Fonte: adaptado de (DIELEMAN, 2016).

Um exemplo simplificado é exibido na Figura 7. Nela, os fatores latentes referem-se às preferências indicadas pelos eixos horizontal e vertical. Seis usuários estão embutidos no espaço de fatores, onde se enquadram em uma das quatro categorias: sério, escapista, voltado para homens e voltado para mulheres. Da mesma forma, as características dos animes referem-se às características das categorias de usuários. A alta correspondência entre os fatores do item e do usuário leva a uma recomendação (KOREN; BELL; VOLINSKY, 2009).

Esses métodos se tornaram populares nos últimos anos por combinar boa escalabilidade com precisão preditiva. Além disso, oferecem muita flexibilidade para modelar várias situações da vida real. Normalmente, o feedback explícito compreende uma matriz esparsa, uma vez que é provável que qualquer usuário tenha avaliado apenas uma pequena porcentagem dos itens possíveis (KOREN; BELL; VOLINSKY, 2009).

Os modelos de fatoração de matrizes mapeiam tanto os usuários quanto os itens

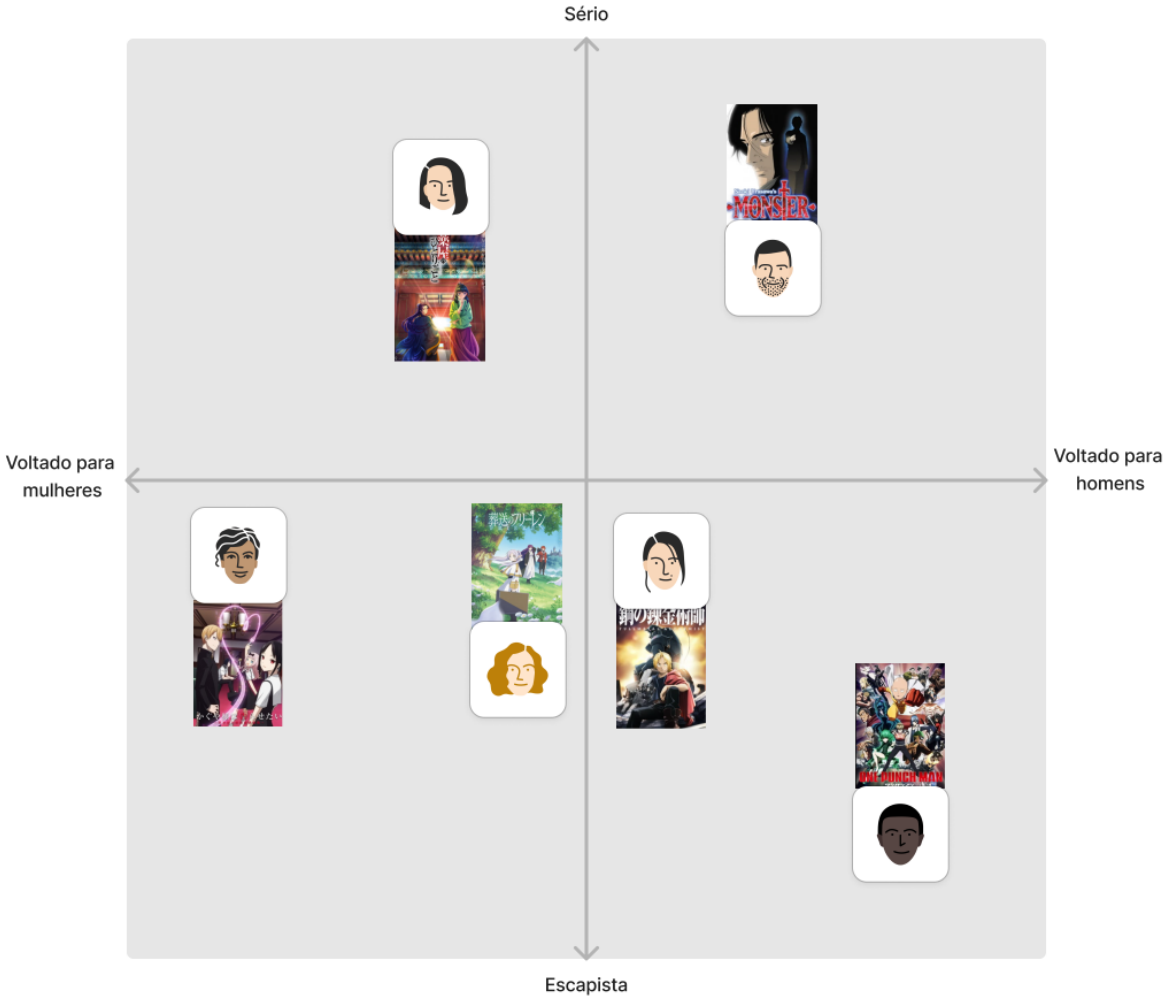


Figura 7 – Uma ilustração simplificada da abordagem de fator latente. Fonte: Adaptado de (KOREN; BELL; VOLINSKY, 2009).

para um espaço de fator latente conjunto de dimensionalidade f, de modo que as interações usuário-item sejam modeladas como produtos internos nesse espaço. Dessa forma, cada item i está associado a um vetor q_i, e cada usuário u está associado a um vetor p_u, ambos pertencendo a R^f (KOREN; BELL; VOLINSKY, 2009). Com isso, temos que a previsão de uma avaliação de u ao item i, que é denotada por r_{ui}, pode ser dada através da estimativa:

$$r_{ui} = q_i^T p_u$$

Isto é, valores altos nas mesmas posições dos dois vetores q_i^T e p_u vão gerar predições de notas também altas, o que indica um forte candidato a recomendação.

Sendo assim, o principal desafio é calcular os cruzamentos de cada item com cada usuário usando os vetores de fatores $q_i ep_u$. Depois que o sistema de recomendação completa esse mapeamento, ele pode facilmente estimar a avaliação que um usuário dará a qualquer

item usando a equação anterior (KOREN; BELL; VOLINSKY, 2009).

Vale mencionar que, conforme exposto acima, a fatoração de matrizes pode ser usada para predições de notas, mas as representações geradas também podem ser utilizadas por algoritmos de aprendizado de máquina e/ou abordagens *user-based* e *item-based* mencionadas anteriormente.

Além disso, modelos mais modernos têm aplicado em conjunto algoritmos como o Gradiente Descendente Estocástico (*Stochastic Gradient Descent* - SGD) para minimização de erro entre a previsão e a avaliação real. Com isso, é possível adicionar uma penalização para evitar *overfitting* (ajuste excessivo aos dados de treinamento), cujos parâmetros são gradualmente ajustados (KOREN; BELL; VOLINSKY, 2009).

2.4.1.3 Sistemas baseados em conteúdo

Modelos de filtragem colaborativa consideram apenas as interações entre usuários e itens, e não consideram atributos do item ou do usuário para fazerem as predições. Por outro lado, nos sistemas de recomendação baseados em conteúdo, as avaliações e o comportamento de consumo dos usuários são combinados com as informações de conteúdo disponíveis nos itens para fazer recomendações (AGGARWAL *et al.*, 2016).

O termo "conteúdo", nesse contexto, refere-se a atributos que, no campo dos animes, podem ser textos oriundos das sinopses, gênero, ano de lançamento, classificação etária indicativa, imagem de capa, entre outros. A partir da extração e tratamento dessas características, é possível gerar matrizes de similaridade entre os itens, que são utilizadas para gerar as recomendações.

Em outras palavras, sistemas baseados em conteúdo não necessitam das avaliações de outros usuários para operar. Afinal, o funcionamento básico exige apenas que sejam conhecidas as características dos itens e o grau de similaridade entre si para que sejam geradas recomendações, como mostra a Figura 8, que ilustra recomendações baseadas em outro anime assistido pelo usuário.

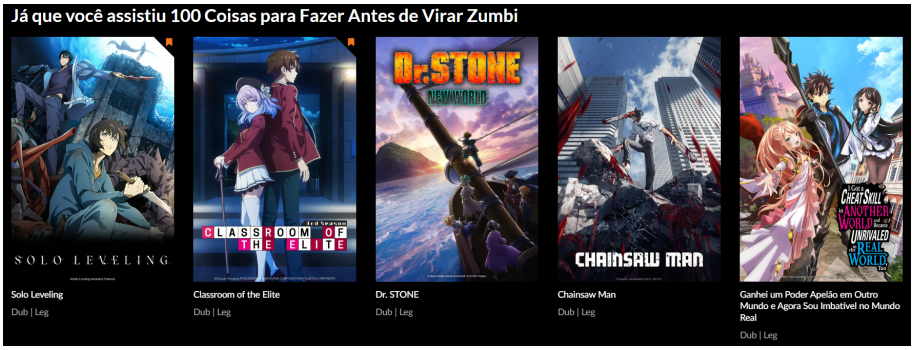


Figura 8 – Recomendações semelhantes a anime assistido. Fonte: Crunchyroll.

Como os sistemas baseados em conteúdo trabalham com uma ampla variedade de

descrições de itens e conhecimento sobre usuários, é necessário converter esses diferentes tipos de dados não estruturados em descrições padronizadas. Os sistemas baseados em conteúdo operam principalmente, mas não exclusivamente, no domínio de texto. Muitas aplicações naturais de sistemas baseados em conteúdo também são centradas em texto. Por exemplo, sistemas de recomendação de animes podem ser baseados em títulos e sinopses. Em geral, métodos de classificação e modelagem de regressão de texto permanecem as ferramentas mais amplamente utilizadas para criar sistemas de recomendação baseados em conteúdo (AGGARWAL *et al.*, 2016).

A depender do tipo de atributo que é utilizado para a similaridade ser calculada, é possível gerar inúmeras categorias de sugestões para os usuários, o que possibilita que plataformas de streaming organizem o conteúdo recomendado de forma inovadora, isto é, para além das divisões tradicionais por gênero. O exemplo da Figura 9 ilustra bem isso, uma vez que exibe recomendações de animes baseadas essencialmente na forma como o título é escrito.

