Adriano Jiang - 12543054

João Eduardo da Paz Silva - 11845514

Gustavo Almeida - 12543892

Lizandro Raposo - 12681351

Otávio Nunes Rosa - 11319037

**Definição de uma questão de pesquisa geral**

Recomendação de Filmes para Usuários Cadastrados: Um Estudo Personalizado

**Definição e coleta dos dados necessários**

Para este projeto, a pesquisa visa criar um sistema de recomendação de filmes personalizado para usuários cadastrados, utilizando a base de dados "The Movies Dataset". Essa base de dados é amplamente utilizada e compreende metadados de mais de 45.000 filmes, incluindo informações como elenco, equipe, palavras-chave do enredo, orçamento, receita, datas de lançamento, idiomas, produtoras e países. Além disso, o conjunto de dados contém 26 milhões de avaliações de aproximadamente 270.000 usuários, com classificações variando de 1 a 5, obtidas do site oficial do GroupLens.  
 Os dados necessários para este projeto foram obtidos do Kaggle, especificamente do conjunto de dados "The Movies Dataset" disponível em<https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/data?select=movies_metadata.csv>. Após adquirirmos o conjunto de dados, selecionamos uma amostra representativa para este planejamento. Esta amostra incluiu 100.000 avaliações de 700 usuários em relação a 9.000 filmes. Os dados foram extraídos diretamente do Kaggle, onde encontramos um subconjunto de dados específicos focados na exploração dos dados. Esta amostra foi obtida através de uma seleção aleatória, garantindo uma representação diversificada dos filmes e das avaliações dos usuários.

**Exploração dos dados**

Os dados estão organizados em quatro arquivos CSV, cada um com um propósito específico. Vou detalhar cada um deles abaixo.

Movies\_metadata

Este arquivo contém informações principais sobre os filmes, incluindo detalhes como orçamento, receita, data de lançamento, idioma original, país de produção e empresas envolvidas na produção.

* *adult (texto, Qualitativo nominal):* indicando se o filme é para maiores de idade.
* *belongs\_to\_collection (texto, dado estruturado):* nome da coleção à qual o filme pertence.
  + *id: texto, qualitativo nominal*
  + *name: texto, não estruturado*
  + *poster\_path: texto, não estruturado,*
  + *backdrop\_path: texto, não estruturado*
* *budget (real, Quantitativo contínuo):* orçamento do filme.
* *genres (texto, dado estruturado):* gêneros do filme.
  + *id: texto, qualitativo nominal*
  + *name: texto, Qualitativo nominal*
* *homepage (texto, não estruturado):* URL da página inicial do filme.
* *id (texto, Qualitativo nominal):* identificador único do filme.
* *imdb\_id (texto, Qualitativo nominal):* identificador único do filme no IMDb.
* *original\_language (texto, Qualitativo nominal):* idioma original do filme.
* *original\_title (texto, não estruturado):* título original do filme.
* *overview (texto, não estruturado):* resumo do enredo do filme.
* *popularity (real, Quantitativo discreto):* popularidade do filme.
* *poster\_path (texto, não estruturado):* caminho para o pôster do filme.
* *production\_companies (texto, dado estruturado):* empresas envolvidas na produção do filme.
  + *id: texto, Qualitativo nominal*
  + *name: texto, não estruturado*
* *production\_countries (texto, dado estruturado):* países de produção do filme.
  + *iso\_3166\_1: texto, Qualitativo nominal*
  + *name: texto, dado estruturado*
* *release\_date (texto, Qualitativo ordinal):* data de lançamento do filme formato AAAA-MM-DD
* *revenue (real, Quantitativo contínuo):* receita do filme.
* *runtime (real, Quantitativo contínuo):* duração do filme.
* *spoken\_languages (texto, dado estruturado):* idiomas falados no filme.
  + *iso\_3166\_1: texto, Qualitativo nominal*
  + *name: texto, dado estruturado*
* *status (texto,Qualitativo nominal):* estado de produção do filme.
* *tagline (texto, não estruturado):* slogan do filme.
* *title (texto, não estruturado):* título do filme.
* *video (texto, Qualitativo nominal):* indicando se há um vídeo promocional do filme disponível.
* *vote\_average (real, Quantitativo contínuo):* média das avaliações do filme.
* *vote\_count (real, Quantitativo discreto):* número de votos recebidos pelo filme.

Keywords

Este arquivo contém palavras-chave do enredo para os filmes.

* *id (texto, Qualitativo nominal):* identificador único do filme.
* *keywords (texto, Dado estruturado):* palavras-chave do enredo do filme.
  + *id: texto, Qualitativo nominal*
  + *name: texto, dado estruturado*

Credits

Este arquivo contém informações sobre o elenco e a equipe de produção de todos os filmes.

