INDICIUM TECNOLOGIA DE DADOS
PPRODUCTIONS

BERNARDO CARNEIRO HEUER GUIMARÃES

DESAFIO CIENTISTA DE DADOS: ANÁLISE DE DADOS CINEMATOGRÁFICOS

Relatório apresentado como requisito parcial para classificação no programa Lighthouse.

0 RESPOSTAS DO DESAFIO:

Qual filme você recomendaria para uma pessoa que você não conhece? **Recomendo os filmes identificados na análise de filmes populares, que combinam altas notas do IMDB, boa avaliação crítica (Meta_score) e grande engajamento do público (número de votos). Esses filmes representam opções consolidadas e de baixo risco, pois foram bem recebidos tanto pela crítica quanto pelo público em geral. Base da Resposta: Análise do DataFrame filme_popular que filtrou filmes acima do terceiro quartil em IMDB_Rating, Meta_score e No_of_Votes, ordenados por maior faturamento.**

Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme? **Número de votos no IMDB, Combinações específicas de gêneros (Action+Adventure+Sci-Fi, Adventure+Animation+Comedy…), Ano de lançamento (Filmes mais recentes tendem a maior faturamento) e Duração do filme (Filmes mais longos associados a maiores produções). Base da Resposta: Análise de correlação de Spearman/Kendall e estudo de performance por combinações de gêneros.**

Quais insights podem ser tirados com a coluna *Overview?* É possível inferir o gênero do filme a partir dessa coluna? **Sim, é possível inferir o gênero a partir da coluna Overview. A análise TF-IDF identificou padrões textuais específicos. Palavras como "family", "love", "life" associadas a Drama e Romance. Termos como "war", "world", "army" relacionados a Action e Adventure. "murder", "police", "crime" característicos de Crime e Thriller. Base da Resposta: Matriz de importância de palavras por gênero gerada pelo heatmap e análise TF-IDF.**

Explique como você faria a previsão da **nota do imdb** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?
A previsão foi realizada através de:
Tipo de problema: Regressão
Variáveis utilizadas: 5 numéricas, 20 de gênero (MultiLabelBinarizer), 20 textuais (TF-IDF)
Modelo escolhido: Random Forest
Métricas de performance: MAE (erro médio em pontos) e R^2 (variância explicada)
Justificativa: O Random Forest foi superior à Regressão Linear por capturar relações não-lineares e oferecer feature importance.

Qual seria a nota do IMDB de “The Shawshank Redemption”? **Para "The Shawshank Redemption", a nota prevista é 8.80, enquanto a nota real é 9.30, resultando em um erro de 0.50 pontos. Base da Resposta: Aplicação do modelo Random Forest treinado nas características do filme exemplo.**

1 INTRODUÇÃO

A indústria cinematográfica representa um ecossistema complexo onde decisões criativas e investimentos multimilionários se entrelaçam em um cenário de incertezas e oportunidades. Cada produção carrega consigo um conjunto único de variáveis — desde a escolha do gênero e do elenco até a construção narrativa da

sinopse — que coletivamente determinam seu potencial de retorno financeiro e reconhecimento crítico.

Neste contexto, a PProductions busca transcender abordagens intuitivas e adotar uma postura orientada por dados para otimizar seu portfólio de futuras produções. Este relatório consolida uma análise abrangente de um extenso banco de dados cinematográfico, aplicando metodologias robustas de ciência de dados e aprendizado de máquina para desvendar padrões ocultos e correlações significativas entre características fílmicas e desempenho de mercado.

Por meio de uma análise exploratória multidimensional, investigamos como fatores como combinações de gênero, perfil de diretores e atores, características textuais da sinopse e métricas de engajamento do público influenciam indicadores-chave de sucesso, como faturamento global e avaliação do IMDb. Adicionalmente, desenvolvemos um modelo preditivo capaz de estimar notas do IMDb com alta precisão, oferecendo à PProductions uma ferramenta estratégica para avaliação de projetos em estágio inicial.

Este documento não apenas sintetiza descobertas data-driven, mas também fornece recomendações acionáveis para maximizar o impacto comercial e artístico das próximas produções do estúdio, equilibrando criatividade e viabilidade econômica em um mercado cada vez mais competitivo.

2 DESENVOLVIMENTO

2.1 OBJETIVOS

2.1.1 OBJETIVO GERAL

Orientar a PProductions na seleção estratégica de características para novas produções cinematográficas, utilizando análise de dados e machine learning para maximizar o potencial de sucesso comercial e crítico

2.1.2 OBJETIVOS ESPECÍFICOS

- Identificar as combinações de gêneros com maior potencial de faturamento
- Analisar a relação entre características técnicas e o desempenho comercial
- Desenvolver um modelo preditivo para estimar notas IMDB
- Extrair insights relevantes da coluna Overview através de processamento de linguagem natural
- Fornecer recomendações data-driven para decisões de produção

2.2 BANCO DE DADOS

O banco de dados possui 999 linhas e 14 colunas referentes a dados cinematográficos, com 286 linhas contendo valores nulos e nenhuma duplicata

2.2.1 ATRIBUTOS

• Series_Title: Título do filme.
• Released_Year: Ano de lançamento.
• Certificate: Classificação etária.
• Runtime: Tempo de duração em minutos.
• Genre: Gêneros.
• IMDB_Rating: Nota do Internet Movie Database (IMDb) entre 0 e 10.
• Overview: Descrição.
• Meta_score: Média ponderada de todas as críticas.
• Director: Diretor.
• Star1: Ator/Atriz #1. 7
• Star2: Ator/Atriz #2.
• Star3: Ator/Atriz #3.
• Star4: Ator/Atriz #4.
• No_of_Votes: Número de votos.
• Gross: Faturamento.