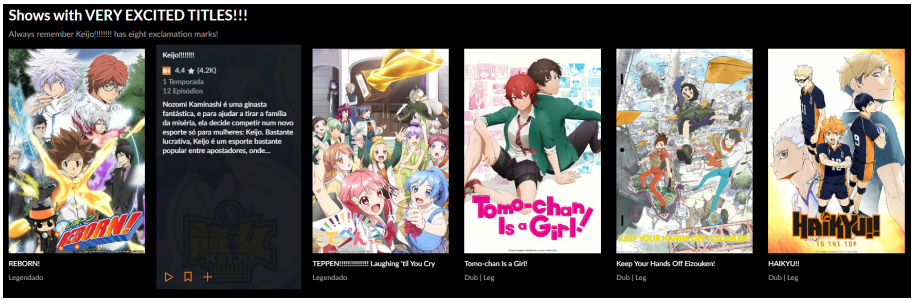


Figura 9 – Bloco de sugestões baseadas em conteúdo. Fonte: Crunchyroll.

Para cada usuário, os dados de treinamento correspondem às descrições dos itens que avaliou. A variável de classe (ou dependente) corresponde às avaliações dadas. Esses dados de treino são usados para criar um modelo de classificação ou regressão, que é específico para o usuário em questão. Esse modelo específico do usuário é usado para prever se o indivíduo irá gostar de um item para o qual sua avaliação é desconhecida (AGGARWAL *et al.*, 2016).

2.4.1.3.1 K-Nearest Neighbors

O algoritmo de aprendizado supervisionado dos K-ésimos Vizinhos mais Próximos (*K-Nearest Neighbors* - k-NN) tem sua utilidade extensamente reconhecida para projetar sistemas de recomendação. Na verdade, é muito comum que uma introdução sobre sistemas de recomendação contenha ao menos uma visão geral sobre o uso do k-NN nesse contexto (NEAMAH; EL-AMEER, 2018).

O primeiro passo na construção desse tipo de sistema é definir uma função de similaridade, que é usada no classificador de vizinho mais próximo. Essa medida de

similaridade é útil para fazer previsões para itens nos quais a avaliação do usuário é desconhecida. Para cada item, seus k-vizinhos mais próximos são determinados. Com isso, o valor médio das avaliações que os k vizinhos receberam pode ser determinado. Esse valor médio, por fim, servirá como a avaliação prevista para o item alvo. Também é possível adicionar ponderações para essa previsão considerando os valores de similaridade, por exemplo (AGGARWAL *et al.*, 2016).

Há muitas vantagens em utilizá-lo, como a boa interpretabilidade, a facilidade de aplicação e versatilidade para se encaixar em problemas tanto de classificação quanto de regressão. A principal distinção é que a classificação é usada para valores discretos, enquanto a regressão é usada para valores contínuos (IBM, 2022). Além disso, ainda que a abordagem k-NN seja simples e intuitiva, ela pode apresentar resultados com alta precisão e ainda ser ajustável para melhorias (NEAMAH; EL-AMEER, 2018).

Como desvantagem, pode ser citada a dificuldade de lidar com escala. À medida que o conjunto de dados cresce, o k-NN torna-se cada vez mais ineficiente, comprometendo o desempenho geral do modelo. Afinal, ele faz parte de uma família de modelos de "aprendizado lento"(*lazy learning*), o que significa que todo cálculo é feito apenas quando uma classificação ou previsão está sendo feita, funcionamento contrário a algoritmos da família *eager learning*, onde o sistema tem uma fase de treinamento propriamente dita que é mais custosa, mas que ocasiona em respostas mais rápidas e eficientes mediante novos *inputs* (IBM, 2022).

Por essa característica, em comparação com outros classificadores, o k-NN tende a ocupar mais memória, o que pode custar caro tanto do ponto de vista de tempo quanto de dinheiro (IBM, 2022). Além disso, tende a ter dificuldade de lidar com dados de entrada de alta dimensionalidade e, consequentemente, pode ser considerado mais propenso ao overfitting (ajuste excessivo aos dados de treinamento) (IBM, 2022).

É importante ressaltar que o k-NN em si não está vinculado necessariamente a um ou outro tipo de sistema de recomendação, como de filtragem colaborativa ou baseada em conteúdo. Conforme mencionado em relação à versatilidade do algoritmo, é possível adaptá-lo a diferentes contextos. Contudo, no domínio deste trabalho, o k-NN está sendo destacado como um potencializador de sistemas de recomendação baseados em conteúdo por possibilitar a classificação de vizinhos mais próximos por meio de atributos dos itens.

2.4.1.4 Sistemas híbridos

Nas seções anteriores, foram expostos três dos principais métodos de recomendação: baseados em popularidade, que recomendam sem personalização os itens mais bem avaliados pelos usuários, colaborativos, que usam as interações usuário-item como forma de identificar similaridade entre usuários e itens, e baseados em conteúdo, que por sua vez utilizam atributos dos itens para calcular a similaridade e realizar recomendações.

Diante dessa diversidade de abordagens, emerge a possibilidade de construção de sistemas híbridos que utilizam características de diferentes métodos a fim de potencializar o desempenho e atingir resultados mais satisfatórios. A título de curiosidade, duas famosas propostas vencedoras no concurso Netflix Prize para aprimoramento do algoritmo de recomendação, conhecidas como "Bellkor's Pragmatic Chaos"e "The Ensemble", eram sistemas sistemas híbridos (AGGARWAL *et al.*, 2016).

Existem três principais maneiras de criar sistemas de recomendação híbridos (AGGARWAL *et al.*, 2016):

1. Design de conjunto (*ensemble*): neste design, os resultados de algoritmos prontos são combinados em uma saída única e mais robusta;

2. Design monolítico: nesse caso, uma recomendação integrada é produzida de forma que pode ser difícil até mesmo visualizar os métodos usados separadamente;

3. Design misto: as recomendações são produzidas por algoritmos diferentes como no design de conjunto, mas apresentadas ao mesmo tempo ao usuário.

Além disso, existem os sistemas com design de conjunto (*ensemble*) cujos recomendadores funcionam de forma paralela, isto é, as previsões individuais são combinadas no final. E também há aqueles em que a saída de um recomendador é usada como entrada para outro, isto é, funcionam de maneira sequencial (AGGARWAL *et al.*, 2016). Os sistemas híbridos que ocorrem de forma paralela e que combinam as avaliações preditas ao fim por meio de uma média ou outras técnicas estatísticas de composição de valor podem ser considerados do tipo ponderado (BURKE, 2003).

Na Figura 10, há um esquema que ilustra o funcionamento de um sistema de recomendação híbrido do tipo *ensemble* paralelo ponderado. Nesse tipo de sistema, os *inputs* do modelo abastecem q recomendadores, que atuam paralelamente, gerando diferentes valores de previsão (P_q), que em seguida compõem o valor final da predição (*output*).

Além de abordagens de *ensemble* serem conhecidas por combinar modelos de diferentes características e obter resultados melhores que modelos individuais (AGGARWAL *et al.*, 2015), no caso de sistemas de recomendação é possível combinar o conhecimento provido pelas interações entre usuário item, e o conteúdo dos itens e perfis dos usuários, o que aumenta a robustez do sistema como um todo.

2.5 Avaliação

Um sistema de recomendação pode ser avaliado a partir de diferentes aspectos como i) o nível de novidade, que está relacionado à capacidade de recomendar ao usuário itens não vistos no passado, ii) a diversidade, que por sua vez se relaciona com o desempenho

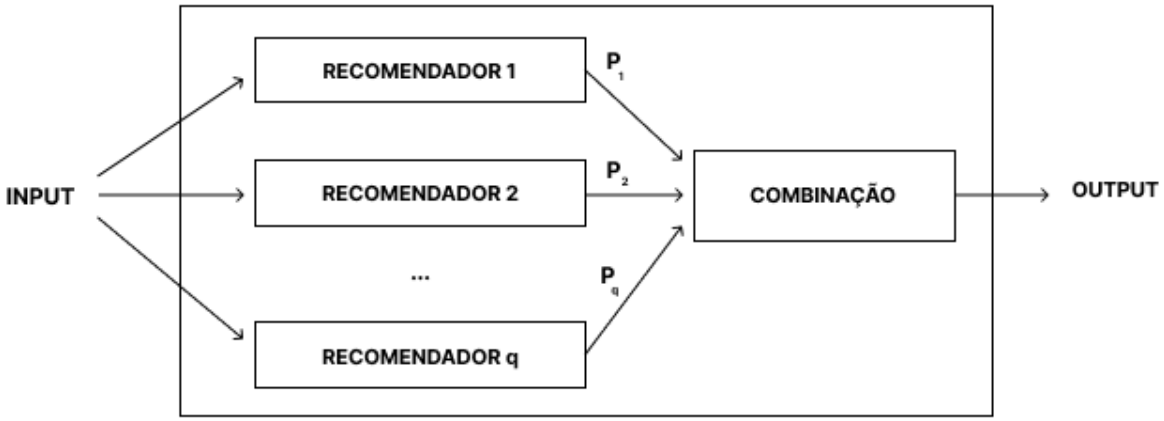


Figura 10 – Exemplo de design de sistema de recomendação do tipo ensemble paralelo ponderado. Fonte: Adaptado de (AGGARWAL *et al.*, 2016).

do sistema ao recomendar uma lista de itens diferentes, mas relevantes e iii) a acurácia do modelo, que diz respeito ao quão assertivas as recomendações geradas foram a partir das ações realizadas pelo usuário (consumir e/ou avaliar bem o item) (AGGARWAL *et al.*, 2016).

Neste trabalho, a recomendação foi avaliada sob a ótica da acurácia, em que é feita uma comparação da predição dos modelos com a nota de fato atribuída pelo usuário a um determinado item no passado.

Os modelos propostos neste trabalho serão avaliados e analisados por meio de duas métricas de qualidade: raiz do erro quadrático médio (Root Mean Squared Error - RMSE) e erro absoluto médio (Mean Absolute Error - MAE). Para ambas as métricas, quanto menor for o valor, melhor foi o desempenho do modelo avaliado.

