INDICIUM TECNOLOGIA DE DADOS
PPRODUCTIONS

CAROLINA EMANUELE SANTOS DE ARAÚJO

DESAFIO CIENTISTA DE DADOS: ANÁLISE DE DADOS CINEMATOGRÁFICOS

SALVADOR - BAHIA
2024

CAROLINA EMANUELE SANTOS DE ARAÚJO

DESAFIO CIENTISTA DE DADOS: ANÁLISE DE DADOS CINEMATOGRÁFICOS

Relatório apresentado como requisito parcial para classificação no programa Lighthouse.

SALVADOR - BAHIA

2024

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O desenvolvimento de filmes envolve altos investimentos e envolvem a seleção minuciosa de múltiplas características, como gênero, diretor, atores e atrizes, escrita de uma descrição chamativa, classificação e tempo de duração, que podem determinar o faturamento e as notas das críticas e sobretudo do IMDB.

Nesse contexto, a aplicação de técnicas de ciência de dados e algoritmos de machine learning são fundamentais para ajudar os estúdios a determinarem o tipo de filme que deve ser lançado para obter um bom faturamento e boas críticas. No caso da PProductions, será utilizado um banco de dados cinematográfico para execução de uma análise exploratória envolvendo a aplicação de técnicas estatísticas, como análise de medidas e inferência estatística, para observar como os dados se comportam e o que é possível supor a partir deles.

Através da utilização dessas ferramentas é possível analisar quais são as principais características que determinam o sucesso de um filme e auxiliam na previsão de como serão os retornos com base nas características selecionadas. Dessa forma, a empresa PProductions poderá escolher quais filmes deveram ser lançados para que se obtenha uma maior margem de lucro e a possibilidade de obter boas notas das críticas e do IMDb.

2 DESENVOLVIMENTO

Nesse capítulo serão discutidos os objetivos da análise, os métodos aplicados em sua execução e os resultados obtidos.

2.1 OBJETIVOS

O objetivo dessa análise é explorar os dados cinematográficos a fim de obter insights e descobrir tendências que sejam relevantes para determinação das características que fazem com que um filme tenha um grande faturamento e ganhe boas notas dos críticos. Além disso, se destina a analisar quais algoritmos de machine learning podem ajudar a realizar previsões quanto aos retornos de um filme.

2.2 BANCO DE DADOS

O banco de dados contém 999 linhas e 14 colunas referentes a dados cinematográficos, com 286 linhas contendo valores nulos e nenhuma linha duplicada.

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.

- Star2: Ator/Atriz #2.
- Star3: Ator/Atriz #3.
- Star4: Ator/Atriz #4.
- No_of_Votes: Número de votos.
- Gross: Faturamento.

2.3 METODOLOGIA

Para desenvolver a análise foi criado um Jupyter notebook utilizando as bibliotecas:

- Pandas: Para manipulação do dataset.
- Matplotlib e Seaborn: Para visualização dos dados.
- Numpy: Para manipulação de dados numéricos.
- Scipy e pingouin: Para inferência estatística.
- Scikit Learn: Para preparação e implementação dos algoritmos de machine learning.
- Optuna: Para otimização dos hiperparâmetros dos modelos de machine learning.
- Pickle: Para salvar o modelo final de machine learning.

Foram utilizadas diversas técnicas de exploração dos dados, tratamentos, análise de medidas estatísticas, visualização gráfica, inferência estatística e aplicação de modelos de regressão com utilização de métricas e gráficos para mensurar o desempenho.

3 ANÁLISE EXPLORATÓRIA DOS DADOS

A análise exploratória dos dados (EDA) foi iniciada pela aplicação de técnicas de tratamentos de dados, de acordo com as primeiras informações obtidas. O primeiro passo foi excluir as 286 linhas contendo valores nulos e a execução de alterações nos valores dos atributos ‘Runtime’, ‘Released_Year’ e ‘Gross’ para passarem a ser colunas do tipo inteiro. Além disso, cada filme teve seus gêneros separados em uma lista, a fim de facilitar futuras manipulações. Também foi aplicada a técnica de vetorização usando o algoritmo da biblioteca scikit learn, TD-IDF, que extraiu em um vetor as 15 palavras com maior frequência relativa no atributo ‘Overview’, sendo elas “boy”, “family”, “father”, “help”, “life”, “love”, “man”, “new”, “old”, “son”, “story”, “war”, “woman”, “world” e “young”, com importâncias em alguns filmes descritas na ***Tabela 1***.

Tabela 1: Importância das palavras extraídas pelo TF-IDF

boy	family	father	help	life	love	man	new	old	son	story	war	woman	world	young
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,50	0,00	0,00	0,43	0,00	0,00	0,48	0,00	0,57	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,75	0,00	0,66	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

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 inconsonâncias nos dados, todos possuem valores válidos, não existe, por exemplo, um ano ou tempo de duração negativo.

