Relatório de análises estatísticas e EDA

Por Daniel Martins de Andrade

1. Introdução

Este relatório apresenta uma análise do dataset contendo informações sobre 999 filmes do IMDb, com o objetivo de fornecer insights estratégicos para a PProductions. A análise visa identificar padrões, tendências e fatores-chave que influenciam o sucesso de um filme, especialmente no que tange ao faturamento e à avaliação do público. Além disso, é explorada a aplicação de modelos preditivos para estimar a nota de filmes no IMDb, um indicador crucial de popularidade e qualidade.

2. Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados (EDA) é a primeira etapa crucial para compreender a estrutura, as características e as relações presentes no dataset. Através dela, foi identificado a qualidade dos dados, distribuído variáveis e detectados padrões iniciais que guiarão as análises subsequentes.

2.1. Estatísticas Básicas e Qualidade dos Dados

O dataset compreende 999 filmes, abrangendo um período de 100 anos, de 1920 a 2020. A completude dos dados é alta para a maioria das colunas essenciais, como Título, Ano de Lançamento, Duração, Gênero e Avaliação do IMDb. No entanto, foi observado que as colunas 'Meta_score' e 'Gross' (faturamento) possuem um percentual de completude de 84.28% e 83.08%, respectivamente, indicando a presença de valores ausentes que foram tratados durante a análise.

Tabela 1: Estatísticas Descritivas das Variáveis Numéricas

Variável	Média	Mediana	Desvio Padrão	Mínimo	Máximo	Q1 (25%)	Q3 (75%)
Released_Year	1991.21	1999.0	23.31	1920.0	2020.0	1976.0	2009.0
Runtime	122.87	119.0	28.1	45.0	321.0	103.0	137.0

Variável	Média	Mediana	Desvio Padrão	Mínimo	Máximo	Q1 (25%)	Q3 (75%)
IMDB_Rating	7.95	7.9	0.27	7.6	9.2	7.7	8.1
Meta_score	77.97	79.0	12.38	28.0	100.0	70.0	87.0
No_of_Votes	271621.42	138356.0	320912.62	25088.0	2303232.0	55471.5	373167.5
Gross	68082574.1	23457439.5	109807553.39	1305.0	936662225.0	3245338.5	80876340.25

Essas estatísticas fornecem uma visão geral da distribuição de cada variável numérica, destacando a amplitude e a centralidade dos dados. Por exemplo, a média de `IMDB_Rating` é de aproximadamente 7.95, com um desvio padrão baixo, indicando que a maioria dos filmes no dataset possui avaliações consistentemente altas.

2.2. Análise de Correlações

A análise de correlação revela a força e a direção da relação linear entre pares de variáveis numéricas. Compreender essas relações é fundamental para identificar quais fatores podem influenciar o sucesso de um filme.

