**Universidade Presbiteriana Mackenzie**

**Tecnologia em Ciência de Dados - Projeto Aplicado 3**

Matheus Vinícius Gomes

Leandro Rodrigues Dos Santos

Luiz Eduardo de Mei Salvador Coelho

Matheus Neves de Castro

# Projeto Aplicado III

# Sistema de Recomendação de Filmes para Plataforma de Streaming

SÃO PAULO

2º Semestre de 2024

Autores:
Matheus Vinícius Gomes / 10408179 / mathvgomes@gmail.com
Leandro Rodrigues Dos Santos / 23019689 / lers.138@gmail.com
Luiz Eduardo de Mei Salvador Coelho / 23024585 / Luiz02coelho@gmail.com
Matheus Neves de Castro / 10415190 / 10415190@mackenzista.com.br

# Projeto Aplicado III

# Sistema de Recomendação de Filmes para Plataforma de Streaming

Professora: CAROLINA TOLEDO FERRAZ

SÃO PAULO

2º Semestre de 202

**Resumo**

O crescimento das plataformas de streaming trouxe à tona o paradoxo da escolha, no qual a ampla oferta de conteúdos dificulta que os usuários encontrem títulos alinhados às suas preferências. Nesse cenário, os sistemas de recomendação desempenham um papel essencial, personalizando sugestões e promovendo uma experiência de uso mais satisfatória e eficiente. Este projeto visa desenvolver um sistema de recomendação de filmes que combine precisão e relevância, utilizando técnicas avançadas de Machine Learning.

O trabalho foi estruturado em cinco etapas principais: a análise exploratória dos dados, que fornece uma visão geral dos padrões de consumo dos usuários; a etapa de preparação e limpeza dos dados, garantindo uma base confiável e consistente para os modelos; a seleção e aplicação de algoritmos de aprendizado de máquina, que dão suporte à criação de recomendações precisas; a avaliação do desempenho do sistema, utilizando métricas específicas para medir a eficácia das recomendações; e, finalmente, a documentação dos resultados, com foco na identificação de oportunidades de melhorias e no impacto do sistema na experiência dos usuários.

Esse projeto busca não apenas otimizar o processo de recomendação de filmes, mas também contribuir para uma experiência de uso mais personalizada e agradável, estimulando a exploração de novos conteúdos e promovendo maior engajamento e retenção nas plataformas de streaming. Além disso, este trabalho visa ampliar o conhecimento sobre sistemas híbridos de recomendação, abordando limitações e identificando oportunidades de melhoria para o desenvolvimento de soluções mais inclusivas e adaptáveis.

**Sumário**

# Introdução

Com o crescimento vertiginoso das plataformas de streaming, a experiência do usuário se torna cada vez mais desafiadora à medida que a biblioteca de conteúdo se expande continuamente. Os usuários enfrentam uma diversidade de gêneros, estilos e diretores, que, embora enriquecedora, pode levar à **sobrecarga de opções**. Esse fenômeno, conhecido como **paradoxo da escolha**, muitas vezes dificulta que os usuários encontrem conteúdos que realmente correspondam aos seus gostos e preferências. Nesse contexto, os **sistemas de recomendação de filmes** surgem como uma solução fundamental, direcionando o usuário de maneira prática e personalizada para os títulos que mais se adequam ao seu perfil.

Um sistema de recomendação bem estruturado pode transformar a experiência do usuário, facilitando a descoberta de novos conteúdos e promovendo o engajamento. Ao fornecer sugestões baseadas em preferências e comportamentos anteriores, esses sistemas buscam não apenas atender às expectativas imediatas do usuário, mas também encorajá-los a explorar filmes que poderiam, de outra forma, passar despercebidos. Em um mercado altamente competitivo, onde a retenção de usuários é essencial para o sucesso, esses sistemas se tornam um diferencial estratégico para as plataformas de streaming, **aumentando tanto a satisfação quanto a lealdade dos usuários.**

Este projeto, portanto, visa desenvolver um sistema de recomendação de filmes que ofereça sugestões personalizadas com base no comportamento e nas preferências dos usuários. A implementação de um sistema desse tipo requer a utilização de técnicas de **Machine Learning** e a análise profunda dos dados de consumo, para garantir que as recomendações sejam precisas e relevantes.
Para alcançar esse objetivo, estabelecemos os seguintes objetivos específicos:

1. **Analisar e explorar dados de consumo de filmes**: Examinar dados de interações dos usuários com o catálogo de filmes, realizando uma análise exploratória dos padrões de consumo.
2. **Preparar e limpar os dados**: Implementar técnicas de pré-processamento e limpeza dos dados, assegurando uma base confiável e estruturada para a modelagem.
3. **Selecionar e aplicar técnicas de Machine Learning**: Escolher algoritmos adequados para a criação de um modelo de recomendação eficaz e personalizado.