2.5.0.1 RMSE

A raiz do erro quadrático médio (RMSE) tem como base o mesmo cálculo do erro quadrático médio (MSE), que é uma métrica de qualidade que permite calcular a média da diferença entre o valor predito e o real. Em outras palavras, para cada par de avaliações reais y_i e preditas $\hat{y}_i$, calcula-se a diferença de uma pela outra, o que por sua vez é elevado ao quadrado e depois dividido pelo número de pares n. Essa exponenciação atua como uma forma de penalizar aqueles valores que se distanciarem muito do real e de gerar sempre valores positivos, possibilitando uma boa análise da qualidade do modelo. O cálculo se dá por meio da seguinte equação:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

.

Contudo, para facilitar a interpretabilidade, no cálculo do RMSE é aplicada uma raiz quadrática para que os resultados estejam na mesma escala do dado original (HODSON, 2022). Portanto, o RMSE é dado por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

2.5.0.2 MAE

Assim como as métricas anteriores, o erro absoluto médio também se baseia na média das diferenças entre os valores reais e preditos. Porém, devido ao fato de haver valores positivos e negativos nessas operações de subtração, é adicionado um módulo. Outra diferença é que essa métrica não contém a exponenciação ao quadrado, o que permite que os valores de saída tenham a mesma escala dos dados originais, além de serem menos sensíveis aos valores discrepantes (outliers). O MAE é dado por:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

2.6 Implantação

Após a fase de Modelagem e Avaliação, o projeto chega a uma última fase: a implantação. Essa fase marca a transição do modelo do ambiente controlado do laboratório para o mundo real, onde estará à disposição de usuários finais. Esse momento pode apresentar diversos desafios em etapas específicas como na integração com a estrutura existente e na construção dos pipelines de dados. Além disso, a forma como o sistema de recomendação se encaixa na experiência dos usuários depende de uma interface adequada e intuitiva.

Um dos maiores desafios da implantação é garantir a performance contínua do modelo ao longo do tempo. O mundo real é dinâmico e os padrões nos dados podem mudar, exigindo monitoramento constante e reavaliação periódica do desempenho. É comum que, para se adaptar à nova realidade em produção, sejam necessários ajustes.

A escalabilidade também é um fator relevante a ser considerado. O modelo precisa ser capaz de lidar com grandes volumes de dados em tempo real, sem comprometer seu desempenho. Isso pode ser especialmente desafiador para alguns métodos de recomendação do que para outros, conforme visto anteriormente. Com isso, podem ser elaboradas estratégias para que diferentes recomendadores entrem em cena a depender da etapa da jornada que os usuários estiverem. Por exemplo, para usuários novos cujas preferências ainda não são conhecidas, sistemas baseados em popularidade podem ter mais protagonismo. Conforme exibem suas decisões e fornecem avaliações, é possível utilizar métodos mais

robustos que permitam a identificação das preferências individuais e, portanto, gerem recomendações personalizadas.

Sobretudo, o gerenciamento de recursos deve ser cuidadoso para que a implantação seja feita de forma eficiente e integrada. Ao superar esses desafios, um sistema de recomendação pode se tornar uma ferramenta poderosa, auxiliando na melhora da experiência dos usuários e na alavancagem de indicadores de negócio favoráveis.

3 TRABALHOS RELACIONADOS

Encontrar os algoritmos de recomendação mais pertinentes para as plataformas de streaming tem gerado ricas investigações, mas é notável que ainda há pouco interesse sobre a aplicação desses algoritmos ao contexto específico de animes (YAO *et al.*, 2021). Adicionalmente, os estudos que se concentraram nessa tarefa encontraram resultados diversos, o que reforça a necessidade de aprofundar essas investigações avaliando múltiplas soluções de recomendação.

Para serem incluídos nesta revisão, os trabalhos relacionados precisaram atender aos seguintes critérios:

- Escrita na língua inglesa ou em português;
- Foco em sistemas de recomendação no nicho de animes;
- Haver descrição de datasets utilizados;
- Haver detalhamentos sobre como foi o pré-processamento;
- Indicações sobre qual ou quais modelos foram empregados;
- Exposições sobre quais foram os resultados obtidos e as métricas utilizadas.

Foram realizadas duas pesquisas no Google Acadêmico com as palavras-chave "recommender system anime"e "sistema de recomendação anime". Os dez primeiros itens de cada pesquisa, ordenados por grau de relevância com relação às palavras-chave, sofreram uma primeira filtragem baseada no título e resumo, resultando em 7 artigos para análise mais profunda derivados da pesquisa em inglês e 4 derivados da pesquisa em português. Mediante análise do texto completo, foi possível identificar se todos os pré-requisitos listados acima foram atendidos, o que resultou em mais um filtro e gerou a seleção de artigos que serão detalhados a seguir.

No trabalho apresentado por Silva (2021), dois conjuntos de dados foram obtidos. O primeiro contendo 12.294 animes e suas informações (como título e gênero) e o segundo com 73.516 usuários distintos e 69.600 avaliações dadas pelos usuários a animes que assistiram. Esses dados foram retirados da plataforma MyAnimeList (MAL) por meio de *web scraping* e disponibilizados na plataforma Kaggle. O objetivo do estudo era avaliar a performance de diferentes técnicas de recomendação, o que inclui recomendações baseadas em popularidade, em conteúdo e colaborativas baseadas em item e em usuário. Para tanto, foi preciso tratar os dados, que passaram por um processo de normalização, além da retirada de usuários que avaliaram menos de 50 animes e animes que receberam menos

de 100 avaliações. Para construir o sistema de recomendação baseado em conteúdo, foi utilizado o K-Means como forma de agrupar os animes com base no gênero, que tiveram suas medidas de similaridade calculadas por meio da distância euclidiana.

No caso da recomendação colaborativa baseada em itens, foi utilizada a Similaridade de Cosseno para encontrar os itens mais similares a cada item previamente avaliado, enquanto que na recomendação colaborativa baseada no usuário, o k-NN foi aplicado para encontrar os usuários mais semelhantes, que tiveram sua similaridade medida pela correlação de Pearson. A filtragem por popularidade, por fim, fez predições de nota baseadas na média de avaliações de cada anime. Para medir os resultados, foram utilizadas as medidas MAE e MSE para avaliação de valores preditos ao passo que uma matriz de confusão e análise de curva ROC foram utilizados para avaliar as classificações.

O sistema de filtragem colaborativa baseado em itens apresentou os menores erros, mas as medidas utilizadas resultaram em valores muito próximos em todos os modelos, o que impossibilitou em primeiro momento a rejeição de algum. Porém, ao analisar a curva ROC, foi possível perceber que a filtragem por popularidade e o modelo baseado em conteúdo tiveram melhores desempenhos. Em específico o modelo de filtragem colaborativa baseada em usuário não se mostrou adequado para o domínio de aplicação, por ter demonstrado uma performance pouco superior a um sistema hipotético que fizesse recomendações aleatórias, além de ser menos eficiente computacionalmente.

Já o estudo conduzido por Yao *et al.* (2021) chegou a conclusões diferentes. Sua hipótese inicial era a de que os usuários escolhiam o que assistir pautados primordialmente na sinopse da obra e sua imagem de capa. Sendo assim, um sistema que tivesse esses parâmetros (dados não estruturados) como principais condutores de recomendações poderia apresentar melhores resultados. Para testar esse cenário, os autores reuniram conjuntos de dados com 81.727 observações de usuários, 192.112 de avaliações e 19.311 contendo metadata de cada anime, obtidos via *web scraping* da MAL. Também foram reunidas 17.335 imagens de capa extraídas da mesma plataforma. O pré-processamento envolveu a remoção de duplicatas, a limitação do escopo de análise com base em alguns pré-requisitos como a exclusão de tipos de anime que não passaram na TV (como filmes e musicais) e que não contivessem uma sinopse válida, assim como usuários que deram uma ou nenhuma avaliação.

Foram testados 5 modelos: três deles baseados em conteúdo e que foram treinados usando diferentes *features* para o cálculo da similaridade como a sinopse e a imagem da capa. Outros dois, de filtragem colaborativa, foram treinados usando avaliações de usuários: um com aplicação do algoritmo SVD (*Singular Value Decomposition*) e outro filtrando com um *Autoencoder*. Para mensurar os resultados, foram realizadas separações de treino e teste e a métrica Hit Rate foi usada como medida em comum para comparar as performances dos diferentes modelos. Nesse estudo, os modelos de filtragem colaborativa apresentaram

um desempenho melhor em comparação com os modelos baseados em conteúdo, com destaque especial para a utilização do Autoencoder, que alcançou a taxa de acerto mais alta entre os cinco modelos testados.

O estudo desenvolvido por (NUURSHADIEQ; WIBOWO, 2020) também ambicionou recomendar animes utilizando mais do que o id de usuários e animes, defendendo a importância de outras informações secundárias de usuários (como a idade) e de animes (como a sinopse). Para validar sua hipótese, reuniu 301.136 avaliações de 116.126 usuários únicos oriundos do MyAnimeList. O pré-processamento envolveu diversas etapas por lidar com dados de diferentes tipos, a saber: discretização de informações secundárias de usuários, *embeddings*, vetorização de texto, GAP e LSTM para redução de dimensionalidade e vetorização. As tabelas do dataset foram combinadas em uma só tabela de forma que, para cada usuário e rating, também houvesse as informações secundárias do usuário e do anime sendo consideradas no treinamento dos modelos.

Para a filtragem colaborativa, foi utilizada a Similaridade de Cosseno, mas também foram testados modelos de filtragem baseados em conteúdo usando Redes Neurais. Modelos baseados em SVD e k-NN foram usados para fins de comparação de resultados mediante testes que foram feitos com um conjunto de teste separado via *cross validation*. MSE e RMSE foram usados como medidas de mensuração de resultado, de forma que comprovou-se que a proposta de um modelo de filtragem colaborativa utilizando dados secundários de usuários em conjunto de técnicas de *deep learning* apresenta menor erro em comparação com os outros modelos.