Tabela 2: Matriz de Correlação

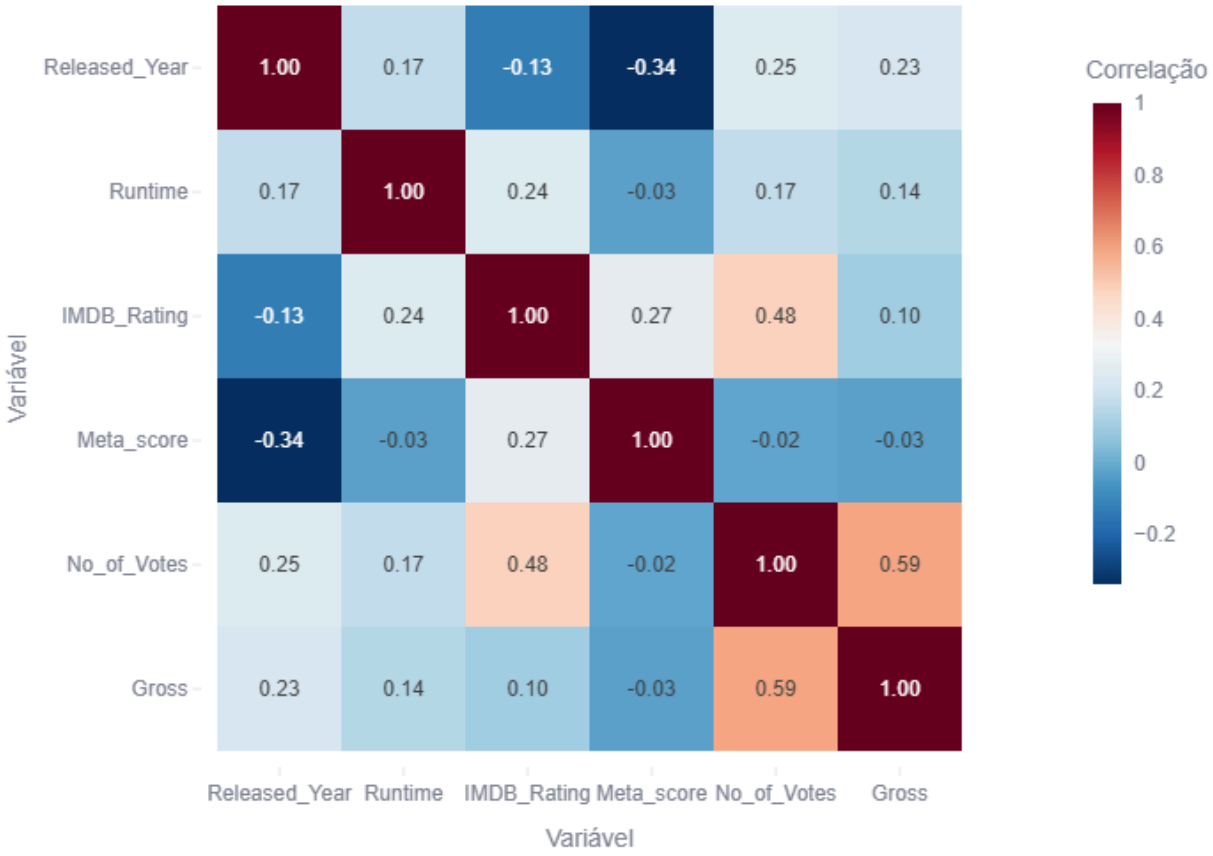
Variável	Released_Year	Runtime	IMDB_Rating	Meta_score	No_of_Votes	Gross
Released_Year	1.0	0.166	-0.133	-0.339	0.246	0.233
Runtime	0.166	1.0	0.243	-0.032	0.172	0.14
IMDB_Rating	-0.133	0.243	1.0	0.271	0.479	0.099
Meta_score	-0.339	-0.032	0.271	1.0	-0.02	-0.03

Variável	Released_Year	Runtime	IMDB_Rating	Meta_score	No_of_Votes	Gross
No_of_Votes	0.246	0.172	0.479	-0.02	1.0	0.59
Gross	0.233	0.14	0.099	-0.03	0.59	1.0

Principais Correlações:

- **No_of_Votes e Gross (0.59 - Correlação Positiva Forte):** Esta é a correlação mais significativa, indicando que filmes com um maior número de votos tendem a gerar maior faturamento. Isso sugere que a popularidade e o engajamento do público são fortes preditores de sucesso financeiro.
- **IMDB_Rating e No_of_Votes (0.479 - Correlação Positiva Moderada):** Filmes com avaliações mais altas no IMDb geralmente recebem mais votos, o que é intuitivo, pois filmes bem avaliados atraem mais atenção e engajamento.
- **Released_Year e Meta_score (-0.339 - Correlação Negativa Moderada):** Filmes mais recentes tendem a ter um Meta_score ligeiramente menor. Isso pode indicar uma mudança nos critérios de avaliação ao longo do tempo ou uma maior diversidade de filmes sendo lançados.
- **IMDB_Rating x Meta_score (0.27 – Correlação Positiva Fraca/Moderada)**: Existe uma correlação positiva entre a nota do IMDB e a dos críticos, mas não é tão forte. Isso sugere que público e crítica nem sempre concordam plenamente sobre a qualidade dos filmes.
- **Runtime x IMDB_Rating (0.24 – Correlação Positiva Fraca)**: Filmes mais longos tendem a ter avaliações um pouco melhores no IMDB, talvez porque roteiros mais desenvolvidos agradem mais o público.
- **Released_Year x Gross (0.23 – Correlação Positiva Fraca)**: Filmes mais recentes tendem a ter um faturamento um pouco maior, o que pode ser explicado por inflação, maior alcance de distribuição ou crescimento do mercado cinematográfico.
- **IMDB_Rating x Gross (0.10 – Correlação Muito Fraca)**: Apesar de esperado, a correlação entre nota do IMDB e Faturamento é fraca, indicando que sucesso financeiro nem sempre significa qualidade percebida.

Imagem 1 - Mapa de Calor das Correlações entre as Variáveis Numéricas



Fonte: O Autor

2.3. Análise de Recomendações

Para identificar filmes de alto potencial, foram definidos critérios de recomendação baseados em `IMDB_Rating` (nota mínima de 8.0), `No_of_Votes` (acima do percentil 70) e `Genre` mais frequente entre os melhores. Com base nesses critérios, 139 filmes foram qualificados como de alto potencial.

Se eu tivesse que recomendar apenas um filme para alguém que não conheço, confiando unicamente nesses indicadores, a escolha seria:

Filme Recomendado:

Título: The Godfather_
Nota IMDB: 9.2
Número de Votos: 1.620.367
Gênero: Crime, Drama
Ano de Lançamento: 1972