Tabela 2: Descrição estatística

index	Released_Year	IMDB_Rating	Meta_score	No_of_Votes	Gross	Runtime (min)
count	712	712	712	712	712	712
mean	1995,738764	7,935674157	77,15449438	353466,2317	78450169,18	123,6671348
std	18,61118224	0,288928075	12,41811467	346450,1612	115068637,2	25,90760688
min	1930	7,6	28	25229	1305	72
25%	1986,75	7,7	69,75	95664,75	6143199	104
50%	2001	7,9	78	235981,5	34850145,5	120
75%	2010	8,1	86	506542,75	102360615	136
max	2019	9,2	100	2303232	936662225	238

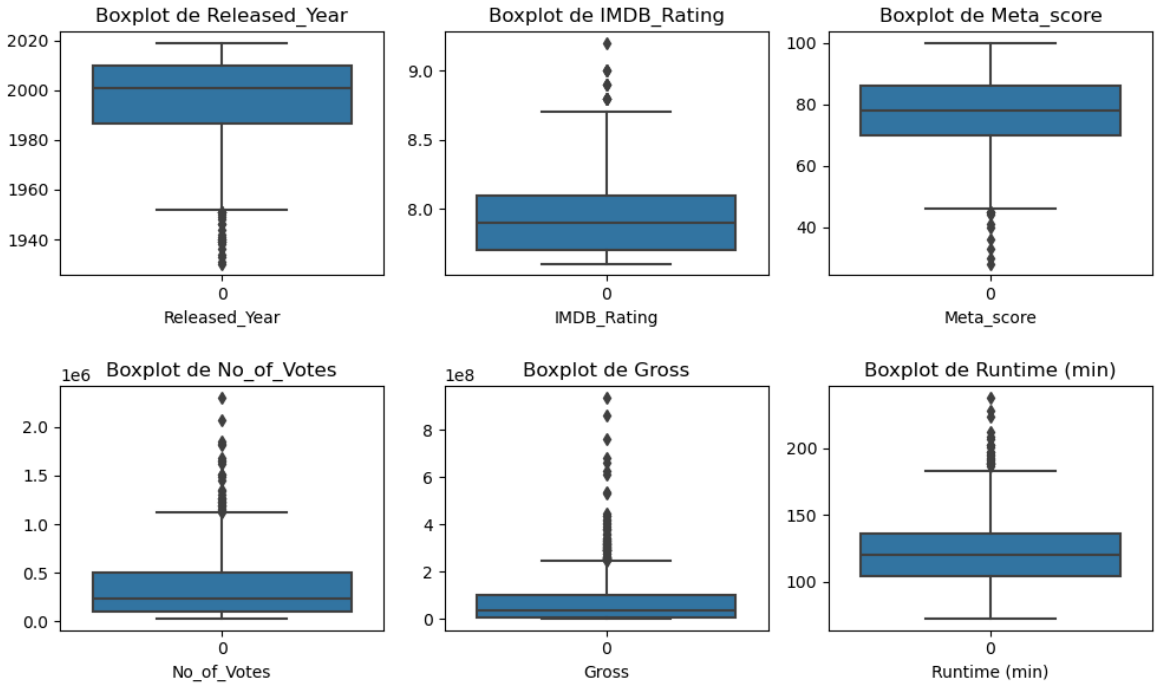
É possível observar que a metade dos filmes do dataset foram lançados entre 1930 e 2001, com média no ano de 1995. Já as notas do IMDb e das críticas possuem mediana em 7.9 e 78 e desvio padrão de, aproximadamente, 0.29 e 12.42, respectivamente. Quanto ao número de votos há uma média de 353466 com máximo de 2303232. Ademais, o faturamento mínimo foi de 1305 dólares e em média de 78450170 dólares. Por outro lado, o tempo de duração desse conjunto de filmes está entre 72 e 238 minutos, com terceiro quartil em 136 minutos.

Para os atributos qualitativos foram calculados as frequências e os percentuais de suas categorias. A classificação etária, por exemplo, teve frequência igual a 182 para categoria U, representado 25,56% dos dados. Os quatro gêneros mais frequentes são Drama, Adventure, Comedy e Crime, sendo que o primeiro categoriza 69,94% dos filmes do banco de dados, lembrando que cada filme pode ter mais de um gênero.

Quanto as palavras principais extraídas pelo método de vetorização, a que apresenta maior importância é ‘life’, com importância relativa de 10.99%. Já na coluna ‘Director’, foi possível observar que os três diretores que mais apareceram no dataset foram Steven Spielberg, em 13 filmes, Martin Scorsese, em 10 filmes e Alfred Hitchcock, em 9 filmes. Já os três atores com maior frequência nessa seleção cinematográfica, independentemente da importância (1, 2, 3 ou 4), foram Robert de Niro, em 16 filmes, Al Pacino e Tom Hanks, ambos em 13 filmes.