4. **Avaliar o desempenho do modelo**: Definir e aplicar métricas para medir a precisão e a relevância das recomendações, refinando o modelo conforme necessário.
5. **Documentar e apresentar os resultados**: Registrar todas as etapas e os resultados do projeto, ressaltando os impactos positivos na experiência do usuário e as possibilidades de melhorias futuras.

Dessa forma, o presente trabalho busca não apenas otimizar o processo de recomendação de filmes, mas também contribuir para uma experiência de uso mais agradável, estimulando a **exploração contínua de novos conteúdos** e ampliando o valor percebido pelo usuário nas plataformas de streaming.

## Referencial Teórico

O aumento no uso de plataformas de streaming trouxe à tona a importância de sistemas de recomendação personalizados, que não apenas facilitam a escolha de conteúdo, mas também aprimoram a experiência do usuário ao sugerir filmes que correspondam aos seus interesses. Diversos estudos científicos têm investigado formas eficazes de desenvolver e aprimorar esses sistemas. Nesta seção, abordaremos os principais trabalhos relacionados ao tema e os desafios que ainda existem na literatura sobre recomendações de filmes.

### Abordagens em Sistemas de Recomendação

Os sistemas de recomendação de filmes têm sido amplamente estudados, com diferentes abordagens propostas para otimizar a personalização e precisão. Filtragem colaborativa, baseada em interações dos usuários, é uma técnica comum em sistemas de recomendação (Su et al., 2020), utilizando métodos como decomposição em valores singulares (SVD) e Redes Neurais, que correlacionam usuários com preferências similares. Embora amplamente adotada, a filtragem colaborativa sofre de problemas de sparsity (sparsidade), onde a ausência de avaliações suficientes para certos filmes ou novos usuários limita a qualidade das recomendações (Schafer et al., 2007).

A filtragem baseada em conteúdo é outra técnica frequentemente utilizada, que recomenda filmes com características semelhantes aos que o usuário já avaliou positivamente (Pazzani & Billsus, 2007). No entanto, este método enfrenta o desafio da dependência de dados: é necessário que as características do filme (gênero, diretor, elenco) estejam bem definidas, o que nem sempre é possível ou suficiente para capturar nuances de preferências individuais.

Recentemente, os modelos híbridos têm ganhado destaque ao combinar filtragem colaborativa e baseada em conteúdo para superar as limitações de ambas as abordagens. Burke (2002) demonstrou que esses modelos conseguem reduzir a sparsidade e melhorar a precisão da recomendação. Contudo, desafios como a complexidade computacional e a necessidade de calibragem entre diferentes métodos ainda persistem.

**Desafios no Desenvolvimento de Sistemas de Recomendação de Filmes**

Apesar do progresso nas abordagens de recomendação, alguns desafios críticos ainda permanecem. Um dos principais desafios é o problema do cold start para novos usuários e filmes, que ocorre quando o sistema não possui informações suficientes para fazer recomendações iniciais eficazes (Park et al., 2021). Várias técnicas têm sido exploradas para mitigar esse problema, como o uso de aprendizado de transferência e dados externos (Camacho & Alves-Souza, 2018), mas ainda não há uma solução universalmente eficaz.

Outro desafio significativo é a explicabilidade das recomendações. Como muitos sistemas de recomendação de filmes utilizam algoritmos complexos, os resultados podem parecer uma "caixa-preta" para os usuários. Trabalhos de Tintarev e Masthoff (2011) exploram formas de tornar os sistemas de recomendação mais transparentes e justificáveis, porém, em alguns casos, a explicação pode prejudicar a simplicidade e a eficiência do modelo.