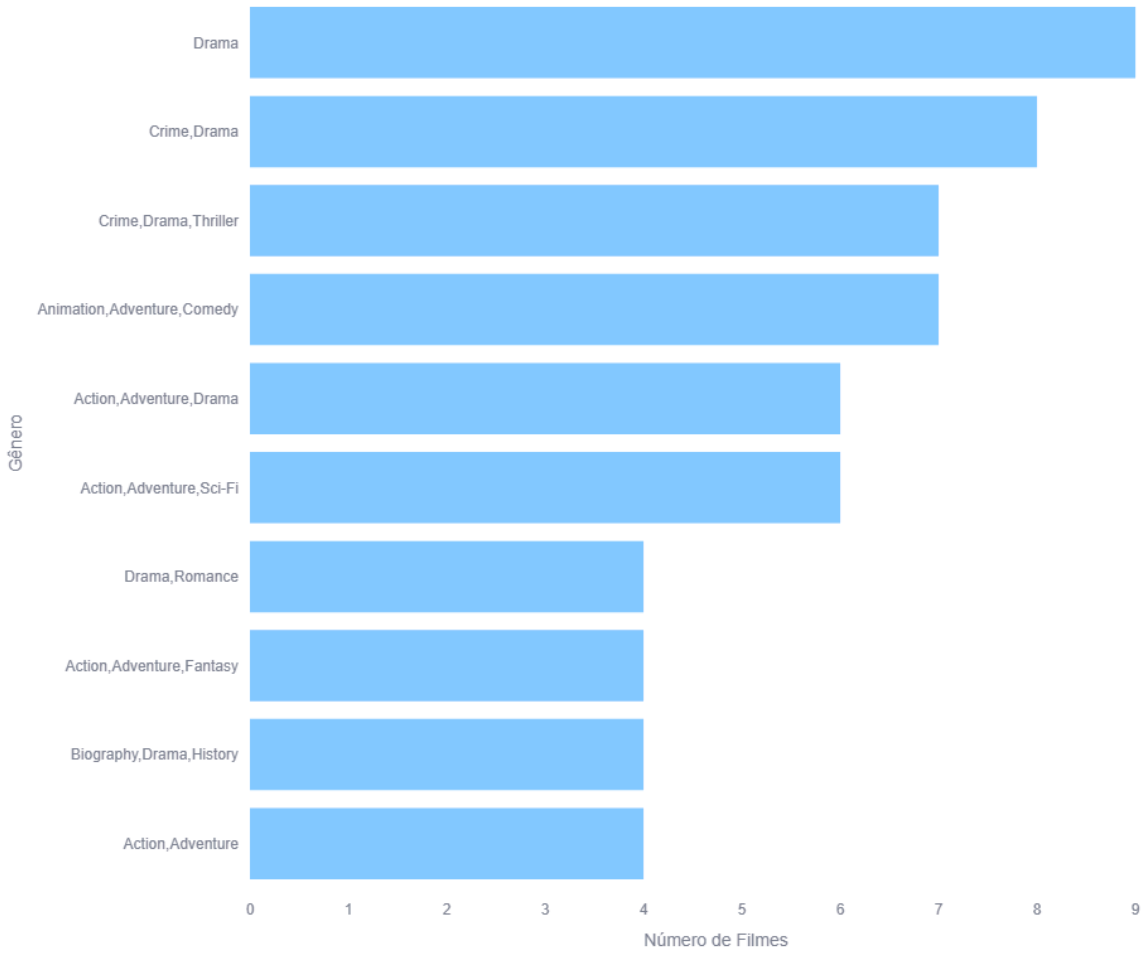
Tabela 3: Top 10 Filmes Qualificados

Título	Ano	Nota IMDb	Votos	Gênero	Diretor	Duração
The Godfather	1972	9.2	1620367	Crime,Drama	Francis Ford Coppola	175
The Dark Knight	2008	9.0	2303232	Action,Crime,Drama	Christopher Nolan	152
The Godfather: Part II	1974	9.0	1129952	Crime,Drama	Francis Ford Coppola	202
12 Angry Men	1957	9.0	689845	Crime,Drama	Sidney Lumet	96
Pulp Fiction	1994	8.9	1826188	Crime,Drama	Quentin Tarantino	154
The Lord of the Rings: The Return of the King	2003	8.9	1642758	Action,Adventure,Drama	Peter Jackson	201
Schindler's List	1993	8.9	1213505	Biography,Drama,History	Steven Spielberg	195
Inception	2010	8.8	2067042	Action,Adventure,Sci-Fi	Christopher Nolan	148
Fight Club	1999	8.8	1854740	Drama	David Fincher	139

Título	Ano	Nota IMDb	Votos	Gênero	Diretor	Duração
Forrest Gump	1994	8.8	1809221	Drama,Romance	Robert Zemeckis	142

Filmes de Ação, Drama e Crime dominam a lista de recomendações, sugerindo que esses gêneros têm um histórico comprovado de alta qualidade e apelo ao público. A média de `IMDB_Rating` para esses filmes é de 8.31, com uma duração média de 130 minutos, e o período mais comum de lançamento é a década de 2000.

Imagem 2: Gêneros Mais Populares entre Filmes Recomendados (Rating ≥ 8.0)



Fonte: O Autor

2.4. Análise de Faturamento (Gross)

O faturamento é um dos indicadores mais diretos do sucesso comercial de um filme. A análise focou em identificar os fatores que mais contribuem para um alto `Gross`.