A utilização de *deep learning* também foi defendida no estudo de (MUTTEPPAGOL, 2021) que, assim como os outros estudos citados, obteve seus conjuntos de dados por meio de extrações do MyAnimeList. Depois da limpeza de dados, que envolveu remoção de valores vazios e de colunas desnecessárias, fusões de tabelas, transformações de valores de datas e gênero, remoção de duplicatas, *encoding*, *embedding* e *batch normalization*, restaram 65.941 usuários únicos e 17.559 animes únicos. A proposta do estudo era mostrar que um modelo de filtragem colaborativa baseado em *deep learning* teria uma performance superior a outros modelos de filtragem colaborativa como os baseados em usuários e em item, o que se mostrou comprovado. Foi realizada mensuração por meio de MSE e MAE, de forma que mesmo os problemas de cold-start e esparsidade dos dados foram resolvidos pela proposta.

4 METODOLOGIA

Neste trabalho, adotou-se o método de Mineração de Dados CRISP-DM, conforme descrito no capítulo 2. Cada seção a seguir detalha uma etapa do método CRISP-DM.

4.1 Entendimento do negócio

No contexto de aplicação, estamos falando de animes, as animações japonesas. Diferentes de animações americanas como as da Disney, que têm um foco maior no público infantil, animes também endereçam tópicos mais maduros e encantam o público adulto em todo o mundo (YAO *et al.*, 2021). Vale ressaltar que animes ganharam popularidade global apenas recentemente, porém, mesmo com uma recência, uma quantidade cada vez maior de animes vem sendo produzida.

A grande quantidade traz consigo a dificuldade em buscar uma opção ideal para assistir. Para isso, sistemas de recomendação podem ser uma funcionalidade valiosa em plataformas de streaming com esse tipo de conteúdo. Porém, a popularidade recente de animes torna mais desafiador o emprego de sistemas de recomendação, já que esse tipo de conteúdo em comparação com outros que já acumularam dados de visualização historicamente e, portanto, podem alimentar de forma mais volumosa modelos de aprendizado de máquina que identificam preferências (JENA *et al.*, 2022). Além disso, a área carece de estudos de diferentes modelagens dos dados e tipos de recomendação.

Por esse motivo, os objetivos deste trabalho são: i) coletar dados de animes que tenham diversos metadados associados, ii) testar e avaliar diferentes formatos e algoritmos para esse sistema de recomendação, e iii) prever com a maior assertividade possível as notas que usuários dariam a animes recomendados.

4.2 Compreensão dos Dados

A etapa de Compreensão de Dados é fundamental para aprofundar o contexto do desafio levantado na etapa de Entendimento do Negócio, bem como para identificar atributos mais promissores e os mecanismos pelos quais eles podem ser acessados e importados (PYLE, 1999). Os objetivos desta seção estão organizados nas seguintes 4 subetapas: i) coleta inicial; ii) descrição dos dados; iii) exploração dos dados; e iv) verificação da qualidade dos dados (IBM, 2022).

4.2.1 Coleta de dados inicial

Os dados foram baixados diretamente da Kaggle[1], plataforma da Google onde ocorrem competições de ciência de dados e aprendizado de máquina. Foram três arquivos com dados brutos extraídos da plataforma MyAnimeList (MAL), uma comunidade online e banco de dados popular para entusiastas de anime e mangá de todo o mundo. A plataforma fornece informações valiosas sobre animes, usuários e as pontuações (avaliações) que esses usuários forneceram para os animes assistidos.

4.2.2 Descrição dos dados

Os dados coletados se subdividem em três conjuntos, a saber: i) conjunto de dados de animes com seus metadados, ii) conjunto de dados de usuários e iii) conjunto de avaliações. Estão incluídos dados até 06/10/2023, com 24.905, 731.290 e 24.325.191 exemplos em cada conjunto respectivamente.

O conjunto de dados de animes inicialmente contava com 24 atributos que descreviam cada anime, como por exemplo título, sinopse e gênero. Já o dos usuários consistia em atributos demográficos como gênero e região, e de consumo, como total de animes assistidos e nota média dada, totalizando 16 colunas. Por fim, o conjunto de dados de avaliações combina em cada linha um par usuário-anime demonstrando qual nota um determinado usuário u deu para um anime a, com 5 colunas contendo dados de identificação das duas entidades.

4.2.3 Exploração dos dados

A fim de potencializar a formulação de hipóteses e fundamentar a escolha de atributos-chave, serão analisadas na etapa de Exploração de dados características dos animes e dos usuários, bem como tendências nas avaliações.

Como ponto de partida, observa-se o crescimento da popularidade dos animes por meio da Figura 11, onde as barras verticais representam a quantidade de animes lançados em cada ano. Percebe-se que o crescimento da produção se tornou cada vez maior, alcançando seu auge histórico em 2017. Essa informação justifica a relevância de desenvolver sistemas que se adequem à entrega e recomendação desse tipo de conteúdo.

[1] <https://www.kaggle.com/>

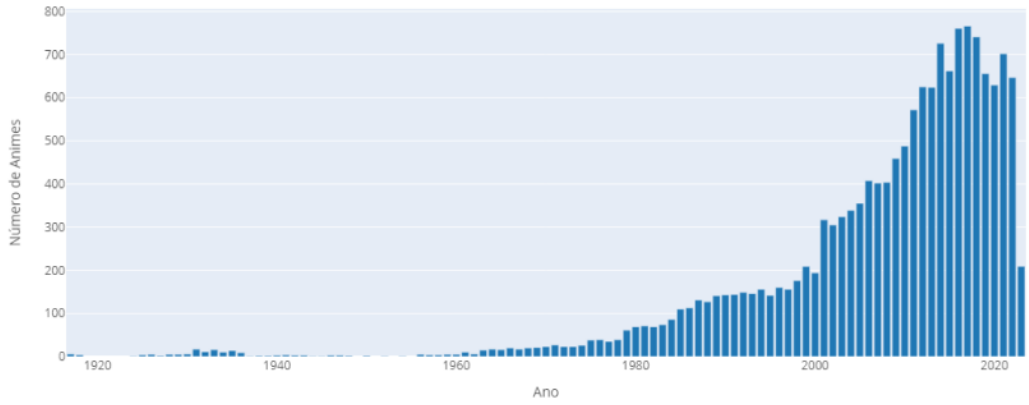


Figura 11 – Número de animes lançados por ano com base na MyAnimeList. Fonte: o autor.

Outro característica notável no conjunto de dados de animes se evidencia ao avaliar as notas médias que cada anime recebeu em conjunto com a quantidade de pessoas que avaliaram (Figura 12). Em geral, a maior parte das avaliações se situa entre 7 e 8,5. O mesmo fenômeno já havia sido identificado no trabalho realizado por Mutteppagol (2021) e pode significar que há um viés inerente nas avaliações, pois as opiniões dos espectadores são frequentemente influenciadas pela massa, o que é referenciado como "comportamento de manada"(*herd behaviour*) na literatura científica (BADDELEY, 2010).

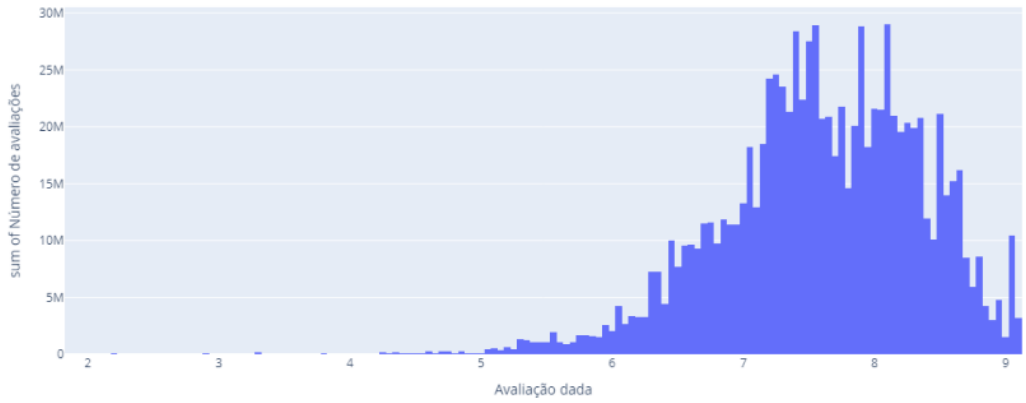
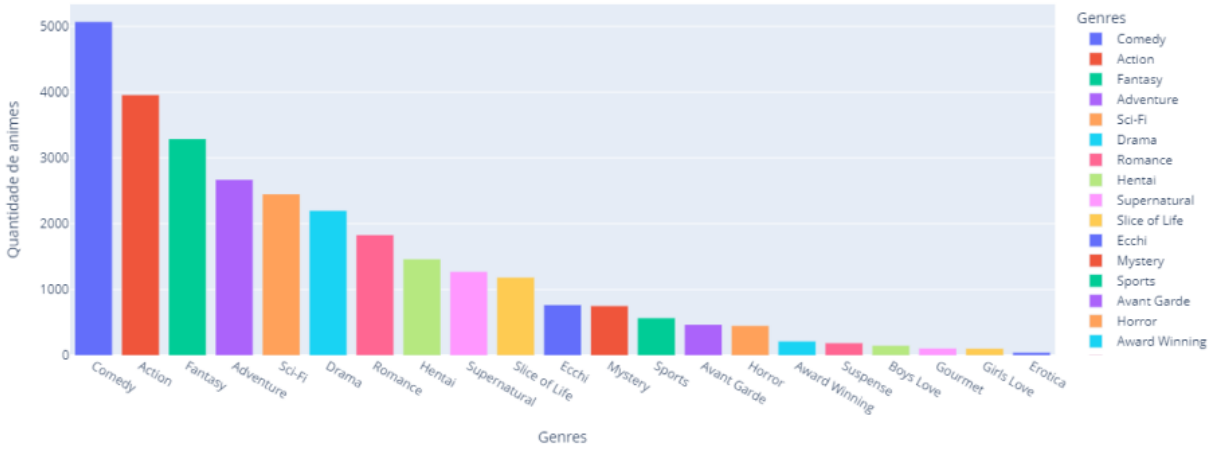


Figura 12 – Quantidade de avaliações recebidas versus nota média do anime com base na MyAnimeList. Fonte: o autor.