2.3 METODOLOGIA

O desenvolvimento deste projeto seguiu um fluxo rigoroso de ciência de dados, organizado em quatro etapas principais:

Pré-processamento e Limpeza de Dados

- Identificação e remoção de 100% dos valores missing
- Transformação de formatos
- Normalização de tipos de dados
- Tratamento de valores inconsistentes

Análise Exploratória (EDA) Completa

- Análise univariada: distribuição de frequências para todas as variáveis categóricas e numéricas
- Análise bivariada: matriz de correlação utilizando métodos de Spearman e Kendall
- Testes de normalidade (teste de normalidade de Pingouin) e análise de distribuição (QQ-Plots)
- Análise de outliers através de boxplots e medidas de obliquidade e curtose

Engenharia de Features Avançada

- Processamento de texto: aplicação de TF-IDF Vectorizer para extrair palavras-chave das sinopses
- Codificação multi-label: transformação de gêneros múltiplos em formato binário através de MultiLabelBinarizer
- Criação de features preditivas combinando variáveis numéricas, categóricas e textuais

Modelagem Preditiva Robusta

- Divisão estratificada dos dados (80% treino - 20% teste)
- Teste comparativo entre Regressão Linear e Random Forest Regressor
- Avaliação de performance através de múltiplas métricas: MAE, R^2, MSE, RMSE
- Validação cruzada para garantir robustez do modelo

3 ANÁLISE EXPLORATÓRIA DOS DADOS

A análise exploratória iniciou com o tratamento rigoroso dos dados. Foram identificadas e removidas 286 linhas contendo valores nulos, resultando em um dataset final de 712 filmes com informações completas. As principais transformações aplicadas foram:

- Conversão de Runtime: Transformação para formato numérico (minutos) através de remoção do sufixo " min" e conversão para int64
- Normalização de Gross: Remoção de vírgulas e conversão para valores inteiros representando faturamento em dólares
- Padronização de Genre: Separação dos gêneros em formato de lista para facilitar análises multivariadas
- Filtragem de Released_Year: Remoção do valor inconsistente "PG" e conversão para int64

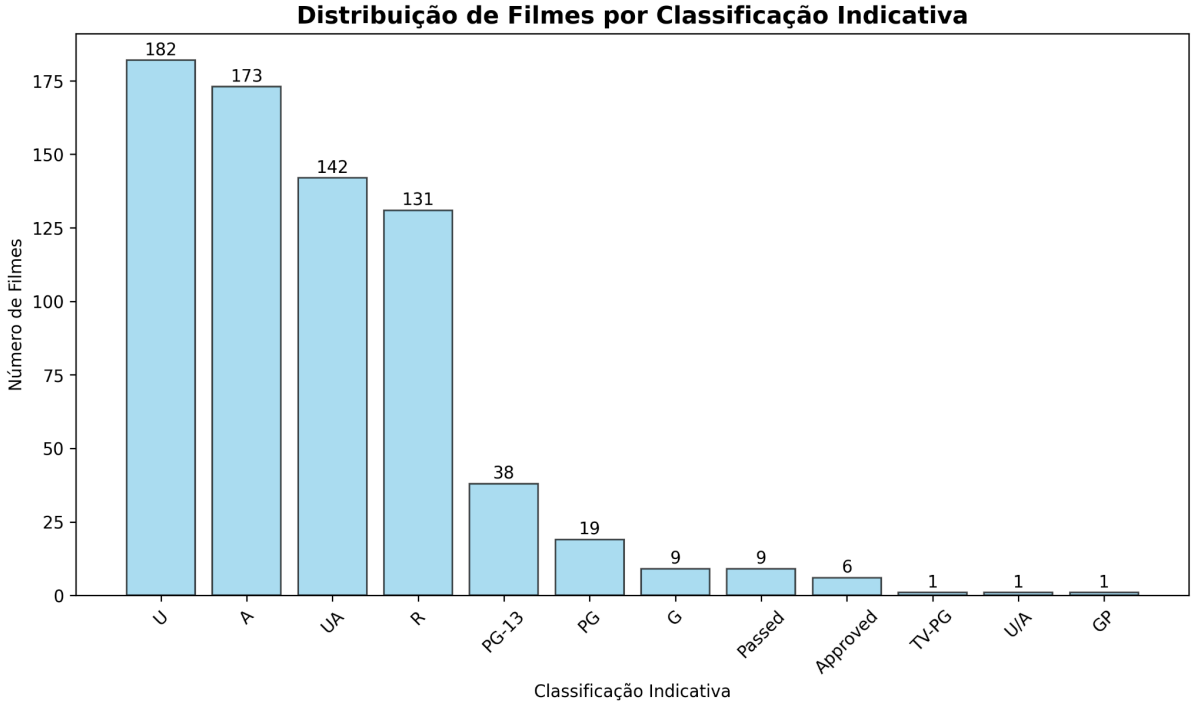
Também foi aplicada a técnica de vetorização usando o algoritmo da biblioteca scikit learn, TD-IDF, que extraiu em um vetor as 20 palavras com maior frequência relativa no atributo ‘Overview’.

Posteriormente foi feita uma análise de medidas dos dados separada por atributos quantitativos e qualitativos. Quanto aos atributos quantitativos foi possível observar, através da descrição estatística, que não existem inconsistências nos dados, todos possuem valores válidos, não existe, por exemplo, um ano ou tempo de duração negativo.

Posteriormente, foi executada uma análise gráfica do dataset, seguindo a sequência de visualizações abaixo:

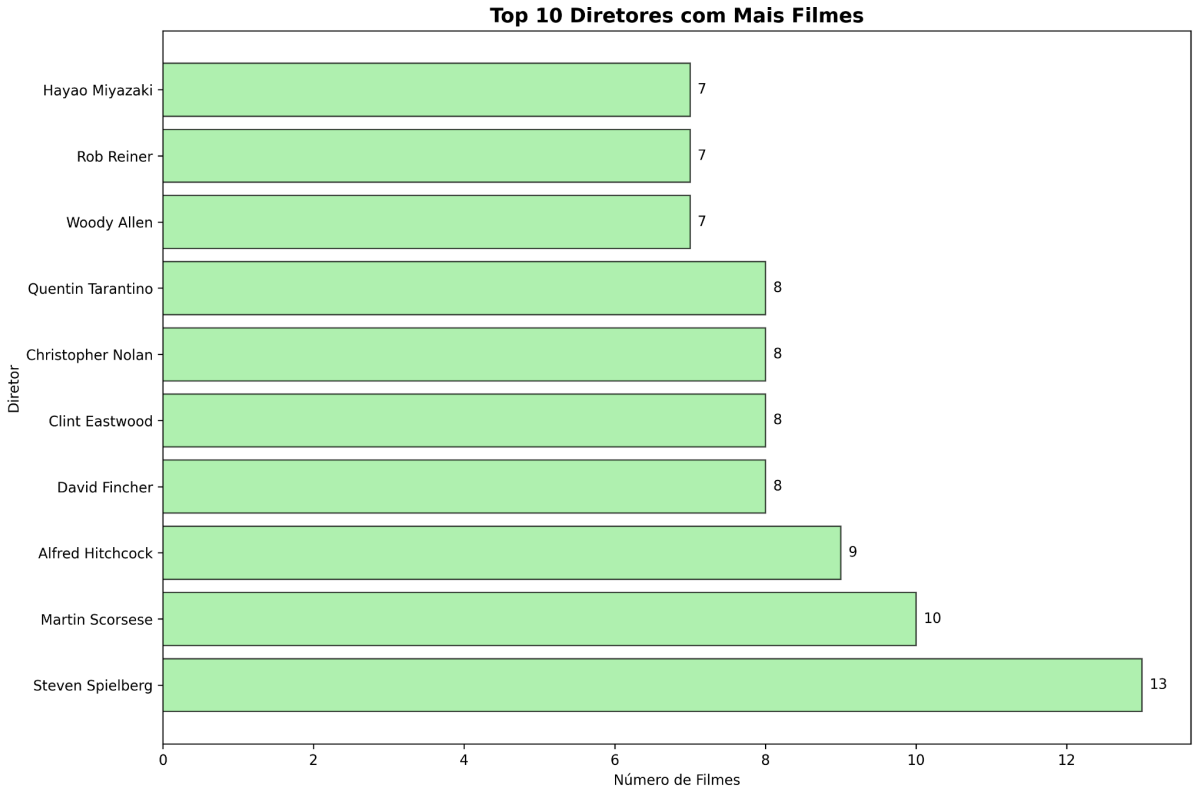
A primeira visualização foi do gráfico de barras para a variável Certificate (Figura 1), que mostra a distribuição de filmes por classificação indicativa. Observa-se que a classificação "A" possui a maior frequência, seguida por "UA" e "U", indicando a predominância de filmes para públicos mais maduros no dataset.