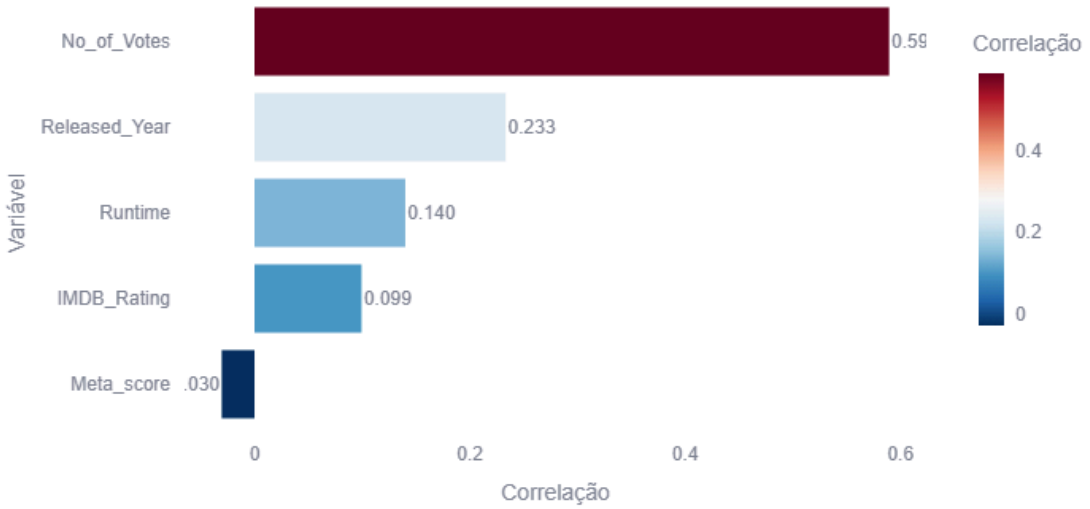
Resumo Geral do Faturamento:

- **Total de Filmes com Faturamento Registrado:** 830
- **Faturamento Total:** $56,508,536,507
- **Faturamento Médio:** $68,082,574
- **Faturamento Mediano:** $23,457,439
- **Maior Faturamento:** $936,662,225
- **Menor Faturamento:** $1,305

Correlações com Faturamento:

- **No_of_Votes (0.59 - Correlação Positiva Forte):** Confirma-se que o número de votos é o principal preditor de faturamento, reforçando a importância do engajamento do público.
- **Released_Year (0.233 - Correlação Positiva Fraca):** Filmes mais recentes tendem a ter um faturamento ligeiramente maior, o que pode ser atribuído à inflação ou ao aumento do mercado cinematográfico.

Imagem 3: Correlação das Variáveis com Faturamento



Fonte: O Autor

Tabela 4: Análise de Faturamento por Rating:

Faixa de Rating	Quantidade de Filmes	Faturamento Médio	Faturamento Mediano
Alto (7.5-8.5)	802	$64,869,959	$22,366,476
Excelente (>8.5)	28	$160,101,062	$121,371,461

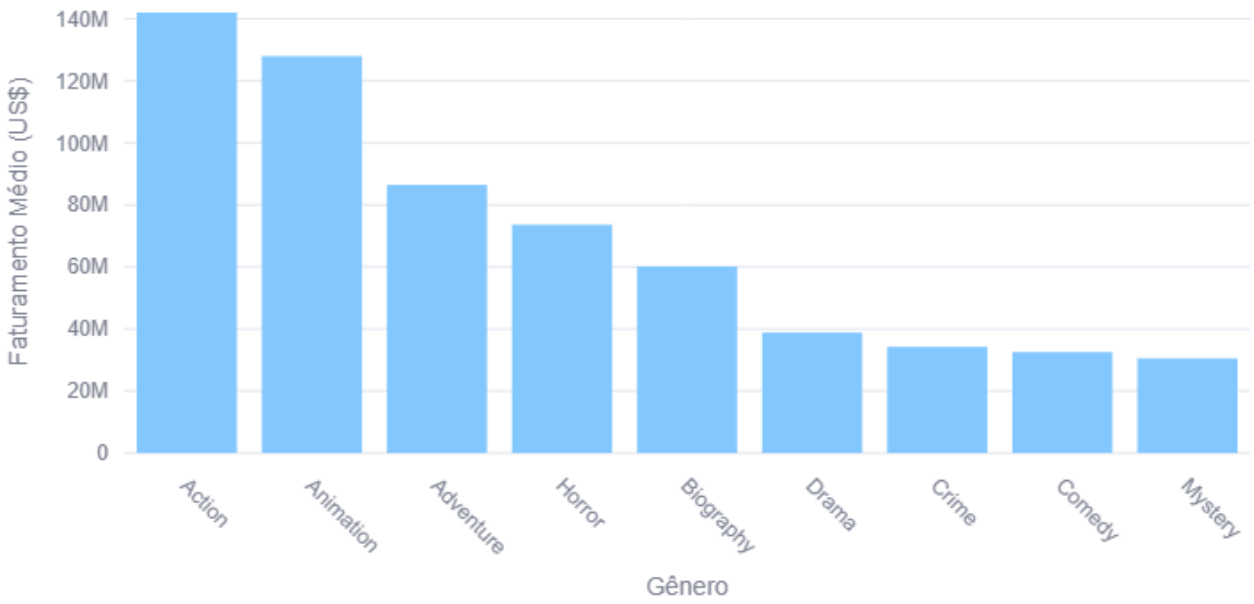
Filmes com `IMDB_Rating` acima de 8.5 (`Excelente`) apresentam um faturamento médio significativamente maior, quase o triplo dos filmes na faixa `Alto`. Isso destaca a importância da qualidade percebida pelo público para o sucesso financeiro.