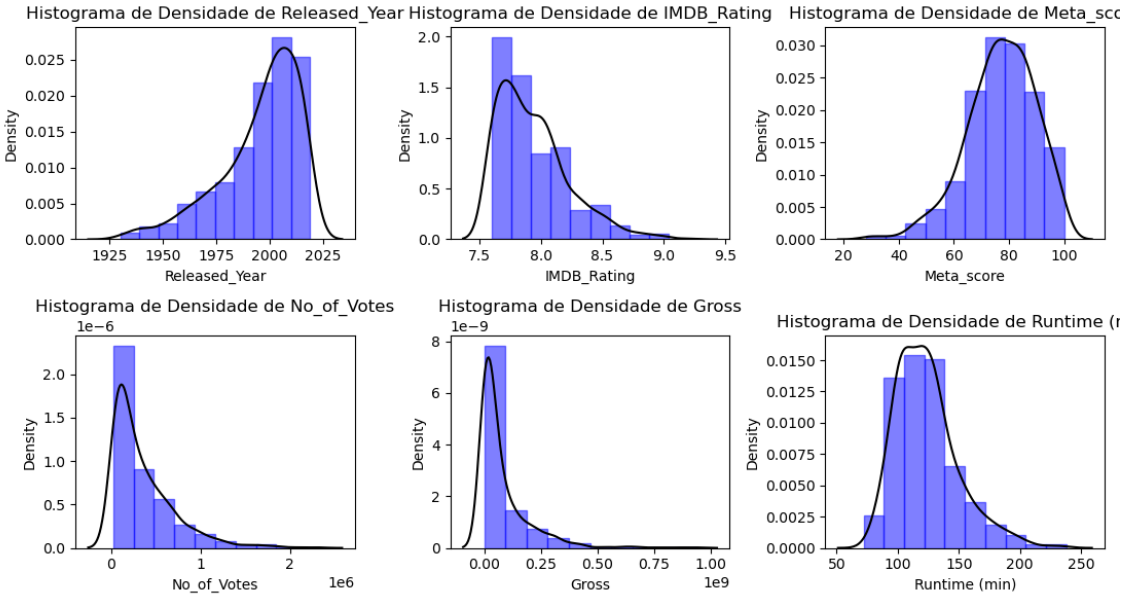
Posteriormente, foi executada uma análise gráfica do dataset, sendo que a primeira visualização foi dos boxplots dos atributos quantitativos, que permite observar a distribuição dos quartis, a assimetria de distribuição e a presença de outliers. O atributo ‘Gross’, por exemplo apresenta uma grande assimetria negativa, com a presença de muitos outliers, devido à grande variância dos valores de faturamento em um intervalo de mil a quase um bilhão de dólares.

Figura 1: Boxplots dos atributos quantitativos



Ainda em relação aos atributos numéricos, foram plotados histogramas para observar a distribuição de densidade de cada um deles, sendo possível observar visualmente que nenhum deles apresenta uma perfeita distribuição normal, todos eles possuem assimetrias.

Figura 2: Histogramas dos atributos quantitativos



Por outro lado, os atributos categóricos foram representados graficamente através de gráficos de barra que mostram a distribuição de suas frequências, possibilitando comparações mais dinâmicas.

Figura 3: Distribuição de 'Certificate'

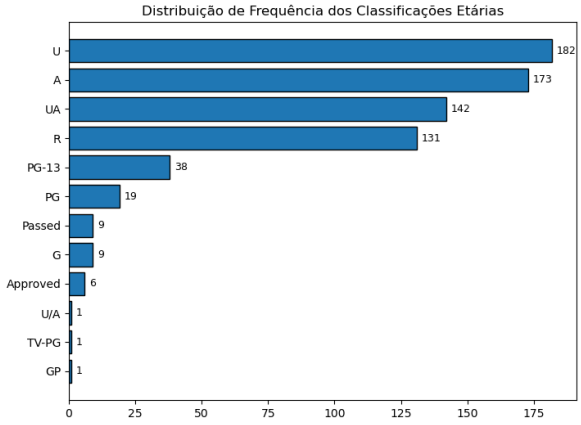


Figura 4: Distribuição de 'Genre'

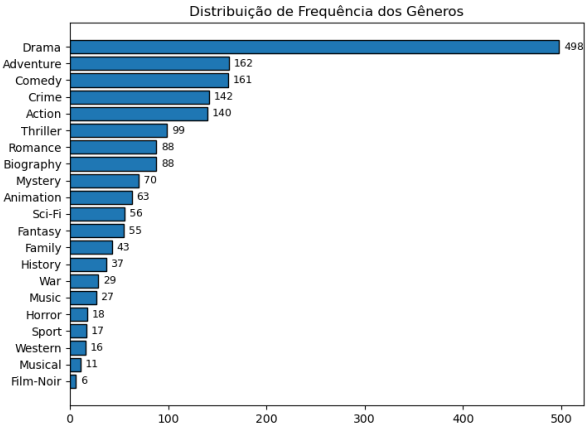


Figura 5: Distribuição das palavras importantes de 'Overview'

