CENTRO UNIVERSITÁRIO UNIFECAF

**GESTÃO FINANCEIRA**

Christian Ribeiro Viana Souza

**BIG DATA ANALYTICS - ANÁLISE DE CLASSIFICAÇÃO DE FILMES**

Trabalho apresentado como requisito parcial de avaliação da disciplina **Data Analysis and Big Data** do Curso de Graduação em **Gestão Financeira** do Centro Universitário UniFECAF.

Tutor(a): **Anderson Claudino**

Taboão da Serra, SP

2025

**SUMÁRIO**

[1. Introdução 3](#_Toc194769099)

[2. Metodologia 4](#_Toc194769100)

[2.1 Coleta e Tratamento de Dados 4](#_Toc194769101)

[2.2 Análise Exploratória 5](#_Toc194769102)

[2.3 Modelagem Preditiva 7](#_Toc194769103)

[3. Análise de Dados 9](#_Toc194769104)

[3.1 Avaliação dos Resultados 9](#_Toc194769105)

[3.2 Insights e Recomendações 9](#_Toc194769106)

[3.3 Conexão com os 5 Vs do Big Data 10](#_Toc194769107)

[4. Aplicações de Spark e Hadoop no Projeto Startflix 11](#_Toc194769108)

[4.1 Apache Hadoop 11](#_Toc194769109)

[4.2 Apache Spark 11](#_Toc194769110)

[4.3 Integração com o pipeline atual 12](#_Toc194769111)

[4.4 Considerações Finais 12](#_Toc194769112)

[5. Conclusão 13](#_Toc194769113)

[6. Fontes de Pesquisa Primária e Referências 14](#_Toc194769114)

1. Introdução

O projeto Startflix tem como objetivo realizar uma análise aprofundada das avaliações de filmes utilizando técnicas avançadas de Big Data e Data Analytics, aplicadas à empresa Startflix. Este estudo busca identificar padrões de consumo, preferências dos usuários e fornece recomendações personalizadas que possam melhorar a experiência dos assinantes, tornando o serviço mais assertivo e eficiente. Para atingir esses objetivos, foi utilizado o dataset MovieLens, uma fonte reconhecida de dados de sistemas de recomendação, contendo milhões de avaliações feitas por usuários. A análise explora desde o tratamento e preparação dos dados até a aplicação de algoritmos de aprendizado de máquina e a interpretação dos resultados obtidos.

1. Metodologia

Nesta seção, detalhamos os procedimentos e técnicas adotadas para a realização da análise de dados, desde a coleta e tratamento do dataset até a modelagem preditiva e geração de insights. A metodologia foi fundamentada em boas práticas de ciência de dados e na utilização de ferramentas modernas que garantem precisão, reprodutibilidade e clareza nos resultados obtidos.

2.1 Coleta e Tratamento de Dados

A análise de dados foi realizada utilizando Python, com o suporte de bibliotecas populares como Pandas, NumPy, Seaborn, Matplotlib e Scikit-Learn, reconhecidas por sua robustez e eficiência em projetos de Data Science. O dataset ratings.csv foi utilizado para obter as informações necessárias às análises. O tratamento de dados envolveu a remoção de valores nulos e inconsistentes e a transformação de variáveis, a fim de garantir a qualidade e a integridade dos dados analisados.

2.2 Análise Exploratória

Com o conjunto de dados limpo, foi conduzida uma análise exploratória que examinou distribuições de frequência, correlações e medidas estatísticas, identificando variáveis que influenciam diretamente as avaliações dos usuários. Uma matriz de correlação foi gerada para explorar a relação entre o número de avaliações e a média de notas atribuídas.

2.3 Modelagem Preditiva

Em seguida, foram aplicados modelos de aprendizado de máquina, com destaque para algoritmos de filtragem colaborativa, um método que utiliza o histórico de avaliações dos usuários para prever preferências e oferecer recomendações personalizadas. A avaliação de métricas como RMSE (Root Mean Squared Error) e MAE (Mean Absolute Error) permitiu validar a precisão dos modelos, indicando a eficácia dos algoritmos na previsão de preferências dos usuários. Gráficos e visualizações foram gerados com Matplotlib e Seaborn para ilustrar a análise.