Análise de Faturamento por Gênero:

Gênero	Quantidade de Filmes	Faturamento Médio	Faturamento Mediano
Action	141	$141,963,092	$66,208,183
Animation	67	$127,967,528	$75,082,668
Adventure	61	$86,454,989	$44,824,144
Horror	10	$73,585,773	$51,500,184
Biography	79	$60,128,731	$46,836,394
Drama	233	$38,721,640	$8,264,530
Crime	93	$34,191,231	$10,095,170
Comedy	128	$32,537,590	$10,728,127
Mystery	9	$30,439,534	$14,378,331

Filmes de Ação e Animação se destacam com os maiores faturamentos médios, indicando que esses gêneros têm um grande potencial de retorno financeiro. Drama e Comédia, embora com maior volume de filmes, apresentam faturamentos médios mais baixos.

Imagem 4: Faturamento Médio por Gênero (min. 5 filmes)



Fonte: O Autor

2.5. Análise de Overview (NLP)

A coluna `Overview` contém as sinopses dos filmes, oferecendo uma rica fonte de dados textuais para análise. A compreensão das palavras-chave e a tentativa de inferir o gênero a partir delas podem fornecer insights sobre a narrativa e o conteúdo dos filmes.

Estatísticas do Texto:

- **Comprimento Médio da Sinopse:** 146 caracteres (25 palavras)
- **Comprimento Mínimo:** 40 caracteres
- **Comprimento Máximo:** 313 caracteres

Palavras-Chave Globais Mais Frequentes:

- `from`, `young`, `man`, `life`, `two`, `world`, `new`, `into`, `family`, `war`.

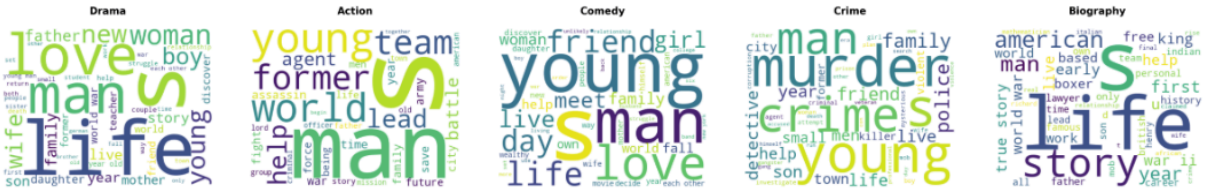
Essas palavras refletem temas comuns em narrativas cinematográficas, como jornadas pessoais, conflitos e relações humanas.