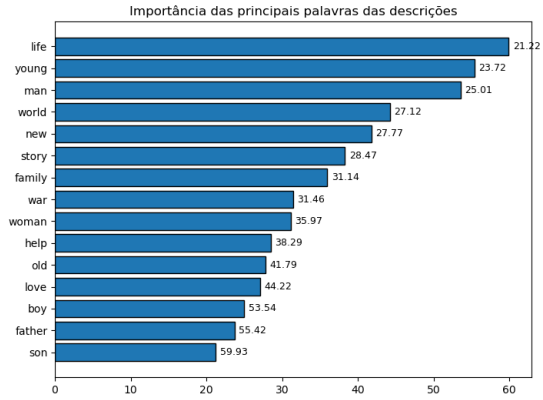


Figura 6: Distribuição de 'Director'

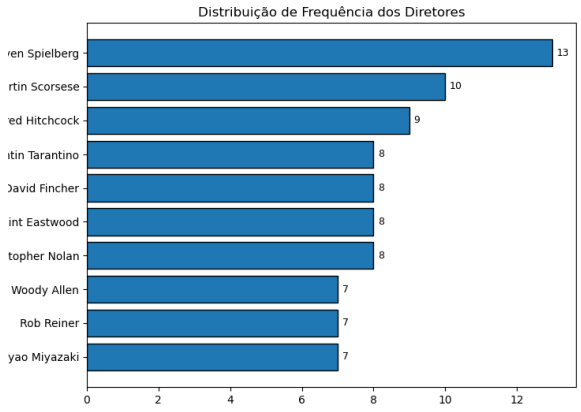
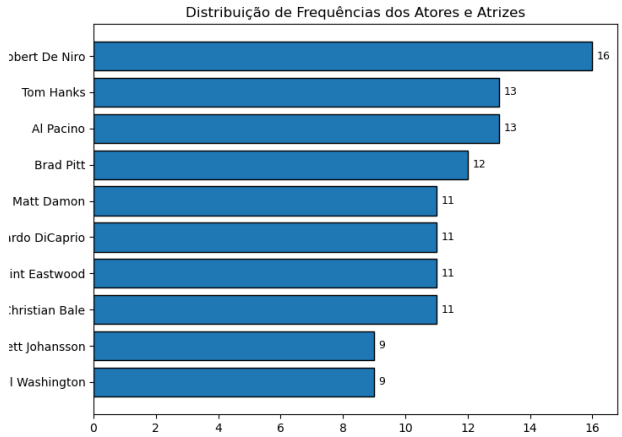


Figura 7: Distribuição de Atores/Atrizes



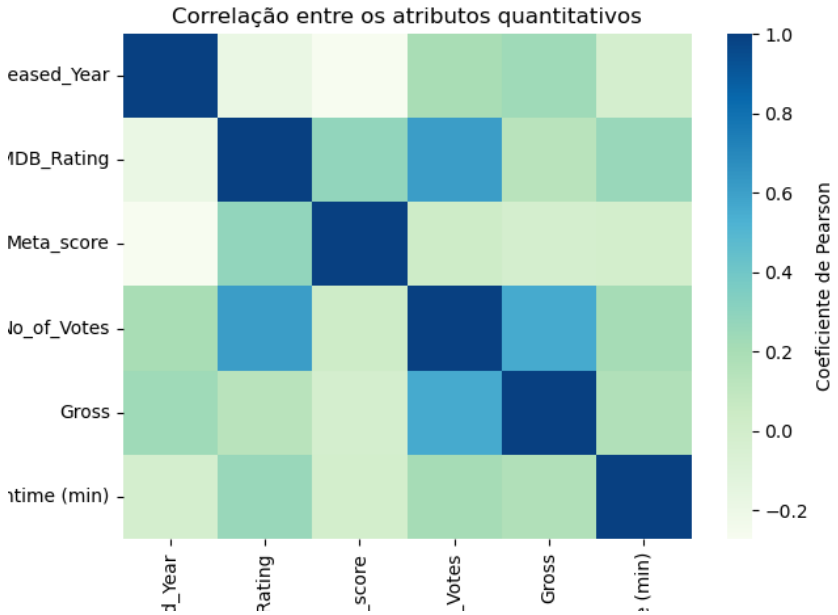
A próxima etapa foi realizar uma análise bivariada dos atributos, para observar a relação entre eles. O primeiro método aplicado foi o cálculo da correlação de

Pearson nas colunas numéricas. Essa técnica possibilitou a observação de que o faturamento dos filmes está atrelado ao número de votos, com uma correlação moderadamente forte e positiva, que pode ser justificada pelo aumento do número de votos devido a popularidade do filme, que implica em um maior faturamento para produtora. A nota da IMDb também possui uma correlação moderadamente forte com o número de votos, além de correções fracas, mais relevantes, com a média ponderada das críticas e com o tempo de duração da obra cinematográfica. As correlações podem ser observadas visualmente pelo heatmap na ***Figura 8***.