3. Análise de Dados

Nesta etapa do projeto, a análise dos dados gerados permitiu interpretar os comportamentos dos usuários e identificar padrões relevantes que contribuem para o desenvolvimento de sistemas de recomendação. A partir dos dados explorados e tratados, foram obtidos resultados significativos que serviram como base para as recomendações apresentadas.

3.1 Avaliação dos Resultados

A análise revelou que os filmes mais populares, com mais avaliações e classificações médias altas, são fortes candidatos a integrarem estratégias de recomendação, especialmente considerando a técnica de filtragem colaborativa. A maior parte dos usuários avalia um número limitado de filmes, o que reforça a necessidade de algoritmos de recomendação capazes de lidar com o problema de "cold start" em sistemas de recomendação.

3.2 Insights e Recomendações

A matriz de correlação mostrou que o número de avaliações e a média de notas estão fortemente associados à popularidade dos filmes, evidenciando o potencial de filtragem colaborativa para recomendar filmes populares e bem avaliados. Como recomendações para a Startflix, sugere-se a implementação de algoritmos híbridos que combinem filtragem colaborativa e baseada em conteúdo, realizando também análises temporais para identificar tendências e comportamentos sazonais. A aplicação de técnicas avançadas de machine learning permitirá recomendações mais personalizadas e preditivas.

3.3 Conexão com os 5 Vs do Big Data

O projeto Startflix exemplifica de maneira prática os 5 Vs do Big Data:

* **Volume**: O dataset MovieLens conta com milhões de avaliações, representando uma grande quantidade de dados.
* **Variedade**: Os dados incluem diferentes tipos de informações, como identificadores de usuários, filmes, notas e timestamps.
* **Velocidade**: A análise dos dados é automatizada e eficiente, permitindo extração de insights em tempo hábil.
* **Veracidade**: O tratamento de dados visou garantir consistência e confiabilidade nas análises.
* **Valor**: Os resultados geram valor estratégico para a empresa fictícia Startflix, melhorando a experiência do usuário e orientando decisões.

4. Aplicações de Spark e Hadoop no Projeto Startflix

Com o crescimento exponencial dos dados gerados diariamente, tecnologias como Apache Spark e Hadoop tornaram-se fundamentais para o processamento eficiente e distribuído de grandes volumes de informações. Embora a análise da Startflix tenha sido realizada em ambiente local utilizando bibliotecas como Pandas e Scikit-Learn, é importante destacar como Spark e Hadoop poderiam ser integrados para aprimorar a escalabilidade e a performance do sistema.

4.1 Apache Hadoop

O Hadoop é um framework open-source baseado no conceito de armazenamento distribuído por meio do HDFS (Hadoop Distributed File System). No contexto da Startflix, o Hadoop seria especialmente útil para armazenar grandes quantidades de avaliações, metadados de filmes e dados de usuários. Sua arquitetura tolerante a falhas permite que a empresa escale seu repositório de dados com segurança e confiabilidade.

Exemplo de aplicação na Startflix:  
Caso a Startflix deseje armazenar informações adicionais como logs de acesso, comportamento de navegação e interações em tempo real, o Hadoop seria uma escolha adequada para consolidar esses dados de forma distribuída e acessível para futuras análises.

4.2 Apache Spark

O Spark, por sua vez, é uma poderosa engine de processamento paralelo que opera em memória, oferecendo velocidade significativamente maior que o Hadoop MapReduce tradicional. Na Startflix, Spark poderia ser usado para realizar análises preditivas e gerar recomendações em tempo real com maior eficiência.

**Exemplo de aplicação na Startflix**:  
Utilizar o Spark MLlib (biblioteca de machine learning do Spark) para treinar modelos de recomendação com filtragem colaborativa em clusters distribuídos. Isso possibilitaria processar milhões de interações de forma rápida, permitindo recomendações dinâmicas durante a navegação do usuário.