* *cast (texto, Qualitativo nominal):* informações sobre o elenco do filme.
* *crew (texto, Qualitativo nominal):* informações sobre a equipe de produção do filme.
* *id (texto, Qualitativo nominal):* identificador único do filme.

Links

Este arquivo contém os IDs TMDB e IMDb de todos os filmes.

* *movieId (texto, Qualitativo nominal):* identificador único do filme no conjunto de dados.
* *imdbId (texto, Qualitativo nominal):* identificador único do filme no IMDb.
* *tmdbId (texto, Qualitativo nominal):* identificador único do filme no TMDB.

Ratings

Este arquivo contém informações sobre as avaliações dos usuários sobre os filmes que ele assistiu.

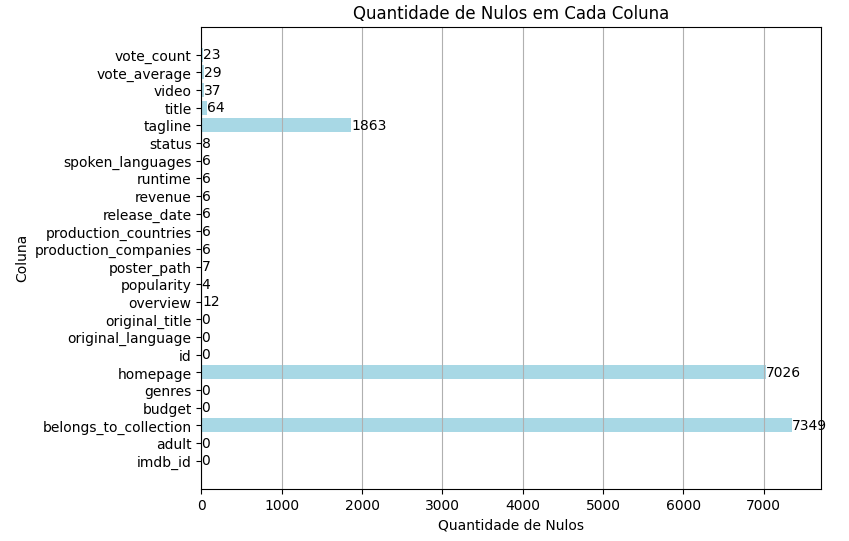
* *userId (texto, Qualitativo nominal):* identificador único do usuário no conjunto de dados.
* *movieId (texto, Qualitativo nominal):* identificador único do filme no conjunto de dados.
* *rating (real, Quantitativo discreto):* Nota de 0 a 5 que o usuário deu ao filme.
* *timestamp* (*real, Quantitativo discreto*): momento no qual o usuário avalia o filme. O timestamp é do tipo Unix, que é a contagem de segundos desde a meia-noite (UTC) de 1º de janeiro de 1970.

Devido à diversidade de arquivos disponíveis, é necessário realizar alguns joins para enriquecer nossas informações, especialmente quando se trata do elenco e da equipe de produção.

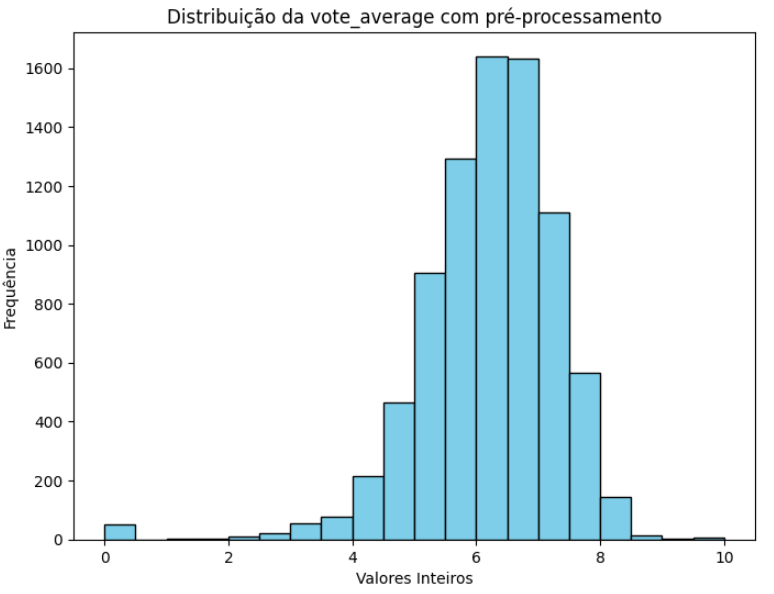
Inicialmente, identificamos várias informações que podem ser relevantes para a elaboração do ranking. No entanto, pode ser desafiador utilizar informações não estruturadas de texto. Neste momento, é prudente concentrar nossa análise nas colunas que contêm informações quantitativas.

A exploração subsequente dos dados, por meio de distribuições, gráficos, entre outros métodos, será conduzida nesta mesma amostra, a fim de evitar comprometer análises futuras que possam depender de um volume maior de dados, como a aplicação de algoritmos de aprendizado de máquina. Para essas análises mais avançadas, faremos uso do restante dos dados disponíveis.