Tabela 3: Correlações de Pearson

index	Released_Year	IMDB_Rating	Meta_score	No_of_Votes	Gross	Runtime (min)
Released_Year	1	-0,178895873	-0,272659177	0,200073817	0,234654369	-0,017883573
IMDB_Rating	-0,178895873	1	0,283993826	0,609444391	0,132396424	0,258984858
Meta_score	-0,272659177	0,283993826	1	0,028575211	-0,014655858	-0,005938411
No_of_Votes	0,200073817	0,609444391	0,028575211	1	0,561532102	0,21321112
Gross	0,234654369	0,132396424	-0,014655858	0,561532102	1	0,168774962
Runtime (min)	-0,017883573	0,258984858	-0,005938411	0,21321112	0,168774962	1

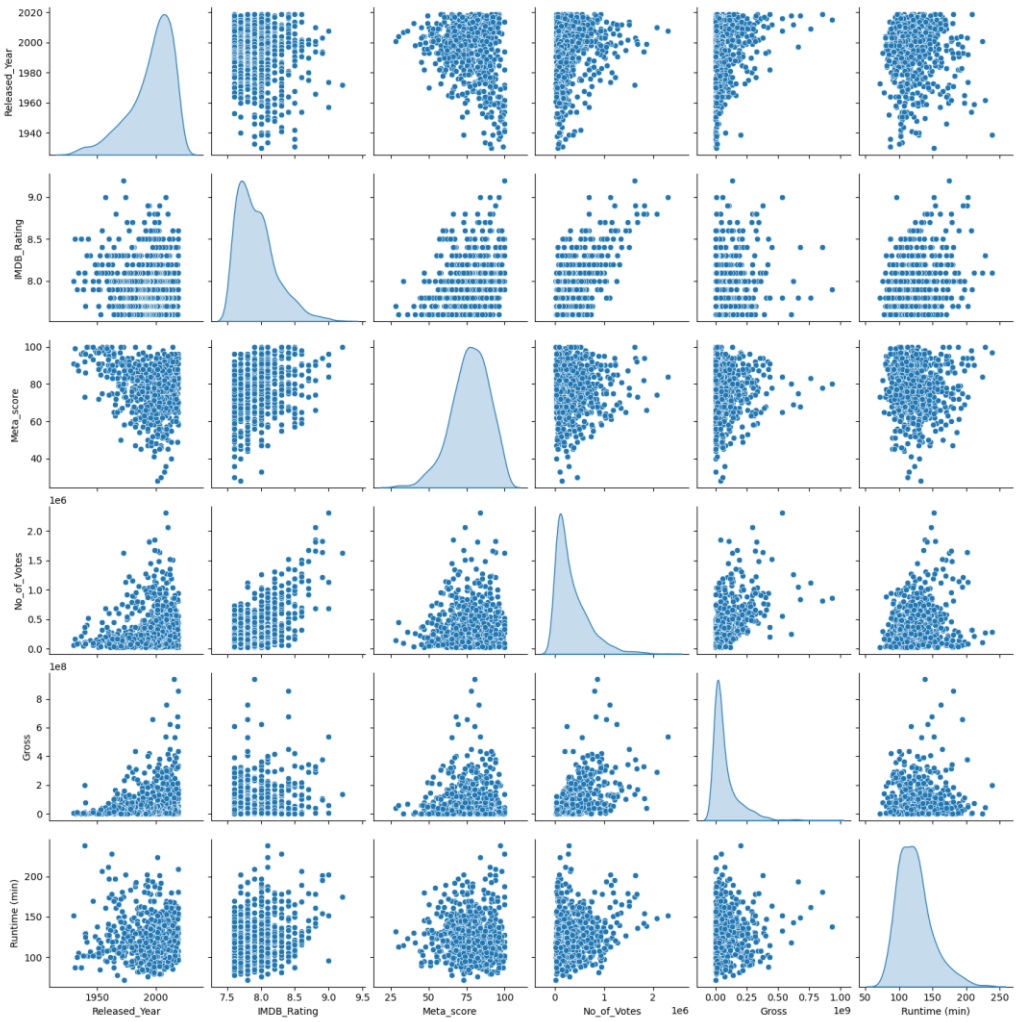
Figura 8: Heatmap de Correlações de Pearson



Além do heatmap, as relações entre os atributos também podem ser observadas na ***Figura 9***, que mostra como os dados se distribuem na associação de cada par de atributos. Essa é uma outra ferramenta para visualizar a presença de uma relação do número de votos com a nota IMDb ou com o faturamento, por exemplo.