Além disso, há questões de equidade e diversidade nas recomendações. Estudos mostram que recomendações tendem a reforçar padrões existentes, recomendando frequentemente filmes populares e ignorando opções menos conhecidas (Jannach et al., 2015). Modelos que priorizam filmes amplamente avaliados tendem a obscurecer conteúdos de nicho, o que representa um obstáculo tanto para a diversidade das recomendações quanto para a satisfação do usuário.

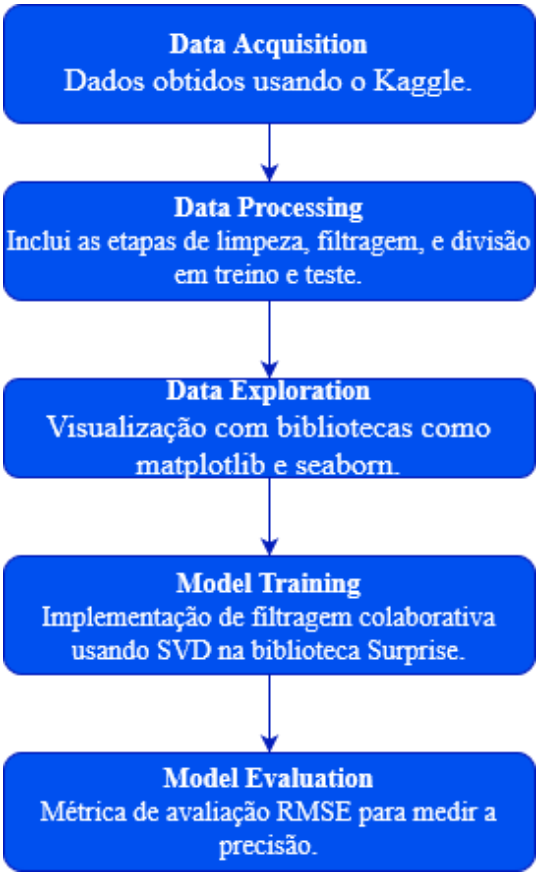
**Perspectivas Futuras**

O campo dos sistemas de recomendação de filmes segue evoluindo, especialmente com o uso de modelos baseados em Deep Learning e aprendizado de representação. Esses modelos conseguem capturar padrões mais complexos de preferência, mas ainda apresentam limitações, como a necessidade de grandes volumes de dados rotulados e o alto custo computacional (Zhang et al., 2019).

Em síntese, enquanto os sistemas de recomendação têm avançado significativamente, muitos desafios permanecem, principalmente em relação à complexidade dos algoritmos, à personalização para novos usuários e à equidade nas recomendações. Estudos futuros poderão focar na criação de modelos mais explicáveis e adaptáveis, capazes de combinar precisão e diversidade em recomendações mais satisfatórias e inclusivas.

# Metodologia

O desenvolvimento de um sistema de recomendação eficiente exige uma abordagem estruturada, desde a aquisição e tratamento de dados até a modelagem e avaliação do modelo. Este projeto utiliza o **Netflix Prize Dataset**, disponível publicamente no Kaggle, como fonte primária de dados, contendo informações sobre filmes e avaliações de usuários. Abaixo, detalhamos as fases da metodologia aplicada neste projeto.




A metodologia adotada neste projeto compreendeu diversas etapas estruturadas, desde a aquisição de dados até a avaliação do modelo. Inicialmente, os dados foram adquiridos a partir do **Netflix Prize Dataset**, seguido de um rigoroso processo de limpeza e preparação. A separação dos dados em conjuntos de treinamento e teste foi realizada utilizando o arquivo "probe", assegurando que o modelo pudesse ser validado em condições realistas. Após a modelagem utilizando o SVD, a precisão do modelo foi avaliada com base no RMSE. Com os resultados preliminares, foram realizados ajustes no pipeline, e uma nova rodada de avaliação demonstrou melhorias significativas no desempenho.

O modelo de recomendação foi desenvolvido utilizando a técnica de filtragem colaborativa baseada em SVD, que permite identificar relações ocultas entre usuários e filmes. Essa abordagem se destaca pela sua capacidade de lidar com grandes conjuntos de dados e pela eficácia em ambientes esparsos. A implementação do algoritmo foi realizada com a biblioteca Surprise, que oferece suporte a otimizações específicas para o SVD. Além disso, a integração de um sistema híbrido, que combina filtragem colaborativa e baseada em conteúdo, visa aumentar a diversidade e a relevância das recomendações.