Ao todo, foram identificados 21 gêneros diferentes de anime e a quantidade de animes em cada um pode ser verificada por meio da Figura 4.2.3. Os gêneros mais populares são "Comédia", "Ação"e "Fantasia", presentes em 42% dos animes. Identificar quais são os gêneros disponíveis possibilitou a análise sobre aqueles que se caracterizam pelo conteúdo adulto. Animes desses segmentos são considerados de nicho e recomendações dos mesmos em plataformas de público genérico podem ofender usuários. Sendo assim, os exemplos desse tipo serão removidos da base na etapa de Preparação dos Dados.



Compreender sobre o público a quem se destina a recomendação também pode ser determinante na elaboração dos sistemas. A partir dos dados obtidos, é possível identificar pela base de dados qual a participação de pessoas que se declara como público masculino, feminino ou não-binário. Conforme demonstrado na Figura 13, a maior parte (56%) do público da MAL é masculino.

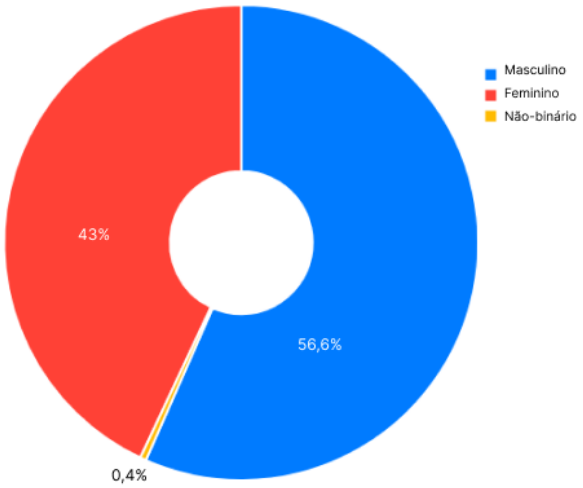


Figura 13 – Participação de usuários por gênero com base na MyAnimeList. Fonte: o autor.

Em termos de idade, pode-se observar na Figura 14 que a maior parte dos usuários está entre os 30 e 34 anos, seguido pelo segundo maior grupo, composto por pessoas de 35 a 39 anos de idade. Isso reforça o quanto esse tipo de conteúdo mobiliza o público adulto, mas também não dispensa o fato de que a plataforma existe desde 2006 (WHOIS. . . , 2024) e pode não ter se mantido popular entre o público mais jovem nos últimos anos. É importante citar que não há qualquer restrição para um indivíduo menor de idade se cadastrar na MAL.

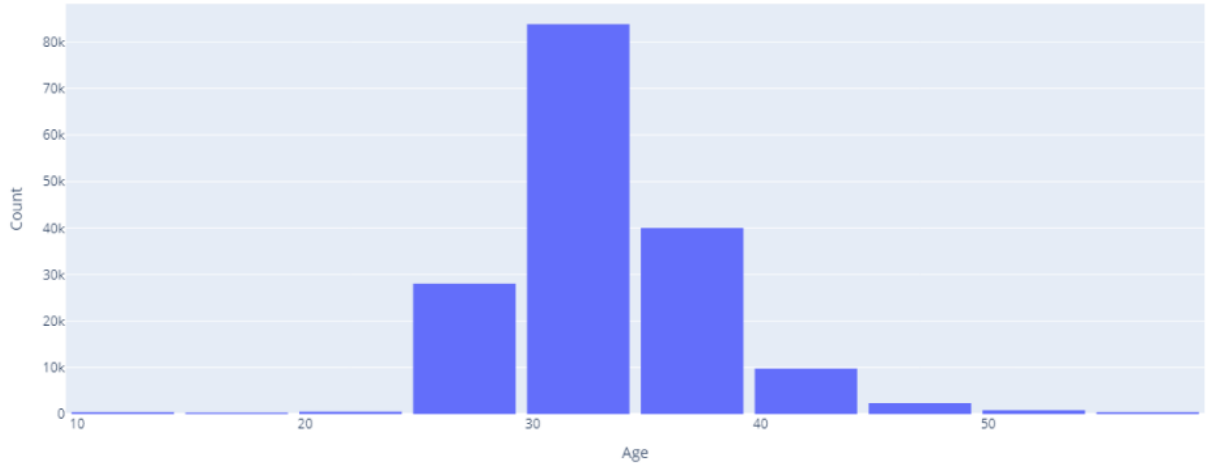


Figura 14 – Distribuição do público por faixa de idade com base na MyAnimeList. Fonte: o autor.

4.2.4 Verificação da qualidade dos dados

O conjunto de dados de avaliações não contém valores nulos nos atributos de identificação de animes e usuários e nem no atributo de avaliação ("*rating*"). Além disso, notou-se que as avaliações têm formato numérico que varia de 1 a 10. A identificação dos animes e dos usuários se dá por um "id"que também é um valor numérico.

Um fator relevante notado no conjunto de dados de avaliações é que ele é muito extenso, contendo mais de vinte milhões de exemplos, o que pode ser desafiador para a memória no processamento dos modelos. Ao todo, são 270.033 usuários e 16.500 animes únicos. Porém, 56% dos usuários deram menos de 50 avaliações e 52% dos animes receberam menos de 100 avaliações. A remoção desses exemplos pode ser determinante para obter algoritmos mais precisos a um custo menor de processamento.

No conjunto de dados de animes, foram identificados cerca de 19% dos animes com valores nulos em gênero e 37% sem nota média. Além disso, o atributo *"Aired"*, que se refere ao período quando o anime foi exibido, contém valores em múltiplos formatos como por exemplo "Apr 3, 1998 to Apr 24, 1999", "Sep 1, 2001", "2002", entre outros, o que exige um tratamento adequado para obter o ano de lançamento.

Os dados de sinopse estão totalmente preenchidos com textos descritivos de cada anime, o que significa que devem passar por algum processo de vetorização antes de serem usados por um algoritmo de aprendizado de máquina. Também não há valores nulos no atributo de popularidade do anime (*"Popularity"*), que diz respeito à posição que o item ocupa no *ranking* de popularidade próprio da MAL e que é representada por um número inteiro.

Já o atributo que diz respeito à classificação etária indicativa de cada anime

apresenta cerca de 2,7% dos exemplos com valor nulo. Por se tratar de um atributo importante para sinalizar quando um item contém conteúdo explícito e pornográfico, deve ser tratado com atenção visto a necessidade de filtro desses itens na etapa de Preparação de dados. Os valores assumidos por esse atributo são textuais e categóricos e podem precisar de algum tipo de codificação para serem processados por algoritmos de aprendizado de máquina.

Por fim, no conjunto de dados de usuários, verificou-se que 69% dos usuários não contêm um valor para gênero, 77% não possuem para data de aniversário (usada para cálculo da idade) e 79% para localidade. Além disso, os dados de data de aniversário vêm em formato completo, como "1991-01-22T00:00:00+00:00", o que exige manipulação até que se chegue a uma faixa etária.

Outro problema no conjunto dos usuários está nos valores assumidos pelo atributo de localização. Frequentemente, os usuários manipulam as respostas preenchendo com um valor irreconhecível (como um conjunto de caracteres especiais) ou expressões que não denotam um local existente (como "AnimeLand"). A falta de formato predefinido também implica em muitas respostas válidas que não são consistentes. Por exemplo, há respostas que contêm o nome de um estado, país, ou até o endereço completo. Dada a complexidade e falta de confiabilidade, o atributo deve ser descartado.

A partir das análises realizadas, foi possível identificar problemas e possíveis riscos para a etapa de Modelagem, bem como direcionamentos para a etapa de Preparação de dados. O conjunto de dados sobre usuários não será utilizado para fins de Modelagem devido ao fato de conter valores nulos na maior parte do conjunto em todos os atributos que caracterizam o perfil do usuário, além da falta de confiabilidade que os valores presentes oferecem. Adicionalmente, devido ao tamanho, o conjunto de dados de avaliações deverá passar por um processo de amostragem antes de, então, ser dividido entre conjunto de treino e de teste.

4.3 Preparação dos Dados

A etapa de Preparação de dados inicia com a responsabilidade de definir de forma mais criteriosa os conjuntos ou subconjuntos de dados que serão utilizados na etapa de Modelagem, além de realizar toda a limpeza, construção, integração e formatação dos dados para serem processados da forma correta. A seguir, a execução dessas tarefas será detalhada em cada seção.

4.3.1 Seleção

Um primeiro exercício de Preparação de dados é utilizar o conhecimento obtido tanto na etapa de Entendimento do negócio mas, principalmente, na etapa de Compreensão de dados, para tomar a decisão sobre quais atributos e conjuntos de exemplos serão utilizados.

Para o domínio deste trabalho, foram removidos exemplos no conjunto de animes que continham conteúdo pornográfico. Esse filtro foi realizado por meio do atributo de classificação etária indicativa, que assinalava animes desse tipo por meio da classificação "RX - Hentai".

Já no conjunto de avaliações, foram realizados outros processos de seleção. Vale lembrar que a base de avaliações é muito volumosa (mais de vinte milhões de exemplos), o que oferece desafios para a memória no que tange ao processamento dos modelos, elevando o custo computacional. Por esse motivo, foi realizada uma amostragem aleatória de 5.000.000 de exemplos iniciais.

Com a base de avaliações amostrada, outro filtro realizado foi em relação aos animes pouco avaliados. O conjunto de dados de animes, que é onde estão os metadados como sinopse e gênero, apenas contém as produções com 100 avaliações ou mais. Isso implica que aqueles com menos que isso no conjunto de avaliações ficariam sem metadados, o que prejudicaria a Modelagem e Avaliação de alguns modelos.

Adicionalmente, conforme visto anteriormente na etapa de Compreensão de dados, mais da metade dos usuários no conjunto de avaliações deram menos de 50 notas. Na prática, isso significa que ao ser construída uma matriz usuário-item com esses usuários incluídos, uma esparsidade muito maior seria enfrentada, o que pode ser prejudicial para a construção de modelos precisos. Portanto, para ter na base mais itens por usuário para o treinamento dos modelos, usuários que deram menos de 50 avaliações foram descartados.