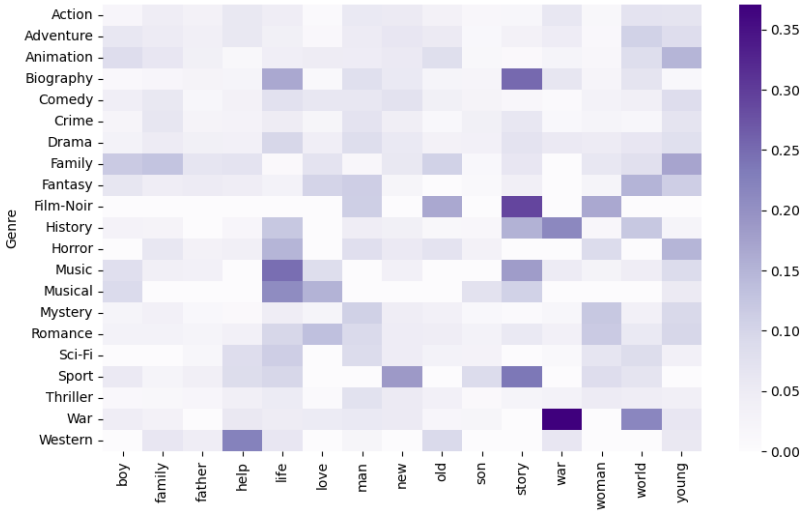
Porém, os gráficos do ano de lançamento em relação aos outros atributos são mais esparsos, mostrando que não apresenta correlações relevantes.

Figura 9: Gráficos de Dispersão



Ademais, também é possível fazer uma comparação entre as palavras extraídas pela vetorização da descrição e o gênero de cada filme. A partir disso, é possível obter a importância média que cada palavra do ‘Overview’ possui para cada categoria de gênero. É possível verificar a variação dessas médias de importâncias através da ***Figura 10***.

Figura 10: Heatmap 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, descritos na ***Tabela 5***, que são muito divergentes de 0, revelando a presença de assimetrias, picos e achatamentos na distribuição, principalmente no atributo ‘Gross’.

Tabela 4: Teste de Normalidade

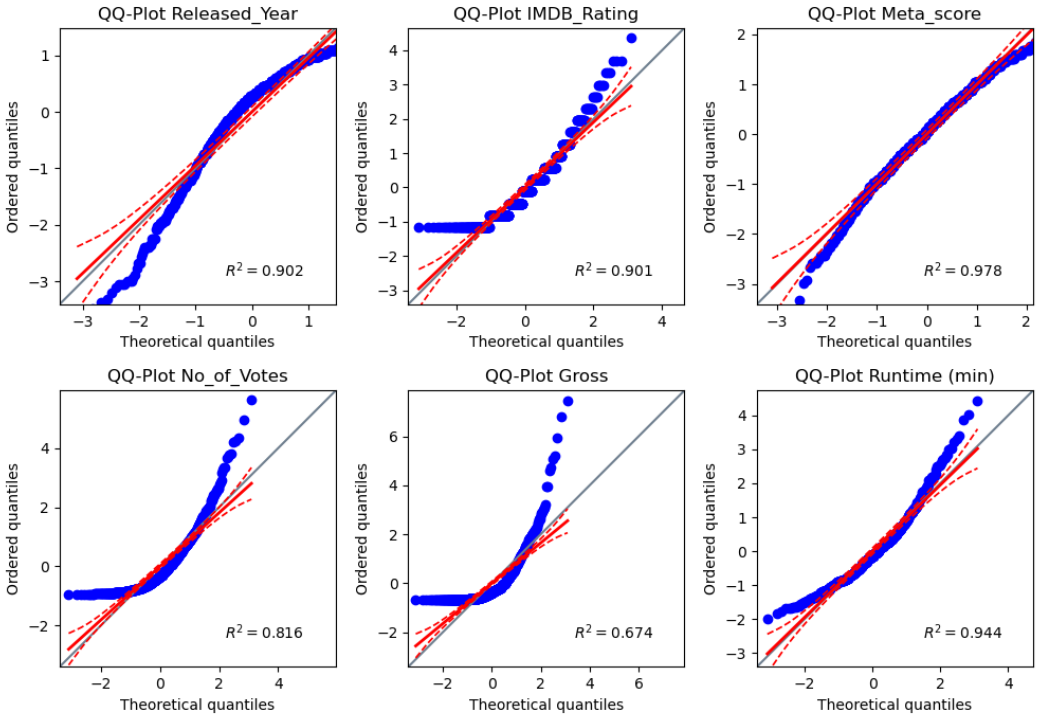
index	W	pval	normal
Released_Year	119,191	0	FALSO
IMDB_Rating	122,521	0	FALSO
Meta_score	40,898	0	FALSO
No_of_Votes	268,552	0	FALSO
Gross	475,325	0	FALSO
Runtime (min)	111,221	0	FALSO

Tabela 5: Obliquidade e Curtose

index	Obliquidade	Curtose
Released_Year	-1,146066074	0,926880607
IMDB_Rating	1,116874231	1,270137443
Meta_score	-0,584246868	0,483402581
No_of_Votes	1,820960302	4,228842476
Gross	2,92765589	12,22377233
Runtime (min)	1,014107679	1,375267038

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 11*** e não apresentam um score R2 que indique um ajuste perfeito, o atributo que mais se aproxima é 'Meta_score'.