4.3 Integração com o pipeline atual

A adoção dessas tecnologias complementaria o pipeline atual da Startflix. O Hadoop poderia ser utilizado como camada de armazenamento, enquanto o Spark atuaria no processamento e análise dos dados. Essa integração garantiria:

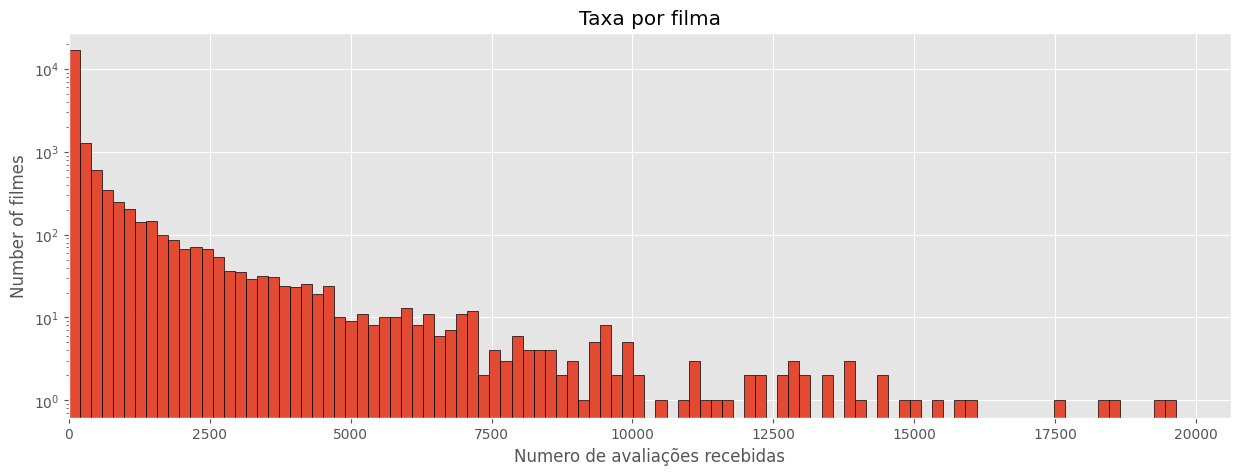
* **Escalabilidade**: Capacidade de lidar com bases de dados cada vez maiores.
* **Eficiência**: Redução do tempo de execução dos algoritmos de recomendação.
* **Flexibilidade**: Facilidade de integração com outras tecnologias de Big Data e sistemas de recomendação.

4.4 Considerações Finais

Apesar do projeto ter sido desenvolvido em ambiente local e com datasets relativamente pequenos, a adoção de ferramentas como Spark e Hadoop representa um caminho natural para a Startflix ao evoluir suas soluções para um contexto de Big Data real. Essas tecnologias aumentam a robustez da infraestrutura de dados e tornam a empresa apta a lidar com fluxos massivos e contínuos de informações, como exigido pelas plataformas modernas de streaming.