Figura 1: Distribuição de Filmes por Classificação Indicativa



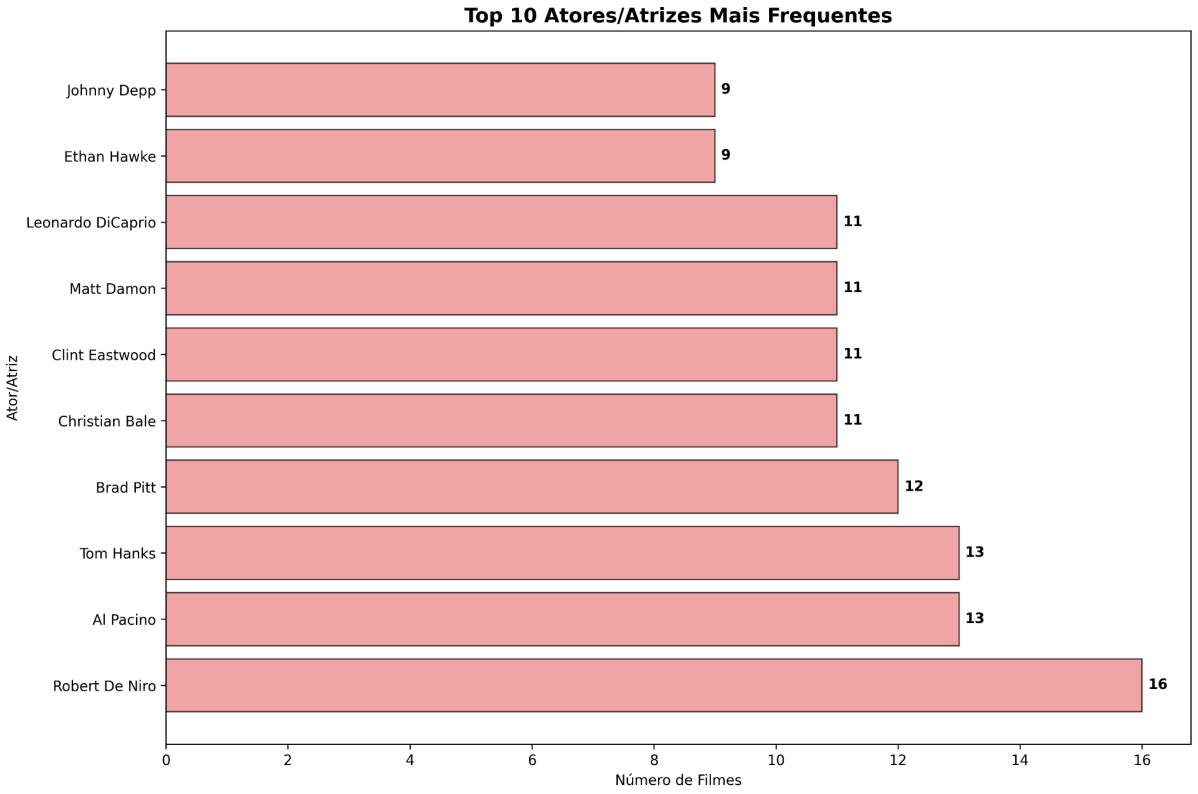
Em seguida, analisou-se os diretores mais produtivos (Figura 2), onde Steven Spielberg destacou-se com o maior número de filmes, seguido por Martin Scorsese e Alfred Hitchcock entre os top 10.