Figura 11: Gráficos QQ-Plot



Posteriormente, foram executados teste 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. Os p-valores dos pares de atributos para cada teste podem ser visualizados na ***Tabela 6***.

Tabela 6: Testes de Correlação se Sperman e Kendall

X	Y	P-valor de Spearman	P-valor de Kendall
Released_Year	IMDB_Rating	0	0
Released_Year	Meta_score	0	0
Released_Year	No_of_Votes	0	0
Released_Year	Gross	0	0
Released_Year	Runtime (min)	0,189	0,179
IMDB_Rating	Meta_score	0	0
IMDB_Rating	No_of_Votes	0	0
IMDB_Rating	Gross	0,407	0,441
IMDB_Rating	Runtime (min)	0	0
Meta_score	No_of_Votes	0,745	0,73
Meta_score	Gross	0,07	0,069
Meta_score	Runtime (min)	0,114	0,107
No_of_Votes	Gross	0	0
No_of_Votes	Runtime (min)	0	0
Gross	Runtime (min)	0	0

Por fim, foram feitas suposições sobre os dados que foram testadas usando o teste não-paramétrico de Mann-Whitney. A primeira suposição é sobre a existência de uma diferença estatística do faturamento dos filmes lançados antes e depois de 1995, que é o ano médio. A execução do teste supracitado teve como resultado uma estatística U igual a 53734 e um p-valor de 0.013, que é inferior ao nível de significância, comprovando a existência de uma heterogeneidade nos valores de faturamento: com o passar dos anos esse valor se tornou cerca de 1 milhão de vezes maior. Essa elevação do valor de faturamento pode ser observada no gráfico de dispersão desses atributos disponível na ***Figura 9***.

Além disso, os faturamentos também foram comparados em amostras dos anos 2015 e 2019. Nesse caso, o valor da estatística U foi 141 e o p-valor 0.842, fornecendo evidências de que a hipótese nula deve ser aceita e que o faturamento de 2015 e 2019 possuem uma distribuição semelhante, com medianas aproximadamente iguais. Essas diferenças estão associadas ao aumento do número de espectadores ao longo do tempo, devido a fatores como uma maior popularização do cinema e a criação de serviços de streamings, por exemplo, que fizeram com que os lucros crescessem exponencialmente.

Outra comparação foi feita em relação a uma amostra do primeiro quartil de faturamento e uma amostra dos dados superiores ao terceiro quartil para observar como se distribuem o número de votos em cada uma delas. O resultado foi de uma

estatística U igual a 1860 e um p-valor igual a 0, que indica que a probabilidade de observar uma uniformidade de distribuição entre as amostras, o que prevê a hipótese nula, é aproximadamente zero.

Portanto, nesse caso a hipótese alternativa é aceita indicando que a mediana do número de votos para os 25% dos valores inferiores de faturamento é completamente diferente para os valores mais altos que estão acima dos 75%. Isso pode ser justificado pelo fato de que o número de votos de um filme está atrelado a sua popularidade, alcance e repercussão, que elevam o número de espectadores e, consequentemente, o número de votos. Vale ressaltar que o atributo ‘No_of_Votes’ está fortemente relacionado com o atributo ‘IMDB_Ratinng’, o que prova a associação entre popularidade e número de votos.

3.1 DISCUSSÃO SOBRE OS RESULTADOS

A partir dessa análise exploratória dos dados é possível responder questionamentos como qual deve ser o filme indicado a uma pessoa desconhecida, que seria um filme com boas notas das críticas e do IMDb e bom número de votos, que indicam a popularidade e sucesso de um filme. Sendo assim, usando medidas estatísticas é possível selecionar 10 filmes que tem alta probabilidade de serem muito populares.

Eles foram extraídos do banco de dados através da utilização de um filtro para filmes com número de votos, média ponderadas das críticas e nota do IMDb acima dos seus respectivos terceiros quartis, sendo sugeridos aqueles que possuem os 10 maiores faturamentos dentro dessa seleção. Os filmes são: ‘The Lion King’, ‘Toy Story 3’, ‘The Lord of the Rings: The Return of the King’, ‘The Lord of the Rings: The Two Towers’, ‘Star Wars’, ‘The Lord of the Rings: The Fellowship of the Ring’, ‘Up’, ‘WALL-E’, ‘Saving Private Ryan’ e ‘Back to the Future’.