Ao examinar nossos arquivos, não encontramos valores nulos nas tabelas de classificações, palavras-chave e créditos. No entanto, notamos que a tabela de links apresenta um pequeno número de valores nulos no identificador tmdbId, apenas 13, mas pode ser um complicador para linkar algumas palavras-chave e créditos de alguns filmes. Já tabela filmes possui algumas colunas com um número considerável de valores nulos, como é evidenciado no gráfico abaixo. Portanto, será necessário pensar em abordagens inteligentes para lidar com esses valores nulos.



É importante observar que, embora existam apenas alguns valores nulos, colunas como *vote\_average* e *vote\_count* também apresentam registros ausentes. Aponto especificamente essas colunas, pois são fortes candidatas a features para nosso modelo.

Analisamos uma das colunas para avaliar quais tipos de pré-processamento ainda seriam necessários antes de utilizá-las. Especificamente, examinamos a coluna 'vote\_average'. Após aplicarmos mais alguns tratamentos nos dados, observamos que, dentre as 9 mil amostras de filmes, 782 instâncias foram filtradas devido a critérios como remoção de valores nulos, textuais, valores fora do intervalo de 0 a 10 ou provenientes de alguma operação de junção de dados. O gráfico a seguir ilustra a distribuição da coluna após a aplicação desses filtros.  


Após estas análises, estamos mais conscientes do que iremos enfrentar neste dataset.

**Refinamento da questão de pesquisa**

A partir de uma análise gráfica e estatística da amostra de dados serão levantadas questões mais específicas, a serem verificadas no conjunto de dados não vistos. Nesse sentido, algumas possíveis questões passíveis de serem exploradas são:

* O que levou o usuário a dar uma nota mais baixa ou mais alta para um filme?
* Quais são os filmes mais bem avaliados? Quais são os filmes mais avaliados? Quais são os filmes mais populares?
* Quais são os assuntos que o usuário mais se interessa? Entre gênero do filme, palavras-chave, elenco, época de lançamento, o título, se é filme adulto, etc.
* o tempo de duração ou o orçamento do filme pode influenciar na nota do usuário?
* Qual o número mínimo de avaliações para entendermos o interesse do usuário? Qual o número mínimo de avaliações para observarmos as tendências de nota do usuário?
* Fatores menores podem influenciar no ranking? Como status do filme, se ele tem slogan, se tem overview, se tem homepage, se existe vídeo promocional do filme, etc.

Entendemos que essas são algumas das possíveis abordagens que poderemos atacar para satisfazer um ranking para o usuário. É possível que haja mais, o que pode ser levado em consideração no processo de geração de features para o modelo, mas já vemos um um leque de possibilidades para melhorar um ranking de filmes para um usuário. É claro que não iremos abordar todas, talvez começar um modelo simples com 3 features seja um começo interessante.

Os modelos de classificação normalmente funcionam prevendo uma score de relevância **s = f ( x )** para cada entrada **x = ( q, d )** onde **q** é uma consulta e **d** é um documento. Assim que tivermos o score de cada documento, podemos classificar os documentos de acordo com essas pontuações.



Para avaliar nossos modelos de classificação, precisamos definir as métricas apropriadas. Considerando que o rótulo é a nota dada pelo usuário a um filme, duas opções que surgem são o Mean Average Precision (MAP) e o Discounted Cumulative Gain (DCG). Uma abordagem inicial seria testar ambas as métricas para determinar qual se adequa melhor ao nosso contexto. Nosso objetivo é desenvolver um modelo simples para o problema, possivelmente uma abordagem pointwise utilizando técnicas de regressão, mas não definimos por enquanto.

**Planejamento em semanas**

**31/04 -** Definição de modelos pointwise a serem abordados e métricas de avaliação de classificação.

**07/05 -** Desenvolvimento de features método 80/20 para implementação e teste rápido.

**14/05 -** Insights e refinamento das features.

**21/05 -** Comparação de diferentes modelos.

**28/05 -** Documentação para entrega do trabalho.

**Referencia**: <https://towardsdatascience.com/learning-to-rank-a-complete-guide-to-ranking-using-machine-learning-4c9688d370d4#:~:text=Pointwise%20and%20pairwise%20approaches%20transform,by%20maximizing%20the%20evaluation%20metric>

**Features**:

* Genero (Bags of words) otavio
* Titulo do filme (tf–idf)
* Idioma (Bags of words)
* Data de lançamento
* Creditos (tf–idf)
* Palavras chave (tf–idf) otavio
* Media do filme

Bags of words and tf–idf: <https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html>

[Link da pasta](https://drive.google.com/drive/folders/14cp0iUooKxzUQOgGtIy4Wmc2lTne5zlO?usp=drive_link)