O título do anime no idioma original não será utilizado uma vez que o id do anime é suficiente para identificação dos mesmos, mas a sua versão em inglês foi selecionada como característica de anime e alimentará alguns modelos. Por outro lado, metadados de anime como "Tipo"(que se refere ao fato de um anime ser lançado na TV ou em outro canal), "Nº de episódios", duração, nome das empresas produtoras, licenciadoras e estúdios não foram consideradas relevantes para o experimento por não serem relevantes neste experimento para caracterizar preferências de usuários.

O atributo "origem"(que denota se foi um anime original ou adaptado dos quadrinhos japoneses, conhecidos como "mangás"), o link de imagem de capa e atributos relacionados ao consumo do anime (posição em *rankings* de popularidade, nº de marcações como "favorito", nº de avaliações e adições à lista pessoal) também foram descartados por serem considerados de menor importância para os modelos que serão testados. Em contrapartida, conforme citado nos trabalhos relacionados, metadados como a imagem de capa já foram utilizados em outros estudos e podem gerar resultados interessantes. Esses atributos apenas não foram priorizados para este trabalho por uma questão de foco.

Com isso, ficou estabelecida a utilização dos seguintes conjuntos de dados e atributos-chave:

- **Conjunto de dados de avaliações:** id do usuário, id do anime e avaliação (*"rating"*).
- **Conjunto de dados de animes:** id do anime, nome do anime em inglês, nota média ponderada recebida (*"score"*), gênero, sinopse, ano de lançamento e classificação etária indicativa.

4.3.2 Limpeza

Na sub-etapa de limpeza, foram endereçados os dados faltantes identificados na etapa de Compreensão dos Dados. Entre os atributos selecionados, havia dados faltantes de nota média de animes (*"Score"*), gênero e classificação etária indicativa. Foram excluídos exemplos que se enquadrassem em qualquer uma dessas condições, resultando em uma diminuição de cerca de 15% na base total de animes.

4.3.3 Construção

Conforme levantado na etapa de Compreensão dos Dados, o atributo "Aired"que originalmente se refere ao período de exibição do anime continha diferentes formatos. Por exemplo, alguns se referiam a uma data específica, outros tinham um período que começava em uma data e terminava em outra e ainda havia aqueles que apenas exibiam um ano, resultando em pelo menos dez tipos de resposta. Tratamentos específicos foram aplicados para que esse atributo fosse utilizado como base para a construção de um novo, chamado "ano de lançamento", que será utilizado na etapa de Modelagem. Ainda assim, cerca de 1% dos animes não pôde ter seu ano de lançamento recuperado e esses casos também foram descartados.

4.3.4 Integração

A integração consiste na junção de conjuntos de dados com exemplos semelhantes, mas com diferentes atributos. Isso ocorreu nesta etapa ao trazer os metadados de animes para o conjunto de dados de avaliações. O "id do anime", presente em ambos os conjuntos, permitiu a junção utilizando esse atributo como chave e identificador único dos animes.

4.3.5 Formatação

Por fim, na sub-etapa de formatação, foram realizadas as conversões de formato dos dados necessárias aos modelos que serão aplicados. Primeiramente, foi realizada uma normalização das notas médias de animes ("*scores*") para uma variação de 0 a 1. A classificação etária indicativa recebeu a aplicação de um *label encoding* para que assuma valor numérico de 0 a 4, onde quanto maior o número, mais alta é a faixa etária mínima indicativa.

Os atributos textuais (sinopse e gênero) também tiveram tratamentos adequados para a correta formatação. As sinopses passaram por um processo de remoção de *stopwords*

e *lematização* antes de serem submetidas ao vetorizador TF-IDF, a partir do qual uma matriz TF-IDF foi criada. Os gêneros, por outro lado, foram tratados como variável categórica distribuída em colunas assumindo o valor de 0 (quando ausente no anime) e 1 (quando presente) por meio do *one-hot encoding*.

4.4 Modelagem

Nesta seção, são apresentados os modelos e configurações dos algoritmos utilizados para gerar as recomendações.

4.4.1 Popularidade (Nota Média Ponderada)

Conforme explicado anteriormente, a nota média dos animes é calculada por meio de um estimador Bayesiano que a própria MyAnimeList utiliza para obter notas ponderadas ('*scores*') levando em consideração o número de avaliações que uma obra teve. Como um modelo sem personalização, o sistema de recomendação baseado em popularidade utilizou essa medida como referência para prever qual nota um usuário qualquer u daria para um anime i.

Em outras palavras, a nota predita para cada usuário em relação a um anime não assistido é exatamente o seu *score*. Se um anime tem uma nota média ponderada alta, significa que ele é popular e portanto deve ser recomendado. Ou seja, a premissa básica é a de que um anime muito bem avaliado por uma quantidade significativa de pessoas provavelmente será recomendado para qualquer usuário nesse tipo de sistema.

Além disso, a característica não personalizável do sistema se evidencia no fato de que apenas o identificador do anime é suficiente para prever qual nota o usuário dará.

4.4.2 Collaborative filtering (*Memory-based*)

Primeiro, foi produzida uma matriz usuário-item, isto é, uma matriz onde cada linha representa um usuário e cada atributo representa um anime, restando aos valores as notas dadas pelos usuários aos animes. Uma vez que usuários tendem a avaliar uma quantidade menor do que o total de conteúdos disponíveis, a matriz gerada teve 4,34% de densidade.

A partir disso, na abordagem baseada em usuário, calcula-se a similaridade entre usuários a partir da nota dada para os mesmos animes. Enquanto na abordagem baseada em item, são calculadas as similaridades entre itens baseando-se nas notas dadas pelo mesmo usuário. Isso possibilita a criação de uma matriz de similaridade para cada abordagem, que é usada como referência para o próximo passo. Vale mencionar que todas as similaridades foram calculadas utilizando a Similaridade de Cosseno e levando em consideração os 50 usuários ou itens mais similares.

Em seguida, uma função executa as seguintes ações a cada linha da tabela de avaliações: i) identifica qual é o par usuário-item, ii) resgata os usuários ou itens mais similares com base nas matrizes de similaridade, iii) verifica quais foram as notas que os usuários mais similares deram para o mesmo item ou as notas que o mesmo usuário deu para itens mais similares, iv) preenche uma lista com as notas e outra lista com as similaridades (como pesos) e v) retorna um valor predito que é a soma das notas multiplicada pelos pesos dividida pela soma dos pesos.

4.4.3 Collaborative filtering (*Model-based*)

Na filtragem colaborativa baseada em modelo, foi utilizado *Singular Value Decomposition* (SVD), conforme popularizado por Simon Funk (FUNK, 2006). Primeiramente, um conjunto de dados é separado para o treino. O próximo passo foi inicializar o modelo SVD e treiná-lo com os dados de treino. Durante o treinamento, o SVD decompõe a matriz de avaliações em três matrizes menores (U, Σ, V) que capturam as características latentes dos usuários e dos itens (animes). Essas características latentes ajudam a identificar padrões e relações escondidas nos dados, mesmo que alguns valores estejam faltando.

- U: Matriz que representa a relação entre os usuários e as características latentes;
- Σ: Matriz diagonal que contém os valores singulares, que representam a importância das características latentes;
- V: Matriz que representa a relação entre os itens (animes) e as características latentes.

Na predição, ie., recomendação, o algoritmo SVD utiliza as matrizes resultantes em uma forma de menor dimensionalidade para estimar as notas que os usuários dariam aos animes que ainda não avaliaram, gerando a matriz de notas preditas $\hat{R}$:

$$\hat{R} = U_k \Sigma_k V_k^T$$

$\hat{R}$ é a matriz de avaliações previstas, onde cada elemento $\hat{R_i}j$ representa a nota prevista do usuário i para o anime j. Por fim, k representa o número de fatores latentes escolhidos para serem mantidos (por padrão, foram 100 neste experimento). Estes fatores latentes são as principais características latentes que capturam a maioria da variabilidade nos dados.

4.4.4 Content-based

A filtragem baseada em conteúdo é uma técnica que utiliza as características de itens já curtidos por um usuário para recomendar novos conteúdos que ele provavelmente vai gostar. O sistema identifica as características dos itens, como gênero e sinopse. Com

base nessas informações, busca por novos itens que possuem características semelhantes, oferecendo uma seleção personalizada de conteúdos com maior probabilidade de serem apreciados.

Para cada usuário, o sistema baseado em conteúdo funcionou a partir das seguintes ações: i) levantar quais animes o usuário assistiu, ii) entre os assistidos, descobrir qual os k animes mais similares (vizinhos mais próximos) ao anime-alvo para o qual deseja-se prever uma nota, iii) recuperar a nota dada para esses animes assistidos mais similares e, iv) calcular média das notas encontradas como a predição para o anime ainda não avaliado.

A premissa é que, por terem características mais similares, a chance de ambos os animes receberem notas similares do mesmo usuário é maior. Com isso, obtém-se notas preditas para todos os pares usuário-item. Para tanto, foram utilizados diferentes atributos de animes para calcular essas similaridades: nota média ponderada recebida ('*score*'), ano de lançamento, classificação etária indicativa, gênero e sinopse. Neste experimento, foram usados os 2 vizinhos mais próximos.

Assim como nos cálculos de similaridade que foram feitos em outros modelos, a métrica de distância usada para encontrar os vizinhos foi a distância de cosseno. Foram testadas duas variações desse modelo, modificando-se os atributos usados no treinamento com o fim de influenciar o cálculo de similaridade e encontrar melhor assertividade, a saber: um utilizando os atributos numéricos (nota média ponderada recebida ('*score*'), ano de lançamento, classificação etária indicativa) e outro utilizando os atributos derivados de texto (gênero e sinopse).

4.4.5 Ensemble

Como uma forma de experimentar a combinação de diferentes sistemas, foi explorada uma versão onde um recomendador baseado em conteúdo foi combinado com um baseado em item (de filtragem colaborativa), formando um sistema de recomendação híbrido do tipo *ensemble*. Nesse caso, o *output* do sistema é uma média simples entre as notas preditas geradas pelos dois recomendadores.