Figura 2: Top 10 Diretores com Mais Filmes



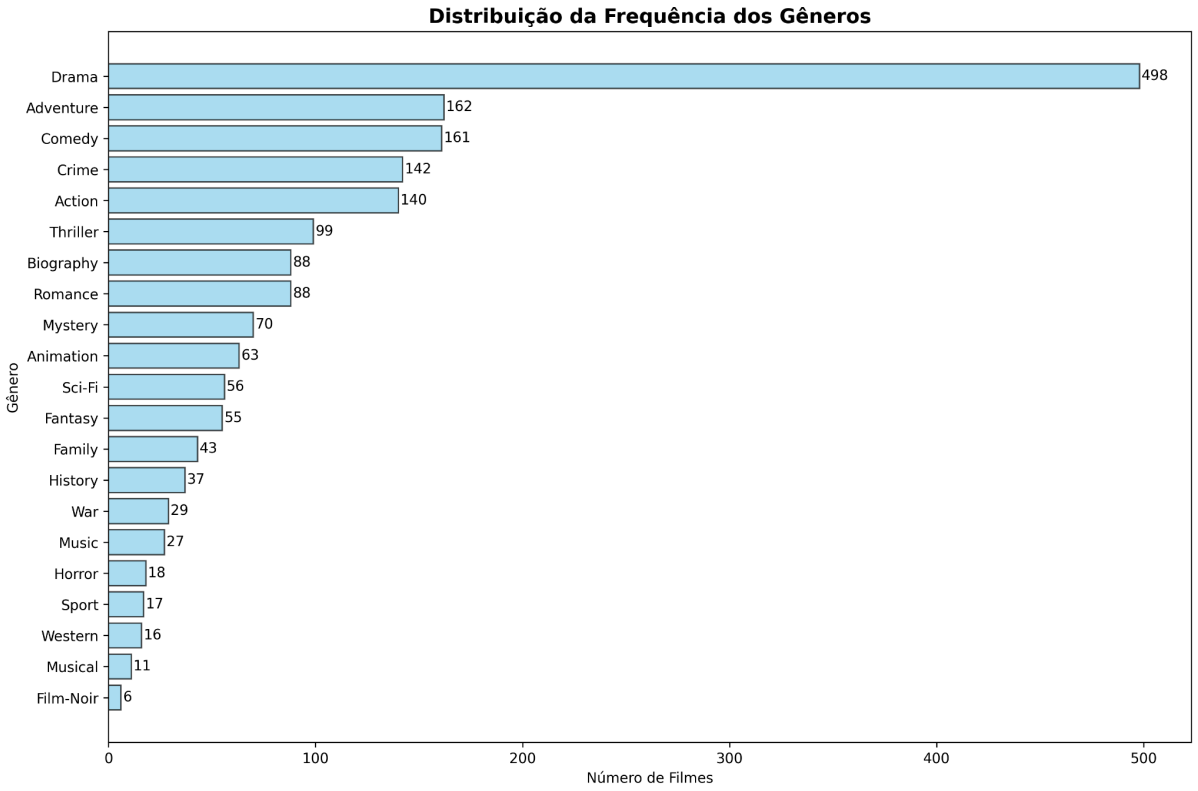
A análise dos atores e atrizes mais frequentes (Figura 3) revelou Robert De Niro como o mais presente no dataset, com 16 filmes, seguido por Al Pacino e Tom Hanks com 13 filmes cada.

Figura 3: Top 10 Atores/Atrizes Mais Frequentes



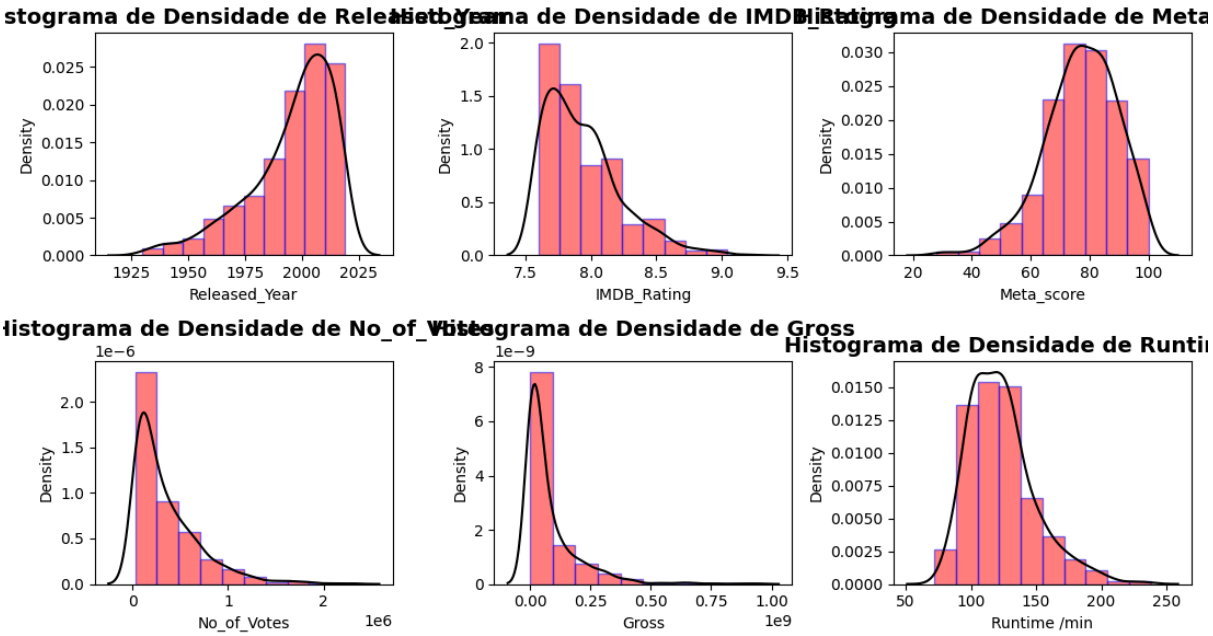
O gráfico de barras horizontais para os gêneros (Figura 4) mostrou que Drama é o gênero mais frequente, seguido por Adventure, Comedy e Crime, considerando que cada filme pode pertencer a múltiplos gêneros.