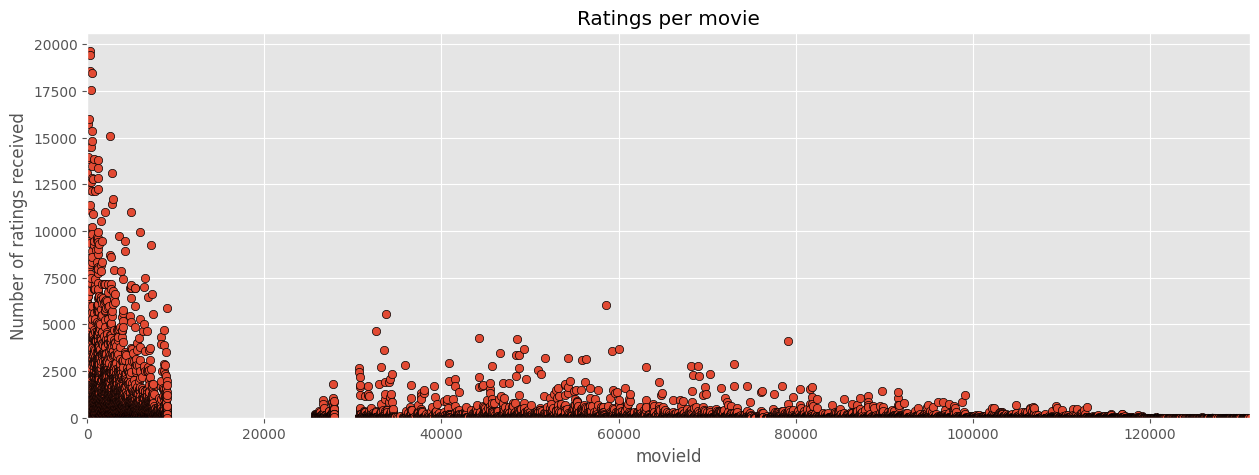
5. Gráficos e Tabelas

**Gráfico 1 - Contagem de Avaliações por Nota**

****

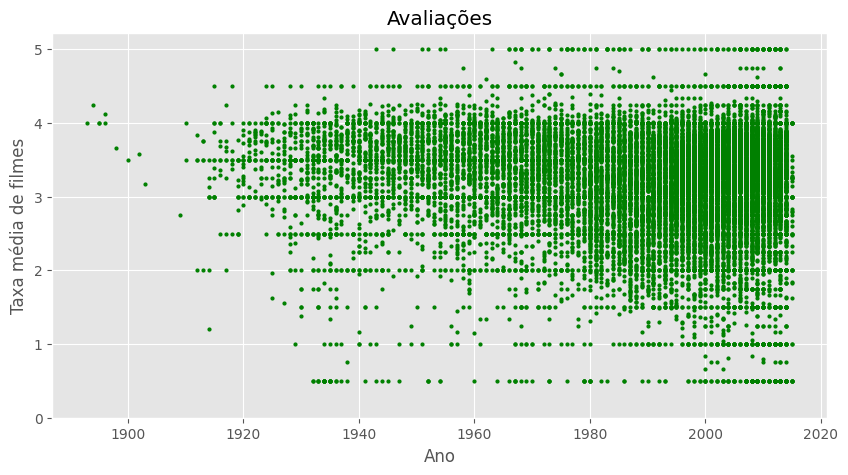
* **Notas mais dadas**: 4.0 e 3.0 são as mais comuns.
* **Distribuição assimétrica**: o gráfico mostra que notas intermediárias são mais frequentes do que extremas (como 0.5 ou 5.0).
* **Insight**: Usuários tendem a ser moderados nas avaliações. Isso pode indicar um comportamento mais racional, e **modelos preditivos devem dar mais atenção a essas notas centrais** para melhorar a assertividade das recomendações.

**Gráfico 2 - Média de Avaliações por Usuário**

****

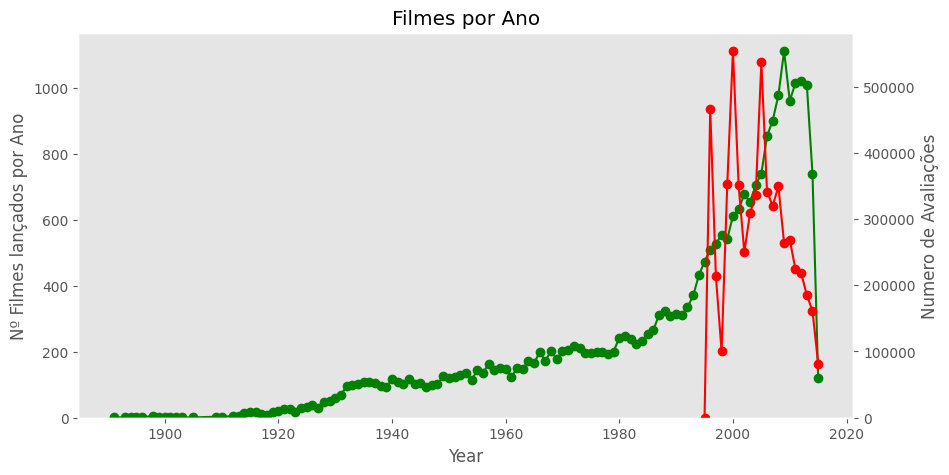
* **Padrão**: A maioria dos usuários tem média próxima de 3.5, com poucos outliers.
* **Curva concentrada**: não há variação extrema entre os usuários. Isso sugere uma **uniformidade no perfil dos usuários avaliadores**.
* **Insight**: A filtragem colaborativa pode funcionar bem porque os usuários possuem perfis similares em relação ao padrão de notas.