Em outras palavras, os recomendadores atuam em paralelo e geram resultados individuais para, posteriormente, essas três notas preditas serem combinadas em uma nota final calculada por meio de uma média simples.

O sistema pode ser considerado híbrido, pois combina diferentes recomendadores, já o termo *ensemble* (design de conjunto) designa um sistema cujos recomendadores funcionam de forma paralela, isto é, as previsões são combinadas apenas no final, gerando um único *output*: a nota que um usuário dará para um anime.

4.5 Avaliação

A amostra do conjunto de avaliações foi dividida em duas partes: uma para treino, representando 60% dos dados, e outra para teste, com os demais 40%. Todos os algoritmos foram avaliados com base em seus resultados no conjunto de teste, que reuniu avaliações de 3.175 usuários únicos a respeito de 564 animes diferentes.

A partir das diferentes abordagens apresentadas neste capítulo, espera-se encontrar qual solução prevê notas de usuários com menor erro. Essa avaliação será melhor detalhada no capítulo a seguir.

* * *

5 AVALIAÇÃO EXPERIMENTAL

Nesta seção são apresentados os resultados obtidos utilizando a metodologia presente no Capítulo 4. Utilizando diferentes abordagens, buscou-se obter previsões das notas que usuários dariam a animes. A partir disso, foi possível mensurar o erro entre a nota predita e a nota que o usuário realmente deu.

Na Tabela 4 são apresentados os resultados, os quais estão divididos por tipo de sistema, abordagem e tipos de atributos. Nas próximas seções, são apresentadas a análise geral e a análise por abordagem. Por fim, serão apresentadas as conclusões e uma análise considerando os algoritmos avaliados e os resultados obtidos.

Tipo	Abordagem	Atributos	RMSE	MAE
Collaborative	Model-based SVD	Ratings	1,27	0,96
Híbrido	Média do output	Gênero e Sinopse + Item-based	1,45	1,13
Collaborative	Item-based	Ratings	1,50	1,12
Content-based	k-NN (k=3)	Gênero e Sinopse	1,52	1,15
Popularidade	Sem personalização	Score	1,52	1,15
Collaborative	User-based	Ratings	1,54	1,17
Content-based	k-NN (k=3)	Class etária, ano e score	1,86	1,42

Tabela 4 – Comparação de Abordagens

5.1 Análise Geral

Pela Tabela 4, nota-se que o sistema de recomendação baseado em *Singular Value Decomposition* demonstrou os menores erros de predição, tanto em termos de RMSE quanto de MAE. Esses resultados estão consonantes com os trabalhos relacionados, em que sistemas deste tipo também tiveram as melhores performances preditivas.

Da mesma forma, o sistema híbrido também esteve entre os melhores resultados, uma vez que combinou abordagens promissoras em sistemas baseados em conteúdo e de filtragem colaborativa. Isso demonstra a potencialidade que sistemas híbridos têm, ao combinar vantagens de diferentes recomendadores.

É importante mencionar que o único sistema de recomendação sem qualquer personalização (o baseado em popularidade) ofereceu um patamar de erro muito competitivo com outras abordagens personalizadas mais sofisticadas, como a filtragem colaborativa baseada em usuário.

5.2 Análise por Abordagem

Ao comparar as abordagens, percebeu-se que a abordagem baseada em conteúdo pode ser promissora ao incorporar metadados de animes, especialmente o gênero e a sinopse. Em muitas ocasiões, usuários podem desejar um tipo específico de conteúdo, que é refletido nesses atributos e pode ser capturado pelo sistema de recomendação ao oferecer conteúdos semelhantes.

Ainda na abordagem baseada em conteúdo, atributos numéricos como a classificação etária indicativa, ano de lançamento e o score não ofereceram um nível de erro competitivo comparativamente às outras abordagens quando utilizados no k-NN. Isso demonstra que, para a recomendação de animes, o nível de maturidade exigido para assistir e a sua recência parecem não ser fatores cruciais para determinar preferências. O score, por outro lado, teve bastante utilidade ao ser inserido no sistema sem personalização, o que indica que, em geral, animes adorados pela massa merecem destaque entre as recomendações.

Além disso, nota-se que a abordagem baseada em item foi mais bem sucedida que a baseada em usuário. Quanto a isso, vale mencionar que, neste experimento, havia muito mais usuários do que itens e, neste tipo de cenário, a filtragem colaborativa baseada em item pode ser mais eficiente, uma vez que comparar todos os usuários entre si pode ser mais caro computacionalmente, além de demorado.

5.3 Conclusões gerais

Quando um novo anime é adicionado no catálogo de uma plataforma de streaming, ele ainda não possui avaliações dos usuários. Sendo assim, nenhum dos métodos de filtragem colaborativa o recomendaria. Na literatura, isso é referido como problema de inicialização a frio (*'cold-start problem'*).

Em contrapartida, um sistema baseado em conteúdo pode utilizar os atributos desse novo item para identificar seu grau de similaridade com os demais e gerar recomendações caso um dado usuário tenha apreciado itens semelhantes anteriormente. Esse exemplo mostra como abordagens baseadas em conteúdo podem ser importantes em um sistema, mesmo que algoritmos de filtragem colaborativa apresentem recomendações mais assertivas na maioria dos casos.

O problema de inicialização à frio também ocorre quando um novo usuário chega à plataforma. Nesse contexto, nenhum feedback foi fornecido ainda e, portanto, tanto abordagens baseadas em conteúdo quanto de filtragem colaborativa terão dificuldade de realizar recomendações.

Para lidar com isso, sistemas baseados em popularidade podem ser promissores. Afinal, os resultados mostram que é improvável que itens muito populares recebam notas muito distintas da sua média ponderada. Isto é, se é um item que agrada a muitos, tem

uma chance relativamente baixa de desagradar um novo usuário.

6 CONCLUSÕES

Animes têm se tornado cada vez mais populares entre os mais diferenciados públicos, fortalecendo a intenção de plataformas de streaming de expandir o catálogo com opções de conteúdo de forma veloz. No entanto, simplesmente aumentar as opções disponíveis não garante uma boa experiência de usuário, implicando no risco de ocorrer até mesmo o oposto, devido à sobrecarga cognitiva da tomada de decisão. Por esse motivo, sistemas de recomendação têm tido sua importância cada vez mais ampliada, uma vez que possibilitam que cada indivíduo tenha uma experiência personalizada e adequada às suas preferências individuais.

Este trabalho de conclusão de curso propôs uma comparação entre diferentes sistemas de recomendação aplicados ao contexto de animes. Por meio de diferentes algoritmos, foram produzidas mais de 500 mil predições de nota com o intuito de mensurar o erro entre as notas preditas e reais e diagnosticar qual abordagem oferece maior assertividade.

Com base nos experimentos realizados, os resultados demonstraram a superioridade do método *Singular Value Decomposition* em comparação com as demais abordagens, o que indica que sistemas de recomendação de filtragem colaborativa tendem a ter maior assertividade. Contudo, devido aos desafios que cada algoritmo enfrenta, como o problema de inicialização em frio (*'cold-start problem'*, não há uma abordagem em específico que se sobreponha às demais em todas as ocasiões, de forma que uma combinação de diferentes recomendadores levando em consideração a etapa da jornada do usuário pode apresentar resultados promissores, o que inclui até mesmo recomendadores sem personalização (baseados em popularidade).

Em outras palavras, pode-se concluir que sistemas de recomendação para plataformas de streaming com animes podem se beneficiar muito de combinações de abordagens. Isso potencializará uma boa experiência de usuário e ajudará a lidar com desafios que os próprios algoritmos enfrentam, especialmente se houver espaço para cada recomendador ser mais protagonista a depender da etapa da jornada do usuário. Usuários recentes podem experimentar uma interface própria para eles, enquanto usuários antigos terão a chance de encontrar recomendações com muita personalização. Com isso, espera-se maximizar sua satisfação.

Ainda assim, este estudo não está isento de limitações. O conjunto de dados de avaliações original não pôde ser utilizado por uma limitação no uso de memória. Além disso, o período de cobertura dos dados pode não abranger todas as nuances que ocorrem em avaliações e no comportamento dos usuários em relação aos animes.

Como implicação prática, é importante destacar que os recomendadores desenvol-

vidos têm potencial para serem aplicados em plataformas que armazenam e distribuem digitalmente animes e outros conteúdos correlatos, como animações, séries, filmes e documentários (plataformas de streaming de conteúdo).

Recomenda-se que trabalhos futuros explorem metadados de usuários, como idade, região de moradia, gênero, raça, entre outros. Além disso, há espaço para a avaliação de muitas outras abordagens, fazendo uso especialmente de aprendizado profundo *(deep learning*) e outras abordagens de *ensemble*. Adicionalmente, pode ser interessante considerar outras formas de feedback além das notas, isto é, tanto feedbacks explícitos de "gostei", "amei"e "não é para mim", quanto implícitos como tempo de visualização.

Finalmente, esta pesquisa forneceu *insights* valiosos e um protocolo robusto para estimular novas pesquisas em sistemas de recomendação, estabelecendo uma base sólida para explorar as inúmeras possibilidades deste campo.

* * *

REFERÊNCIAS

AGGARWAL, C. C. An introduction to data classification. **Data classification: algorithms and applications**, v. 125, n. 3, p. 142, 2014.

AGGARWAL, C. C. *et al.* **Data mining: the textbook**. [*S.l.: s.n.*]: Springer, 2015. v. 1.

AGGARWAL, C. C. *et al.* **Recommender systems**. [*S.l.: s.n.*]: Springer, 2016. v. 1.

AHUJA, R.; SOLANKI, A.; NAYYAR, A. Movie recommender system using k-means clustering and k-nearest neighbor. *In*: IEEE. **2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)**. [*S.l.: s.n.*], 2019. p. 263–268.

AJITSARIA, A. **Build a recommendation engine with collaborative filtering**. 2022. Available at: <https://realpython.com/build-recommendation-engine-collaborative-filtering>.

ARYOSETO, G. S.; MARDIANTO, I.; ARIWIBOWO, A. B. Recommendation system for mental health article on circle application. **Intelmatics**, v. 3, n. 1, p. 1–6, 2023.