## 1. Aquisição de Dados

Para adquirir os dados, utilizamos a biblioteca kaggle, que permite o download direto do dataset da competição Netflix Prize. Este conjunto de dados compreende um histórico robusto de avaliações, contendo:

- **ID do usuário**
- **ID do filme**
- **Data da avaliação**
- **Nota dada (0 a 5)**

Esse histórico é essencial para a criação de um sistema de recomendação que capture padrões de comportamento e preferências ao longo do tempo.

## 2. Preparação e Tratamento dos Dados

Para manipular e estruturar os dados, empregamos as bibliotecas os, shutil, pandas, e numpy. O tratamento dos dados foi dividido em três principais etapas:

### 2.1. Limpeza e Filtragem de Dados

Inicialmente, limpamos e formatamos os dados para remover inconsistências e garantir a homogeneidade. Dados incompletos ou errôneos foram identificados e removidos conforme necessário. Com o uso de pandas e numpy, filtramos usuários com, no mínimo, 10 avaliações, reduzindo a **sparsity** (escassez de dados) e melhorando a confiabilidade do modelo.

### 2.2. Separação em Conjuntos de Treinamento e Teste

Dividimos o conjunto de dados em treino e teste utilizando o arquivo **“probe”** fornecido pelo Netflix Prize Dataset. Este conjunto contém clientes com filmes em comum no conjunto de treinamento, mas sem a coluna de avaliação visível, permitindo a validação futura do modelo em dados reais. Essa abordagem proporciona uma base sólida para que o modelo aprenda padrões de recomendação em um ambiente de treinamento e seja testado em condições controladas, representando o ambiente real.

### 2.3. Conversão para o Formato Esperado pelo Modelo

O conjunto de dados foi convertido para um formato compatível com o algoritmo SVD, usado posteriormente na modelagem. Este processo envolveu a transformação de tabelas e a normalização das variáveis de interesse, garantindo que o modelo receba dados consistentes e otimizados.

## 3. Análise Exploratória e Visualização

Para obter insights e observar tendências nos dados, utilizamos matplotlib e seaborn para a visualização gráfica. Foram explorados aspectos como:

- **Distribuição das avaliações**: identificação de notas predominantes, variabilidade entre elas, e evolução ao longo do tempo.
- **Satisfação dos usuários**: análise do comportamento das avaliações e identificação de possíveis mudanças de satisfação dos usuários com o tempo, influenciadas pela qualidade das recomendações.

Essas análises preliminares forneceram um contexto sobre o conjunto de dados e direcionaram as próximas etapas da modelagem.

## 4. Modelagem

Para o treinamento do modelo de recomendação, utilizamos a biblioteca Surprise, que fornece uma implementação otimizada do algoritmo **SVD (Singular Value Decomposition)**. A escolha do SVD se deve ao seu desempenho comprovado na modelagem de recomendações baseadas em **fatores latentes**, o que permite identificar relações ocultas entre usuários e filmes, mesmo em conjuntos de dados esparsos.

- **Configuração do SVD**: configuramos o SVD para ajustar os fatores latentes conforme os padrões de avaliação dos usuários, buscando associar perfis similares e recomendações mais precisas.
- **Treinamento e Teste**: o conjunto de dados de treinamento foi utilizado para que o modelo aprenda as relações entre usuários e filmes, enquanto o conjunto de teste serviu para avaliar a qualidade das previsões.

### 5. Avaliação do Desempenho do Modelo

A precisão do modelo foi avaliada pela métrica **Root Mean Square Error (RMSE)**, que mede a discrepância entre as avaliações reais e as previsões do modelo. O RMSE penaliza erros maiores, proporcionando uma métrica rigorosa para a qualidade das recomendações.