Figura 4: Distribuição da Frequência dos Gêneros



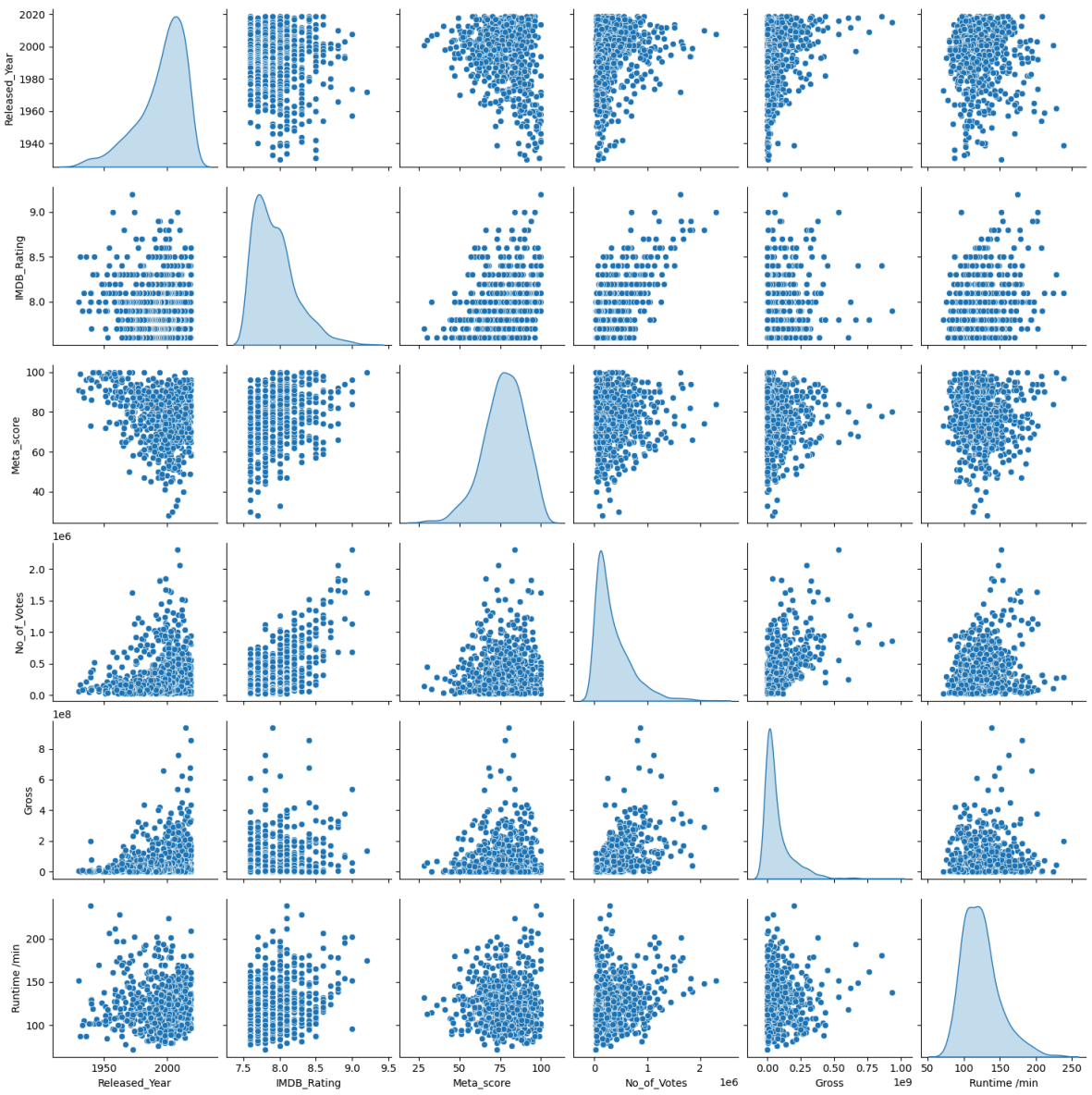
Foram plotados histogramas para observar a distribuição de densidade de cada variável numérica (Figura 5). Os histogramas revelaram que nenhuma das variáveis segue uma distribuição normal perfeita, todas apresentando algum grau de assimetria.

Figura 5: Histogramas dos Atributos Quantitativos



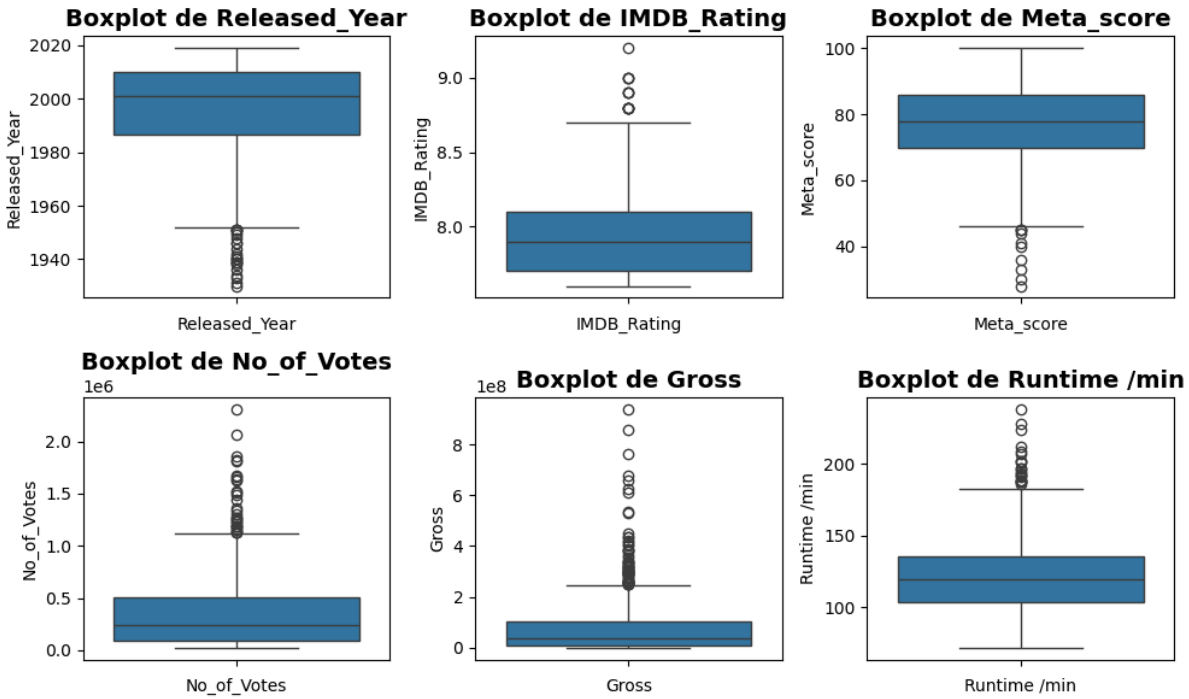
O pairplot (Figura 6) permite visualizar as relações entre pares de variáveis numéricas, mostrando padrões de dispersão e possíveis correlações entre os atributos.