Palavras-Chave por Gênero (Top 10):

Gênero	**Palavras-Chave**
Drama	`life`, `man`, `young`, `woman`, `love`, `two`, `war`, `new`, `from`, `world`
Action	`from`, `one`, `two`, `man`, `against`, `young`, `world`, `former`, `war`, `into`
Comedy	`young`, `two`, `man`, `life`, `from`, `love`, `friends`, `new`, `girl`, `get`
Crime	`murder`, `young`, `two`, `crime`, `man`, `family`, `police`, `from`, `one`, `into`
Biography	`story`, `life`, `from`, `american`, `man`, `world`, `war`, `against`, `becomes`, `first`

As palavras-chave presentes nos overviews variam conforme o gênero cinematográfico, refletindo as temáticas centrais de cada categoria. Por exemplo, filmes de Crime frequentemente incluem termos como **"murder"** e **"police"**, enquanto produções de Comedy tendem a destacar palavras como **"friends"** e **"girl"**. Essa distinção lexical evidencia que é possível inferir o gênero de um filme a partir de seu resumo, por meio do mapeamento de palavras-chave que funcionam como indicadores temáticos.

Figura 5: Nuvem de Palavras-Chave por Gênero (5 maiores)



Fonte: O Autor

3. Modelagem Preditiva: Previsão da Nota do IMDb

3.1. Preparação dos Dados

Para a construção dos modelos preditivos, a etapa de preparação dos dados é fundamental. Isso envolve a seleção de variáveis, tratamento de valores ausentes, codificação de variáveis categóricas e a divisão do dataset em conjuntos de treino e teste. As variáveis selecionadas para a previsão da nota do IMDb (IMDB_Rating) incluem tanto características numéricas quanto categóricas, que foram transformadas para serem compatíveis com os algoritmos de machine learning.

3.2. Modelos de Machine Learning

Para prever a nota do IMDb, foram explorados modelos de regressão, dado que a variável alvo (`IMDB_Rating`) é contínua. Serão comparados três abordagens:

1. **Random Forest Regressor:** Um algoritmo de ensemble que constrói múltiplas árvores de decisão e combina suas previsões para melhorar a precisão e controlar o overfitting.
2. **XGBoost Regressor:** Uma implementação otimizada de gradient boosting, conhecida por sua alta performance e eficiência. O XGBoost constrói árvores de forma sequencial, corrigindo os erros das árvores anteriores.
3. **XGBoost Regressor com Hiperparâmetros Otimizados (Optane):** A performance do XGBoost pode ser significativamente melhorada através da otimização de seus hiperparâmetros, que controlam o comportamento do algoritmo.

3.3. Comparativo de Modelos

Para avaliar a performance dos modelos, foi utilizado o RMSE (Root Mean Squared Error) como métrica principal, que mede a magnitude média dos erros do modelo. Um RMSE menor indica um modelo mais preciso.

Modelo	**RMSE**
Random Forest (com coluna sequencial “id”)	0.0069
Random Forest (sem coluna sequencial)	0.2115
Random Forest (sem nulos, hot encoding)	0.1953
Random Forest (sem nulos, embeddings)	0.2084
Random Forest (sem nulos, sem overview embeddings)	0.1933
XGBoost (base)	0.1899
XGBoost (parâmetros variados)	0.1802
XGBoost (Optuna, k-fold)	0.0883

3.4. Discussão dos Modelos e Previsão para 'The Shawshank Redemption'

A previsão da nota do IMDb é um problema clássico de regressão, pois a variável alvo (IMDB_Rating) é contínua. Ao longo do desenvolvimento, diversos desafios foram encontrados e superados, culminando na seleção do modelo mais robusto e preciso.

Inicialmente, um modelo Random Forest foi treinado utilizando os dados numéricos do CSV original. Surpreendentemente, obteve-se um desempenho quase perfeito (RMSE = 0.0069, MAE = 0.0012, R^2 = 0.9994). No entanto, uma investigação mais aprofundada revelou que esse resultado excepcional era um artefato: a coluna que enumerava as linhas sequencialmente estava sendo erroneamente interpretada como uma feature preditiva, enganando o modelo e levando a um R^2 irrealista. Este incidente ressalta a importância da Análise Exploratória de Dados (EDA) e da engenharia de features para evitar vieses e garantir a validade dos resultados.