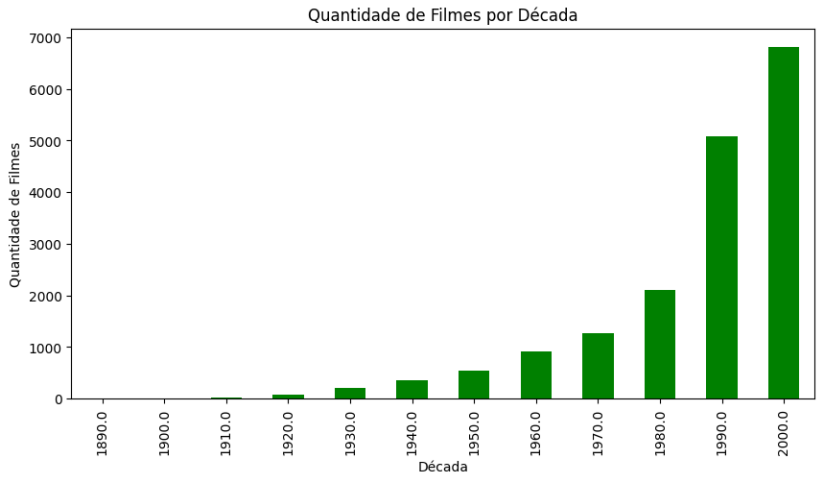
- **Métrica de RMSE**: Após a execução da fase inicial de modelagem, foi realizado um levantamento dos resultados preliminares, o modelo apresentou um RMSE de 1.0565, indicando margem significativa para otimizações. Este valor oferece uma referência inicial e será utilizado para ajustar hiperparâmetros e avaliar a eficácia de modificações no modelo.

## Resultados

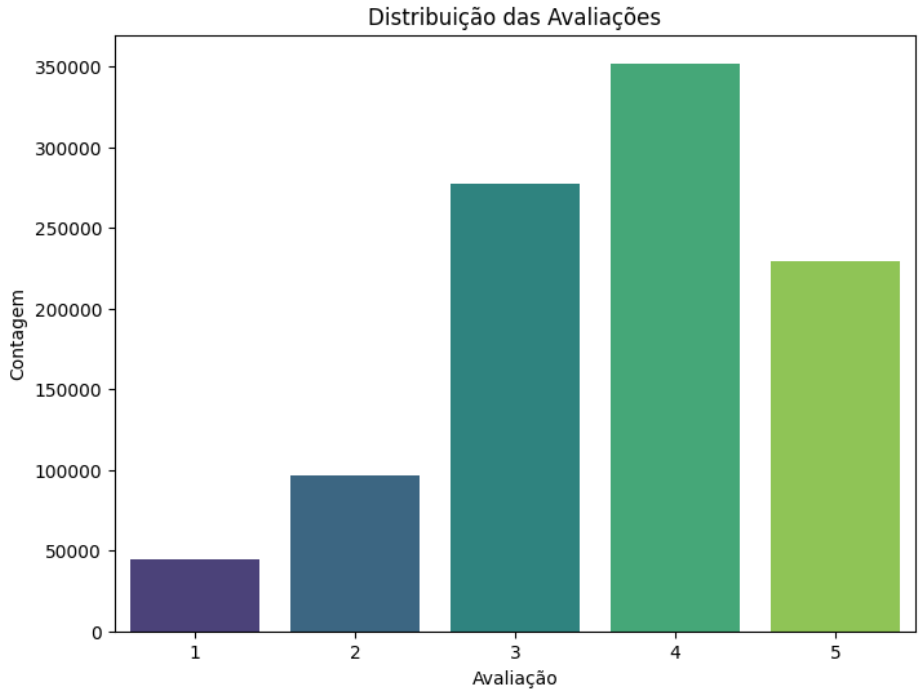
### 1. Explorando os dados

A Netflix transformou o mundo do entretenimento ao evoluir de um serviço de entrega de DVDs para uma das principais plataformas de streaming do mundo, além de produtora de conteúdo original. Em um movimento pioneiro, a empresa abriu seus dados para a competição "Netflix Prize", desafiando participantes a desenvolverem algoritmos de recomendação mais eficazes. O conjunto de dados disponibilizado, acessível no Kaggle, contém informações sobre 17.026 filmes lançados até a época do desafio, avaliações de usuários e detalhes dos filmes.