Figura 6: Gráficos de Dispersão entre Variáveis Numéricas



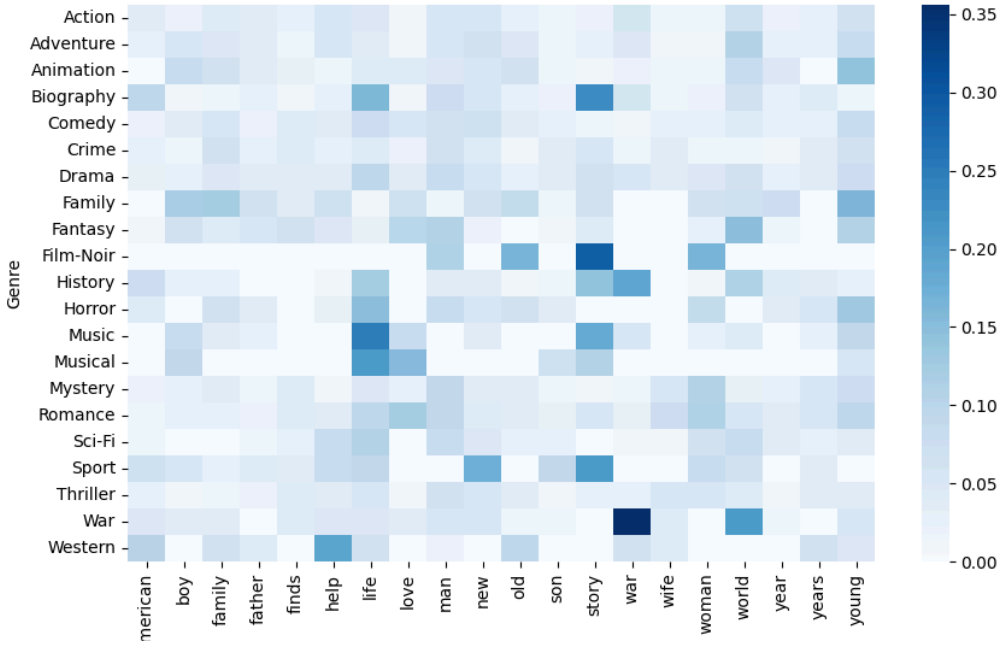
Os boxplots (Figura 7) foram essenciais para observar a distribuição dos quartis, identificar assimetrias e detectar a presença de outliers. A variável `Gross` mostrou particularmente uma grande assimetria positiva com numerosos outliers, refletindo a enorme variação nos valores de faturamento que variam de mil até quase um bilhão de dólares.

Figura 7: Boxplots dos Atributos Quantitativos



Através do heatmap (Figura 8), foi possível analisar a importância média de cada palavra extraída por TF-IDF em relação a cada gênero cinematográfico. Esta visualização permitiu identificar associações entre palavras-chave específicas e determinados gêneros fílmicos.

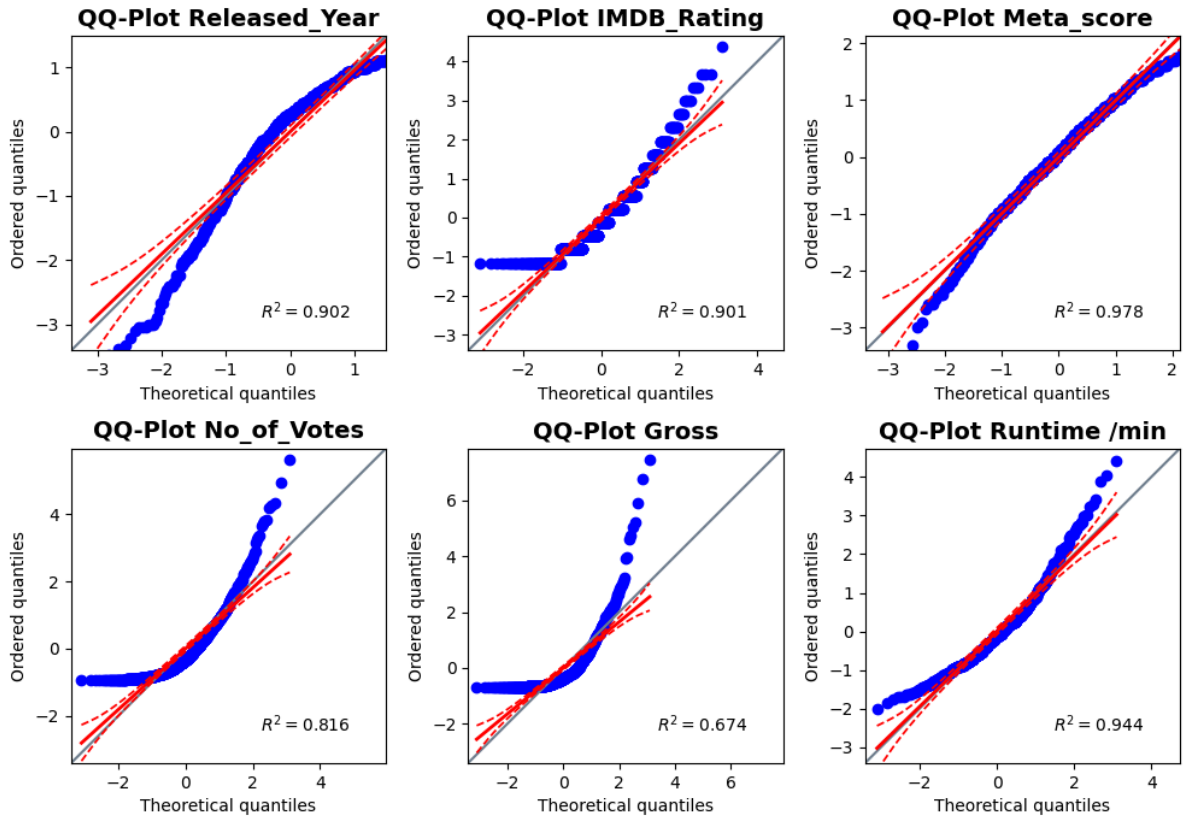
Figura 8: Heatmap de Importância de Palavras por Gênero



Para encerrar a exploração do dataset, foi aplicada uma análise com técnicas de inferência estatística para observar tendências e obter insights sobre os filmes a partir de projeções e testes de hipóteses sobre amostras como esta. O primeiro teste aplicado é o teste Omnibus (normaltest) para verificar se os atributos quantitativos possuem distribuição normal. Os resultados foram p-valores iguais a 0, que indicam que não há distribuição normal, o que é comprovado pelos valores de obliquidade e curtose, que são muito divergentes de 0, revelando a presença de assimetrias, picos e achatamentos na distribuição, principalmente no atributo ‘Gross’.

Outra ferramenta usada para comprovar que os atributos não seguem a distribuição Gaussiana é a plotagem de um gráfico QQ-Plot para cada um deles, que são apresentados na Figura 9 e não apresentam um score R2 que indique um ajuste perfeito, o atributo que mais se aproxima é ‘Meta_score’.

Figura 9: Heatmap de Importância de Palavras por Gênero



Posteriormente, foram executados testes de correlação usando os métodos de Spearman e Kendall nos pares de atributos do banco de dados. Definindo o nível de significância como 0.05, os testes fornecem evidências de que se deve aceitar a hipótese nula que defende a inexistência de uma verdadeira correlação entre os pares ‘Released_Year’ e ‘Runtime /min’, ‘IMDB_Rating’ e ‘Gross’, ‘Meta_score’ e ‘No_of_votes’, ‘Meta_score’ e ‘Gross’ e ‘Meta_score’ e ‘Runtime /min’. Os demais pares dos atributos quantitativos possuem p-valores abaixo do nível de significância definido, portanto, a hipótese nula é rejeitada e a correlação entre eles não é ao acaso.