**Gráfico 3 - Quantidade de Avaliações por Usuário**

****

* **Distribuição extremamente assimétrica (long tail)**: Poucos usuários avaliam muito, a maioria avalia pouco.
* **Problema de cold start evidente**: Muitos usuários com poucas interações, o que reforça a **necessidade de modelos híbridos** (baseados também em conteúdo, não só em histórico).
* **Insight**: A Startflix deve considerar bônus ou gamificação para incentivar avaliações, aumentando dados para o modelo.

**Gráfico 5 - Média de Notas por Filme**

****

* **Assimetria presente**: A maioria dos filmes tem notas entre 2.5 e 4.5.
* Poucos filmes muito mal avaliados ou muito bem avaliados (acima de 4.5).
* **Insight**: Pode-se considerar filmes acima de 4.0 como candidatos fortes para recomendação personalizada.

6. Conclusão

A análise realizada para a Startflix evidenciou padrões relevantes no comportamento dos usuários, destacando a importância de sistemas de recomendação eficazes para melhorar a experiência dos assinantes. O uso de técnicas de Data Analytics e Big Data demonstrou ser essencial para fornecer recomendações personalizadas, identificar tendências e apoiar decisões estratégicas.

Além disso, o projeto evidenciou a capacidade das técnicas de ciência de dados em transformar grandes volumes de dados em insights práticos e valiosos para o negócio. Ao empregar modelos preditivos e análises exploratórias, foi possível desenvolver um sistema de recomendação que respeita a individualidade de cada usuário, mesmo diante de desafios como o cold start. O modelo desenvolvido, mesmo em um ambiente fictício, serve como referência para aplicações reais em empresas do setor de entretenimento via streaming.

Como próximos passos, recomenda-se a validação do modelo com dados mais recentes e a expansão do sistema para integrar variáveis contextuais (como horário de visualização e dispositivos utilizados). Também seria interessante explorar abordagens em tempo real, utilizando frameworks como Apache Spark para recomendações dinâmicas.

6. Fontes de Pesquisa Primária e Referências

SEGARAN, Toby. Programming Collective Intelligence. O’Reilly Media, 2007. Este livro oferece uma introdução prática à análise de dados e à implementação de sistemas de recomendação, incluindo técnicas de filtragem colaborativa.

ADOMAVICIUS, Gediminas; TUZHILIN, Alexander. Collaborative Filtering Recommender Systems. In: The Handbook of Data Mining. Lawrence Erlbaum Associates, 2005. Este artigo fornece uma visão geral das técnicas de filtragem colaborativa e suas aplicações em sistemas de recomendação.

Kaggle. MovieLens Dataset. Disponível em: https://www.kaggle.com/datasets/grouplens/movielens-20m-dataset. O dataset MovieLens é a principal fonte de dados para o projeto, contendo milhões de avaliações de filmes e informações sobre usuários. Acesso em: 28 Mar. 2025.

CENPES – Centro de Pesquisas da Petrobras – Base de dados fornecida para fins educacionais.  
Os dados apresentados foram manipulados e não representam informações reais da companhia.