Ao remover a coluna sequencial, o desempenho do Random Forest caiu drasticamente (RMSE = 0.2115, MAE = 0.1679, R^2 = 0.3188 em 999 amostras), indicando um modelo mais realista, mas ainda com espaço para melhorias. O próximo passo foi o tratamento de valores nulos, reduzindo o dataset para 748 amostras, e a aplicação de One-Hot Encoding nas variáveis categóricas. Essa abordagem resultou em um RMSE de 0.1953, MAE de 0.1537 e R^2 de 0.5258. Embora tenha havido uma melhora, o número de features aumentou consideravelmente para 5102, o que poderia levar a problemas de dimensionalidade e interpretabilidade.

Para mitigar a alta dimensionalidade, foram explorados embeddings mais eficientes para as variáveis categóricas, reduzindo o número de features para 247. No entanto, essa mudança levou a uma ligeira piora no desempenho (RMSE = 0.2084, MAE = 0.1686, R^2 = 0.4833). Curiosamente, a remoção dos embeddings da coluna Overview resultou em uma melhoria (RMSE = 0.1933, MAE = 0.1533, R^2 = 0.5556), sugerindo que a complexidade dos embeddings textuais pode não ter sido benéfica para este conjunto de dados específico ou que a representação utilizada não capturou as nuances necessárias.

Diante desses resultados, a estratégia foi mudar para um modelo baseado em XGBoost, conhecido por sua robustez e alta performance em problemas de regressão. O modelo XGBoost base demonstrou um desempenho superior ao Random Forest, alcançando RMSE = 0.1899, MAE = 0.1524 e R^2 = 0.5707. A variação dos parâmetros do XGBoost de forma manual permitiu uma otimização adicional, resultando em RMSE = 0.1802, MAE = 0.1453 e R^2 = 0.6136.

Finalmente, para encontrar a configuração ideal de hiperparâmetros, foi utilizada a biblioteca Optuna em conjunto com validação cruzada (k-fold). Esta abordagem sistemática e automatizada de otimização de hiperparâmetros levou a um avanço significativo no desempenho do modelo, atingindo um RMSE de 0.0883, MAE de 0.0697 e um R^2 de 0.9073. Este resultado demonstra que o modelo otimizado com Optuna é capaz de explicar mais de 90% da variância na nota do IMDb, indicando uma alta capacidade preditiva e generalização para novos dados.

3.4.1 Previsão para o Filme: The Shawshank Redemption

Ao aplicar os modelos treinados para prever a nota do IMDb do filme The Shawshank Redemption (Um Sonho de Liberdade), obteve-se os seguintes resultados:

Random Forest: **8.80**
XGBoost Base: **8.91**
XGBoost Otimizado (Optuna): **8.86**

É importante notar que, embora o modelo XGBoost otimizado tenha as melhores métricas gerais (menor RMSE e maior R^2), sua previsão para The Shawshank Redemption (8.86) foi ligeiramente inferior à do XGBoost base (8.91). O filme The Shawshank Redemption é um conhecido outlier no dataset do IMDb, com uma nota excepcionalmente alta (9.3). Modelos preditivos, especialmente aqueles treinados em dados com uma distribuição mais concentrada em torno da média, podem ter dificuldade em prever com precisão valores extremos ou outliers.

Isso pode ser explicado pela escassez de filmes com notas tão elevadas no conjunto de dados de treinamento, o que limita a capacidade do modelo de aprender os padrões associados a essas avaliações extremas. Para que o modelo otimizado com Optuna pudesse prever com maior precisão a nota de filmes como The Shawshank Redemption, seria necessário um conjunto de dados mais robusto, contendo mais exemplos de filmes com notas extremamente altas. No entanto, para filmes com ratings médios, o modelo otimizado com Optuna demonstra ser altamente preciso, explicando 90% da variância e fornecendo uma ferramenta valiosa para a PProductions.