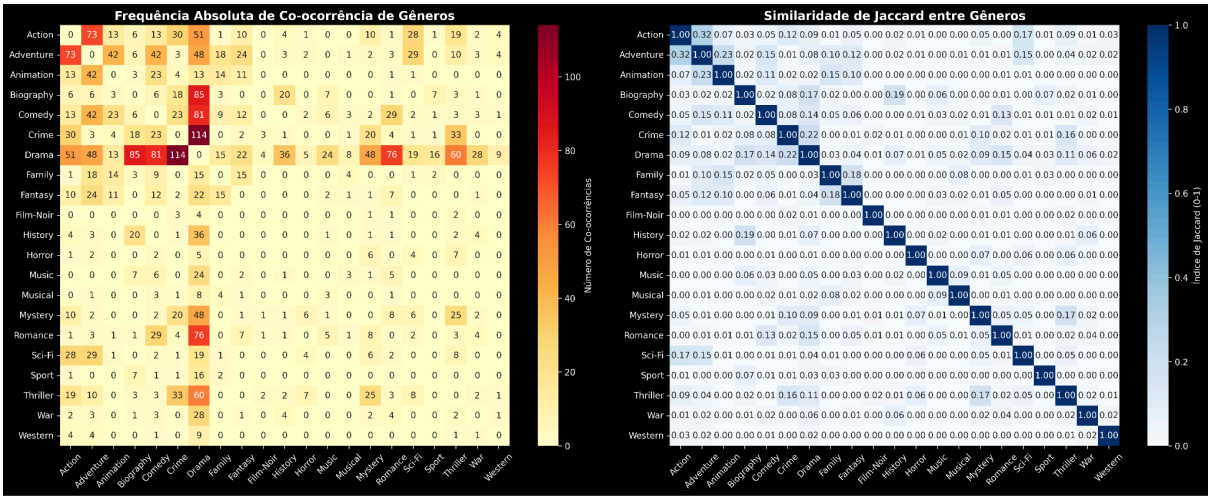
Para compreender as relações entre os diferentes gêneros cinematográficos, foi construída uma matriz de co-ocorrência que quantifica quantas vezes cada par de gêneros aparece conjuntamente nos filmes do dataset. A diagonal principal foi zerada para focar nas combinações entre gêneros distintos.

A matriz revelou combinações frequentes como Drama + Romance (76 filmes), Comedy + Drama (81 filmes) e Crime + Drama (114 filmes), indicando que o gênero Drama serve como base para diversas combinações.

Complementarmente à análise de co-ocorrência absoluta, calculou-se o Índice de Jaccard para medir a similaridade entre os gêneros. Esta métrica varia entre 0 e 1,

onde valores mais próximos de 1 indicam maior similaridade na co-ocorrência relativa.

Figura 10 e 11: Matriz de Co-ocorrência de Gêneros (Frequência Absoluta), Matriz de Similaridade de Jaccard entre Gêneros



Analisou-se as 15 combinações mais frequentes de gêneros (com pelo menos 2 gêneros) em relação ao seu desempenho comercial e de avaliação:

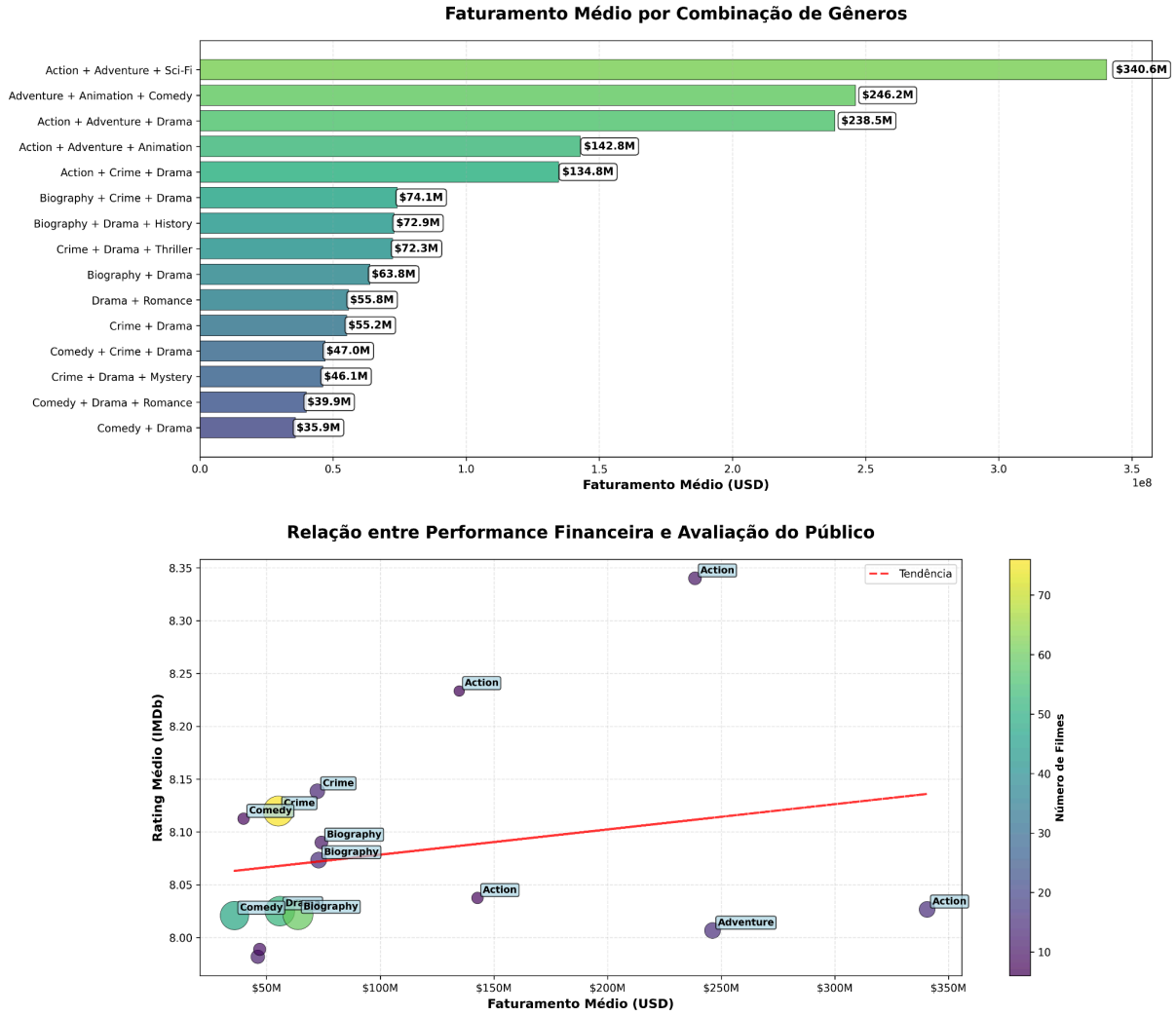
Tabela 1: Performance das Principais Combinações de Gêneros