BADDELEY, M. Herding, social influence and economic decision-making: socio-psychological and neuroscientific analyses. **Philosophical Transactions of the Royal Society B: Biological Sciences**, The Royal Society, v. 365, n. 1538, p. 281–290, 2010.

BURKE, R. Hybrid systems for personalized recommendations. *In*: SPRINGER. **IJCAI Workshop on Intelligent Techniques for Web Personalization**. [*S.l.: s.n.*], 2003. p. 133–152.

CARBONE, F. **Crunchyroll ultrapassa os 13 milhões de assinantes <https://www.adrenaline.com.br/games/crunchyroll-ultrapassa-13-milhoes-assinantes/>**. 2024. Último acesso em 15 de Janeiro de 2024.

CASTRO, L. N. D.; FERRARI, D. G. **Introdução à mineração de dados**. [*S.l.: s.n.*]: Saraiva Educação SA, 2017.

CHAPMAN, P. *et al.* Crisp-dm 1.0: Step-by-step data mining guide. **SPSS inc**, v. 9, n. 13, p. 1–73, 2000.

CHATTERJEE, T. K. **Why using CRISP-DM will make you a better Data Scientist? <https://www.mygreatlearning.com/blog/why-using-crisp-dm-will-make-you-a-better-data-scientist/>**. 2022. Último acesso em 10 de Fevereiro de 2024.

CNI, C. N. D. I. Perfil do consumidor: Consumo pela internet. retratos da sociedade brasileira, 9(51). 2022.

COLES, S.; JR, P. J. R. Inferência estatística. 2023.

DALL'ARA, J. **Alto consumo de pirataria é favorecido pela desigualdade econômica no País <https://jornal.usp.br/atualidades/desigualdade-economica-e-um-dos-fatores-responsaveis-pelo-alto-consumo-de-pirataria-no-pai>**. 2022. Último acesso em 15 de Janeiro de 2024.

DIELEMAN, S. Keynote: Deep learning for audio-based music recommendation. *In*: . [*S.l.: s.n.*], 2016. p. 1–1. ISBN 978-1-4503-4795-2.

Estadão Conteúdo. **Netflix (NFLX34) tem alta no lucro e no número de assinantes no 2º trimestre, mas receita decepciona <https://www.infomoney.com.br/mercados/netflix-nflx34-acoes-alta-no-lucro-e-no-numero-de-assinantes-no-2o-trimestre-mas-receita-dec>**. 2023. Último acesso em 15 de Janeiro de 2024.

Felipe Vinha. **Sony finaliza compra da Crunchyroll e deve unir streaming com Funimation <https://tecnoblog.net/noticias/2021/08/10/sony-finaliza-compra-da-crunchyroll-e-deve-unir-streaming-com-funimation/>**. 2021. Último acesso em 15 de Janeiro de 2024.

Folha de São Paulo. **Netflix atinge 118,9 milhões de assinantes <https://www1.folha.uol.com.br/mercado/2018/04/netflix-atinge-1189-milhoes-de-assinantes.shtml>**. 2018. Último acesso em 15 de Janeiro de 2024.

FUNK, S. **Netflix Update: Try This at Home**. 2006. Online. Acessado: 13 de junho de 2024. Available at: <https://sifter.org/~simon/journal/20061211.html>.

HODSON, T. O. Root mean square error (rmse) or mean absolute error (mae): When to use them or not. **Geoscientific Model Development Discussions**, Göttingen, Germany, v. 2022, p. 1–10, 2022.

HOTZ, N. **What is CRISP DM? <https://www.datascience-pm.com/crisp-dm-2/>**. 2023. Último acesso em 10 de Fevereiro de 2024.

IBM. **IBM SPSS Modeler CRISP-DM Guide <https://www.ibm.com/docs/en/spss-modeler/18.1.1?topic=spss-modeler-crisp-dm-guide>**. 2021. Último acesso em 21 de Março de 2024.

IBM. **What is the k-nearest neighbors (KNN) algorithm? <https://www.ibm.com/topics/knn>**. 2022.

IMDB, H. c. **Ratings FAQ**. IMDb.com, 2023. Available at: <https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV#>.

JENA, A. *et al.* Recommendation system for anime using machine learning algorithms. *In*: **Proceedings of the International Conference on Innovative Computing & Communication (ICICC)**. [*S.l.: s.n.*], 2022.

JI, Y. *et al.* A re-visit of the popularity baseline in recommender systems. *In*: **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval**. [*S.l.: s.n.*], 2020. p. 1749–1752.

JúNIOR DANIEL MATTOS, l. P. L. F. L. T. R. W. F. **Panorama do Mercado de Vídeo por Demanda no Brasil <https://www.gov.br/ancine/pt-br/oca/publicacoes/arquivos.pdf/informe-vod2022.pdf>**. 2022. Último acesso em 10 de Fevereiro de 2024.

KAHNEMAN, D.; TVERSKY, A. Prospect theory: An analysis of decision under risk. **Econometrica**, [Wiley, Econometric Society], v. 47, n. 2, p. 263–291, 1979. ISSN 00129682, 14680262. Available at: <http://www.jstor.org/stable/1914185>.

KNAUER, L. **OTT CHURN: EVERYTHING YOU NEED TO KNOW <https://www.brightcove.com/en/resources/blog/ott-churn-everything-you-need-know/>**. 2019. Último acesso em 15 de Janeiro de 2024.

KOREN, Y.; BELL, R.; VOLINSKY, C. Matrix factorization techniques for recommender systems. **Computer**, IEEE, v. 42, n. 8, p. 30–37, 2009.

Ligia Mello. **Streaming 2023 <https://cdnstar.b-cdn.net/wp-content/uploads/2023/08/23HB_STR001.pdf>**. 2023. Último acesso em 15 de Janeiro de 2024.

MADHUKAR, M. Challenges & limitation in recommender systems. **International Journal of Latest Trends in Engineering and Technology (IJLTET)**, v. 4, n. 3, p. 138–142, 2014.

MASE, H.; OHWADA, H. A collaborative filtering incorporating hybrid-clustering technology. *In*: IEEE. **2012 International Conference on Systems and Informatics (ICSAI2012)**. [*S.l.: s.n.*], 2012. p. 2342–2346.

MUTTEPPAGOL, V. M. **A deep learning recommender system for anime**. 2021. Tese (Doutorado) — Dublin, National College of Ireland, 2021.

NEAMAH, A. A.; EL-AMEER, A. S. Design and evaluation of a course recommender system using content-based approach. *In*: IEEE. **2018 International Conference on Advanced Science and Engineering (ICOASE)**. [*S.l.: s.n.*], 2018. p. 1–6.

NUURSHADIEQ; WIBOWO, A. T. Leveraging side information to anime recommender system using deep learning. *In*: **2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)**. [*S.l.: s.n.*], 2020. p. 62–67.

Patrick Macias. **Crunchyroll Brings Anime to Brazil! <https://www.crunchyroll.com/news/latest/2012/11/1/crunchyroll-brings-anime-to-brazil3>**. 2021. Último acesso em 15 de Janeiro de 2024.

Paula Filizola. **Mercado de plataformas de streaming valerá US**$1 trilhão em 2027.<>. 2021. Último acesso em 15 de Janeiro de 2024.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. Data quality assessment. **Communications of the ACM**, ACM New York, NY, USA, v. 45, n. 4, p. 211–218, 2002.

PODER360. **Brasil é o 5º país que mais consome pirataria, diz pesquisa <https://www.poder360.com.br/midia/brasil-e-o-5o-pais-que-mais-consome-pirataria-diz-pesquisa/>**. 2022. Último acesso em 15 de Janeiro de 2024.

PRAMANA, R. *et al.* Systematic literature review of stemming and lemmatization performance for sentence similarity. *In*: IEEE. **2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)**. [*S.l.: s.n.*], 2022. p. 1–6.

PYLE, D. **Data preparation for data mining**. [*S.l.: s.n.*]: morgan kaufmann, 1999.

QADER, W. A.; AMEEN, M. M.; AHMED, B. I. An overview of bag of words; importance, implementation, applications, and challenges. *In*: IEEE. **2019 international engineering conference (IEC)**. [*S.l.: s.n.*], 2019. p. 200–204.

RAMOS, J. *et al.* Using tf-idf to determine word relevance in document queries. *In*: CITESEER. **Proceedings of the first instructional conference on machine learning**. [*S.l.: s.n.*], 2003. v. 242, n. 1, p. 29–48.

ROSSI, R. G. **Classificaçao automática de textos por meio de aprendizado de máquina baseado em redes**. 2016. Tese (Doutorado) — Universidade de São Paulo, 2016.

SARWAR, B. *et al.* Item-based collaborative filtering recommendation algorithms. *In*: **Proceedings of the 10th international conference on World Wide Web**. [*S.l.: s.n.*], 2001. p. 285–295.

SCHÜTZE, H.; MANNING, C. D.; RAGHAVAN, P. **Introduction to information retrieval**. [*S.l.: s.n.*]: Cambridge University Press Cambridge, 2008. v. 39.

SCHWARTZ, B. **Barry Schwartz sobre o paradoxo da escolha <https://www.ted.com/talks/barry_schwartz_the_paradox_of_choice/transcript?language=pt-BR&subtitle=pt-br>**. 2005. Último acesso em 15 de Janeiro de 2024.

SILVA, L. M. Algoritmos de recomendação: estudo aplicado a streaming de anime. 2021.

TUKEY, J. W. *et al.* **Exploratory data analysis**. [*S.l.: s.n.*]: Reading, MA, 1977. v. 2.

TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. **Journal of artificial intelligence research**, v. 37, p. 141–188, 2010.

WHOIS Lookup MyAnimeList Check Domain <https://whois.domaintools.com/myanimelist.net>. 2024. Último acesso em 31 de Maio de 2024.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. *In*: MANCHESTER. **Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining**. [*S.l.: s.n.*], 2000. v. 1, p. 29–39.

YADAV, D. **Categorical encoding using Label-Encoding and One-Hot-Encoder**. 2019. Available at: <https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-o

YAO, S. T. S. *et al.* Recommender algorithm for japanese animes. 2021.