CENTRO UNIVERSITÁRIO NOBRE
TECNÓLOGO EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

CÉSAR HENRIQUE SOUSA LIMA
GUSTAVO DA SILVA SANTOS

ANÁLISE DE DADOS DE SÉRIES E FILMES DA NETFLIX

FEIRA DE SANTANA
2025

CÉSAR HENRIQUE SOUSA LIMA
GUSTAVO DA SILVA SANTOS

ANÁLISE DE DADOS DE SÉRIES E FILMES DA NETFLIX

Trabalho para avaliação didática da disciplina de Data Science, do curso de Análise e Desenvolvimento de Sistemas.

Professor Orientador: Gledston Carneiro da Silva.

FEIRA DE SANTANA
2025

1. Descrição do Projeto

1.1. Objetivo

O objetivo deste projeto foi realizar uma **análise exploratória** e construir um **modelo preditivo** para estimar a pontuação IMDb de séries e filmes da Netflix. Este projeto é importante porque permite identificar padrões nos dados, compreender as preferências do público e fornecer insights valiosos para a indústria do entretenimento.

1.2. Fontes de Dados

Os dados utilizados foram obtidos a partir do conjunto de dados **"Netflix TV Shows and Movies"** disponibilizado no Kaggle. Este dataset contém informações como título, ano de lançamento, duração, gênero, classificação etária, país de produção, pontuação IMDb e quantidade de votos.

1.3. Métodos Planejados

Para atingir os objetivos do projeto, foram planejadas as seguintes etapas:

- **Coleta e preparação dos dados**: Tratamento de valores ausentes, remoção de duplicatas e normalização de variáveis.
- **Análise exploratória**: Visualização dos dados, identificação de padrões e correlação entre variáveis.
- **Modelagem preditiva**: Treinamento de um modelo **Random Forest Regressor** para prever a pontuação IMDb.
- **Avaliação do modelo**: Métricas de desempenho como **MAE, MSE e R^2 Score**.

2. Desenvolvimento do Projeto

2.1. Coleta de Dados

Os dados foram importados a partir do arquivo CSV original e passaram por um processo de limpeza:

- **Remoção de valores nulos** utilizando médias, medianas ou modas.
- **Eliminação de registros duplicados** para evitar distorções.
- **Conversão de variáveis categóricas** usando **One-Hot Encoding**.

2.2. Análise Exploratória de Dados

Durante a análise exploratória, foram identificados diversos padrões, incluindo:

- **Distribuição da pontuação IMDb**: A maioria dos títulos tem notas entre 5 e 8.
- **Gêneros mais comuns**: Drama, Comédia e Ação foram os mais populares.
- **Relação entre votos e pontuação**: Títulos com mais votos tendem a ter avaliações mais equilibradas.
- **Diferenças entre produções de diferentes países**: Filmes e séries de determinados países apresentam tendências de pontuação distintas.

2.3. Desenvolvimento do Modelo

- **Seleção de recursos**: Foram utilizadas variáveis como **ano de lançamento, duração, número de temporadas, votos no IMDb, classificação etária e gênero**.
- **Divisão dos dados**: Os dados foram separados em **80% para treinamento e 20% para teste**.

- **Modelo escolhido**: **Random Forest Regressor**, devido à sua capacidade de capturar relações não lineares e lidar com variáveis categóricas de forma eficiente.

2.4. Avaliação do Modelo

As métricas de desempenho foram:

- **Erro Absoluto Médio (MAE):** Indica a diferença média entre as previsões do modelo e os valores reais.
- **Erro Quadrático Médio (MSE):** Mede a magnitude do erro, penalizando grandes diferenças.
- **R^2 Score:** Avalia o quanto o modelo explica a variabilidade dos dados.

Os resultados indicaram que o modelo tem um bom desempenho, mas ainda pode ser refinado.

3. Considerações Finais

3.1. Desafios Enfrentados

Os principais desafios durante o projeto foram:

- **Tratamento de valores ausentes**, especialmente em variáveis categóricas.
- **Escolha e codificação de features**, garantindo que os dados fossem interpretáveis pelo modelo.
- **Interpretação dos resultados**, especialmente no impacto de diferentes variáveis na previsão da pontuação IMDb.

3.2. Aplicabilidade dos Resultados

Os resultados obtidos podem ser utilizados para:

- **Auxiliar plataformas de streaming** na seleção de conteúdo que tende a ter boas avaliações.
- **Direcionar estratégias de marketing** com base nos gêneros e características mais bem avaliadas.
- **Analisar padrões de consumo** e preferências do público ao longo dos anos.

3.3. Próximos Passos

Para aprimorar o projeto, algumas melhorias podem ser feitas:

- **Incluir mais fontes de dados**, como redes sociais e avaliações de outras plataformas.
- **Testar novos algoritmos**, como redes neurais e XGBoost, para melhorar a precisão da previsão.
- **Analisar fatores específicos**, como o impacto de diretores, roteiristas e atores nas notas IMDb.

Conclusão

Este projeto forneceu insights valiosos sobre os padrões de avaliação de filmes e séries da Netflix. O modelo preditivo teve um desempenho satisfatório, mas ainda pode ser aprimorado. As práticas aplicadas podem ser expandidas para outras plataformas de streaming e serviços que busquem entender as preferências do público.