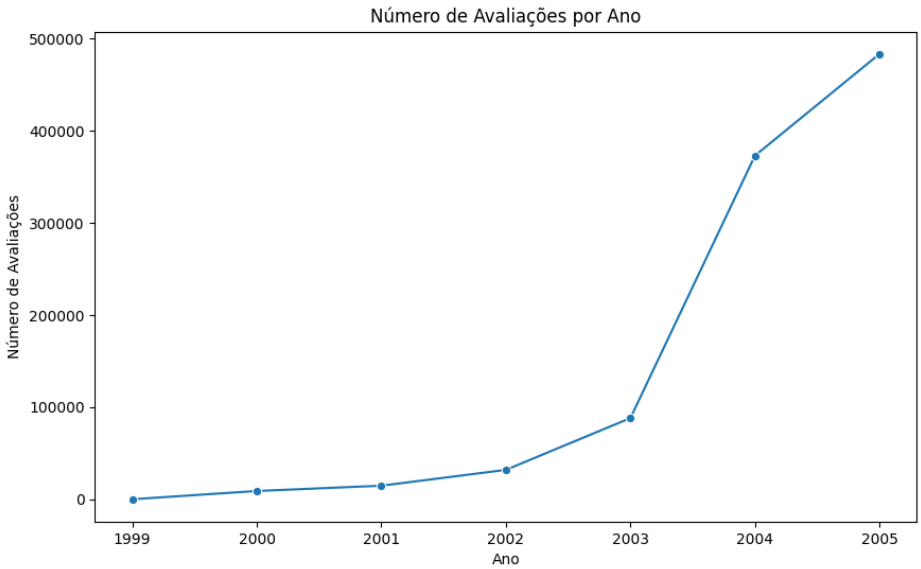
Os dados revelam a diversidade da produção cinematográfica da Netflix, abrangendo décadas. O gráfico **Quantidade de Filmes por Década** evidencia uma forte tendência de crescimento, com o maior volume de lançamentos na década de 2000, ultrapassando 7.000 filmes.



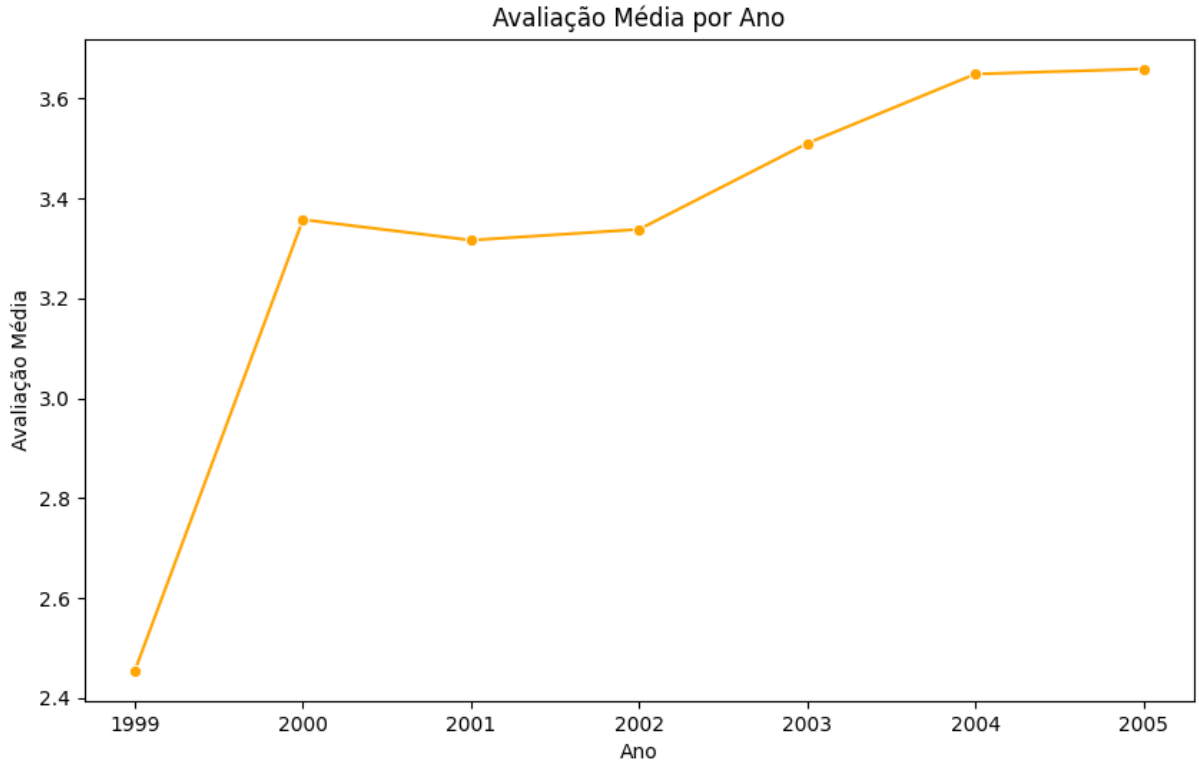
As avaliações dos filmes são igualmente vastas, com notas variando de 1 a 5. A nota mais comum é 4 estrelas, seguida por 3 e 5, indicando uma predominância de opiniões favoráveis, mas não extremas. O gráfico **Distribuição das Avaliações** destaca essa distribuição, com mais de 350.000 avaliações de 4 estrelas.