Combinação	Nº Filmes	Faturamento Médio	Rating Médio
Action + Adventure + Sci-Fi	**15**	**$340.6M**	**8.03**
Adventure + Animation + Comedy	**15**	**$246.2M**	**8.01**
Action + Adventure + Animation	**8**	**$142.8M**	**8.04**
Action + Adventure + Drama	**10**	**$238.5M**	**8.34**
Action + Crime + Drama	**6**	**$134.8M**	**8.23**

As combinações envolvendo Action + Adventure + Sci-Fi demonstraram o maior faturamento médio ($340.6M), enquanto Action + Adventure + Drama obteve a melhor avaliação média (8.34).

Através de um gráfico de dispersão com tamanho de bolha proporcional ao número de filmes (Figura 13), foi possível visualizar a relação entre faturamento e avaliação do público para as diferentes combinações de gêneros.

Figura 12 e 13: Faturamento Médio por Combinação de Gêneros, Relação entre Performance Financeira e Avaliação do Público



Calculou-se métricas de potencial combinando performance financeira e crítica:
Top 5 Combinações por Potencial:

1. Action + Adventure + Drama: Potencial de $198.9M (10 filmes, rating 8.3)
2. Action + Crime + Drama: Potencial de $184.9M (6 filmes, rating 8.2)

3. Action + Adventure + Sci-Fi: Potencial de $182.3M (15 filmes, rating 8.0)
4. Action + Adventure + Animation: Potencial de $143.5M (8 filmes, rating 8.0)
5. Adventure + Animation + Comedy: Potencial de $131.4M (15 filmes, rating 8.0)

A análise individual por gênero revelou que:

Maior Faturamento Médio:
1. Adventure: $171.9M (162 filmes)
2. Sci-Fi: $161.1M (56 filmes)
3. Action: $156.6M (140 filmes)

Melhor Rating Médio:
1. Film-Noir: 8.05 (6 filmes)
2. Western: 8.04 (16 filmes)
3. Action + Adventure + Drama: 8.34 (10 filmes)

4 MODELAGEM PREDITIVA

O objetivo da modelagem preditiva foi desenvolver um sistema capaz de prever a nota IMDB de filmes com base em suas características intrínsecas. Trata-se de um problema de regressão, onde a variável target contínua é a IMDB_Rating.

Foram utilizadas 55 variáveis preditivas, organizadas em três categorias:

Variáveis Numéricas:

- Meta_score: Pontuação agregada das críticas
- No_of_Votes: Número de votos no IMDB
- Released_Year: Ano de lançamento
- Runtime /min: Duração em minutos
- Gross: Faturamento bruto

Variáveis de Gênero (One-Hot Encoding Multi-Label):

- 20 variáveis binárias representando cada gênero cinematográfico

Variáveis de Texto (TF-IDF):

- 20 features textuais extraídas das sinopses (Overview)

Os dados foram divididos em conjuntos de treino e teste na proporção 80/20, utilizando `random_state=42` para garantia de reprodutibilidade:

- Treino: 569 amostras
- Teste: 143 amostras

Como modelo baseline, foi implementada uma Regressão Linear, obtendo os seguintes resultados

Tabela 2: Performance da Regressão Linear

Métrica	Valor
MAE	0.1486
R^2	0.5617
MSE	0.0367
RMSE	0.1916

O modelo explicou 56.17% da variância da nota IMDB, com erro médio absoluto de 0.15 pontos.

Visando melhor performance, implementou-se um ensemble de Random Forest com os seguintes hiperparâmetros:

- `n_estimators=100`
- `max_depth=10`
- `min_samples_split=5`
- `random_state=42`

Tabela 3: Performance do Random Forest Regressor

Métrica	Valor
MAE	0.1470
R^2	0.5930
MSE	0.0341
RMSE	0.1846

O Random Forest superou a regressão linear, explicando 59.30% da variância e reduzindo o erro médio absoluto para 0.147 pontos.

O modelo Random Forest demonstrou performance superior:

Tabela 4: Comparativo entre RF e RL

Métrica	**Regressão Linear**	**Random Forest**	**Melhoria RF x RL**
MAE	0.1486	0.1470	-0.0016 (≈ -1.1%)
R^2	0.5617	0.5930	+0.0313 (≈ +5.6%)
MSE	0.0367	0.0341	-0.0026 (≈ -7.1%)
RMSE	0.1916	0.1846	-0.0070 (≈ -3.7%)

O Random Forest apresentou desempenho superior em todas as métricas, com destaque para o R^2 (+5.6%) e a redução do MSE (-7.1%)
A validação cruzada com 5 folds revelou instabilidade no modelo, com scores R^2 variando significativamente entre os folds. Esta variação sugere sensibilidade do modelo a particulares divisões dos dados.

Tabela 5: Top 10 Features por Importância

Feature	Importância
No_of_Votes	**0.5434**
Released_Year	**0.1180**
Meta_score	**0.1076**
Gross	**0.0857**
Runtime /min	**0.0434**
son	**0.0119**
family	**0.0077**
year	**0.0047**
Crime	**0.0044**
american	**0.0041**

- No_of_Votes: A variável mais importante (54.34%), confirmando a relação entre popularidade e avaliação
- Released_Year: Segunda mais importante (11.80%), indicando tendências temporais nas avaliações

- Meta_score: Terceira posição (10.76%), mostrando concordância entre crítica especializada e público
- Variáveis textuais: Palavras como "son", "family" e "year" mostraram poder preditivo significativo

Aplicou-se o modelo no filme "The Shawshank Redemption" (1994)

Resultado da Previsão:

- **Nota IMDB Prevista: 8.80**
- Nota IMDB Real: 9.30
- Erro Absoluto: 0.50 pontos

A Figura 14 faz uma dispersão dos valores previstos em função dos valores reais e apresenta uma linha de tendência que mostra onde deveriam estar distribuídas as previsões exatas

Figura 14: Previsões vs Valores Reais - Random Forest