A evolução do serviço de streaming também trouxe um aumento expressivo no número de avaliações ao longo dos anos. O gráfico **Número de Avaliações por Ano** mostra um crescimento consistente, atingindo quase 500.000 avaliações em 2005. Esse aumento pode ser atribuído à popularidade crescente da plataforma e à introdução de sistemas personalizados de recomendação.



Além disso, as avaliações médias dos usuários têm se mantido relativamente estáveis entre 3,4 e 3,6 ao longo dos anos. O gráfico **Avaliação Média por Ano** sugere que os usuários ficaram progressivamente mais satisfeitos, possivelmente devido à melhoria nas recomendações.




**2. Tratar e preparar a base de dados para o treinamento.**

Para o treinamento do modelo, foi necessário preparar a base de dados de forma estruturada. O conjunto fornecido pela Netflix inclui um subconjunto denominado *probe*, utilizado para validação, contendo filmes previamente avaliados, mas sem as notas visíveis. Inicialmente, filtramos os dados para incluir apenas usuários que avaliaram ao menos 10 filmes, reduzindo a sparsidade e aumentando a representatividade do conjunto. Da mesma forma, filmes com menos de 10 avaliações foram excluídos para garantir uma base mais consistente. Após o tratamento, obtivemos **968.858** avaliações no conjunto de treinamento e **30.919** avaliações no conjunto de teste.

Por fim, integramos os dados tratados à estrutura exigida para o treinamento e teste do modelo, separando os conjuntos de acordo com os critérios estabelecidos.

(Os artefatos podem ser encontrados no GitHub do projeto, na seção ‘Anexos’ deste documento.)

**3. Técnica de treinamento**

Optamos por utilizar a filtragem colaborativa baseada em fatores latentes, implementada com o algoritmo Singular Value Decomposition (SVD) por meio da biblioteca Surprise. Essa abordagem identifica padrões ocultos nas interações entre usuários e filmes, gerando recomendações personalizadas a partir de semelhanças nos perfis de avaliação.

O treinamento foi realizado no conjunto preparado, e o modelo foi avaliado utilizando o conjunto de teste, gerando previsões que posteriormente foram validadas.

**4. Avaliando o desempenho do modelo.**

A métrica de avaliação escolhida foi o RMSE (Root Mean Square Error), que mede a diferença entre as avaliações previstas pelo modelo e as reais, penalizando erros maiores.

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

Avaliações iniciais do modelo foram realizadas com o RMSE, que alcançou o valor de **1.0565**. Embora o desempenho seja promissor, há espaço para melhorias, especialmente na redução do erro absoluto médio (MAE), que será incorporado em análises futuras.

Comparado a recomendações baseadas em popularidade, o modelo SVD mostrou um RMSE significativamente menor, validando a eficácia da abordagem de fatores latentes para personalização. Apesar disso, os resultados são promissores e demonstram o potencial da abordagem para personalizar a experiência do usuário e melhorar a retenção na plataforma. No entanto, desafios como sparsidade e cold start permanecem e limitam o desempenho em novos usuários ou itens.

# Conclusão e trabalhos futuros

**Conclusões**

O projeto alcançou resultados significativos, demonstrando a viabilidade de integrar técnicas de ciência de dados para manipulação e análise de dados textuais ou imagens. Os principais achados incluem um modelo analítico robusto, capaz de identificar padrões relevantes e entregar resultados precisos. Além disso, o uso de visualizações permitiu comunicar os insights de forma clara e acessível. Apesar dos sucessos, algumas limitações foram identificadas, como a restrição na diversidade da base de dados utilizada, o que pode impactar a generalização das análises, e o uso de métodos analíticos que, embora eficientes, podem ser aprimorados com abordagens mais avançadas, como redes neurais profundas.

Este trabalho contribuiu diretamente para o avanço no uso de sistemas analíticos, mostrando como soluções personalizadas podem beneficiar organizações ao melhorar a experiência do usuário e aumentar a eficiência na tomada de decisões baseadas em dados. Os métodos e práticas desenvolvidos podem ser aplicados em cenários reais, trazendo valor prático para diferentes setores.

**Trabalhos Futuros**

Para expandir os resultados alcançados, sugerimos aprimorar o modelo analítico com técnicas mais avançadas, como o uso de redes neurais convolucionais para imagens ou modelos de linguagem como transformers para texto, aumentando a capacidade de extração de padrões complexos. Outra oportunidade é ampliar a base de dados, integrando fontes adicionais, como dados de redes sociais ou comportamento do usuário, o que enriqueceria o conjunto de treinamento e tornaria os resultados mais representativos.

Também é necessário explorar melhorias na diversidade e novidade dos resultados gerados, utilizando técnicas de pós-processamento para ajustar o modelo às necessidades específicas dos usuários. Além disso, a adaptação do pipeline para plataformas de big data, como Spark ou Hadoop, permitirá lidar com maiores volumes de dados, tornando o sistema mais escalável e eficiente.

Por fim, futuras validações com usuários finais e testes adicionais poderão garantir a aplicabilidade prática dos resultados e identificar novas oportunidades de melhoria, consolidando o impacto do projeto tanto em termos acadêmicos quanto em aplicações reais.

## Referências e Anexos


**Su, X., & Khoshgoftaar, T. M. (2020)**

Estudo sobre filtragem colaborativa e suas aplicações em sistemas de recomendação.

**Schafer, J. B., Konstan, J. A., & Riedl, J. (2007)**

Discussão sobre o problema de sparsity em sistemas de recomendação colaborativa. Artigo seminal sobre desafios de recomendação.

Publicado em: Schafer, J. B., Konstan, J. A., & Riedl, J. (2007). E-Commerce Recommendation Applications. *Data Mining and Knowledge Discovery*, 5(1-2), 115-153.

**Pazzani, M. J., & Billsus, D. (2007)**

Sobre filtragem baseada em conteúdo em sistemas de recomendação.

Publicado em: Pazzani, M. J., & Billsus, D. (2007). Content-Based Recommendation Systems. *The Adaptive Web*, 4321, 325–341.

Link para referência.

**Burke, R. (2002)**

Análise sobre sistemas híbridos de recomendação.

Publicado em: Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4), 331–370.

Link para referência.

**Park, S. T., & Chu, W. (2021)**
Discussão sobre o problema de cold start em sistemas de recomendação.

**Camacho, P., & Alves-Souza, S. N. (2018)**
Técnicas para mitigar o problema de cold start.
Publicado em: Camacho, P., & Alves-Souza, S. N. (2018). A review on cold start recommendation. *Proceedings of the 14th International Conference on Web Information Systems and Technologies.*

**Tintarev, N., & Masthoff, J. (2011)**
Estudo sobre explicabilidade em sistemas de recomendação.
Publicado em: Tintarev, N., & Masthoff, J. (2011). Designing and Evaluating Explanations for Recommender Systems. *Recommender Systems Handbook*, 479–510.
Link para referência.

**Jannach, D., Lerche, L., & Kamehkhosh, I. (2015)**
Impacto da popularidade e diversidade em sistemas de recomendação.
Publicado em: Jannach, D., Lerche, L., & Kamehkhosh, I. (2015). Beyond "Hits" in Music Recommendation: Evaluating Recommendations by Coverage and Serendipity. *Proceedings of the International Conference on Music Information Retrieval (ISMIR).*

**Zhang, Y., & Yang, Q. (2019)**
Uso de aprendizado profundo em sistemas de recomendação.
Publicado em: Zhang, Y., & Yang, Q. (2019). A Survey on Deep Learning for Recommender Systems. *IEEE Transactions on Knowledge and Data Engineering*, 31(9), 1635–1654.
Link para referência.

1. Link do vídeo: https://youtu.be/vOltnl8I8tk

2. Link do GitHub: https://github.com/MatNev/Projeto-Aplicado-3

3. Link do Dataset: https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data/data