Essa filtragem foi escolhida porque esses atributos estão fortemente associados a popularidade de um filme, sendo que alguns deles também estão atrelados a expectativa de um alto faturamento para um filme. Isso pode ser comprovado pelo teste de inferência estatística de Mann-Whitney provou a existência de uma relação entre o aumento o faturamento da popularidade dos filmes, associada

ao número de votos e por outro lado a popularização das obras cinematográficas com o passar dos anos. A análise da correlação também confirma essa hipótese, indicando uma correlação positiva moderadamente forte entre o faturamento e o número de votos, com fortes evidências de que sua associação não é ao acaso, assim como sua associação com o ano de lançamento do filme, com o qual possui uma correlação positiva fraca.

Outro importante resultado obtido nessa análise é a extração das 15 palavras mais importantes do ‘Overview’ de um filme, obtidas através de uma vetorização feita a partir da multiplicação entre a frequência de uma palavra em um ‘overview’ de um filme pela frequência inversa dessa palavra em todo as descrições dos filmes do banco de dados. As 15 palavras mostram que os filmes dessa amostra têm uma tendência em abordar temas relacionados a vida, a mundos de realidades paralelas ou situações que possam afetar o mundo, a guerras, a histórias de família, especialmente pais e filhos, a histórias sobre ajuda e empatia e histórias de amor.

É possível obter a partir da média, uma medida estatística, a importâncias que geralmente cada palavra tem para filmes de determinados gêneros. A partir disso, é possível associar palavras a gêneros ao selecionar aquelas que possuem o maior valor de importância média. Sendo assim a palavra ‘world’, é a principal palavra dos gêneros Action, Adventure e Fantasy. Já ‘young’ está associada principalmente a Animation, Comedy e Family. ‘story’ possui a maior importância média dos gêneros Biography, Film-Noir e Sport. ‘man’ está atrelado as categorias Crime e Thriller. Por outro lado, Drama, Horror, Music, Musical e Sci-Fi se associam com mais a palavra ‘life’. ‘war’ se associa a History e War. Por fim, os gêneros Mystery, Romance e Western tem como palavras de maior importância ‘woman’, ‘love’ e ‘help’, respectivamente.

As tendências e insights obtidas a partir dessa análise exploratória são fundamentais para compreender como um filme é recebido pelo público e os fatores que influenciam seu faturamento. Além disso, essas características possibilitam realizar predições sobre a nota que ele receberá pelo IMDb, por exemplo, como será discutido no próximo tópico.

4 MODELAGEM PREDITIVA

Técnicas de machine learning podem ser utilizadas para prever a nota do IMDb de um filme a partir de suas características. Nesse caso, serão utilizados os atributos ‘Released_Year’, ‘Certificate’, ‘Meta_score’, ‘Director’, 'Star1', 'Star2', 'Star3', 'Star4', 'No_of_Votes', 'Gross', 'Runtime (min)'. Além disso foram criadas colunas de classificação dos gêneros para Drama, Adventure, Comedy, Crime e outros, sendo elas 'GenreDrama', 'GenreAdventure', 'GenreComedy', 'GenreCrime', 'GenreOthers'. As colunas das palavras extraídas pela técnica TF-IDF, 'boy', 'family', 'father', 'help', 'life', 'love', 'man', 'new', 'old', 'son', 'story', 'war', 'woman', 'world', 'young', também serão utilizadas.

A seleção dessas colunas tem como objetivo reunir as principais características do filme a fim de prever as notas do IMDB, ressaltando que o número de votos se associa a popularidade do filme, os diretores, atores e atrizes podem chamar atenção do público, as palavras do overview podem cativar mais espectadores, o faturamento pode ser um indicativo da boa repercussão do filme. O data frame resultante foi separado em atributos previsores e atributo preditivo (‘IMDB_Rating’) e manipulado com a utilização do LabelEncoder para codificar as colunas categóricas e StandardScaler para padronização das colunas numéricas, mantendo sem alterações as colunas categóricas binárias, como as colunas de classificação dos gêneros.

Os algoritmos de machine learning testados para essa previsão são de regressão, tendo em vista que o atributo alvo é um dado contínuo. Inicialmente, os dados são divididos em conjunto de treino, validação e teste, sendo que o segundo será utilizado para implementar uma validação cruzada para obter o desempenho médio dos modelos Linear Regression, Lasso, Ridge, Random Forest Regressor, Multi Layer Perceptron Regressor, Decision Tree Regressor, Support Vector Machinhe Regressor, Gradient Boosting Regressor e Ada Boost Regressor.

Para isso o conjunto de validação foi divido por um KFold de 10 splits e foi usado como score o valor negativo da raiz do erro quadrático médio (neg_root_mean_squared_error), porque esse algoritmo procura maximizar o score, que nesse caso é o negativo de uma penalização quadrática, mas com a métrica na mesma unidade das notas reais. A ***Tabela 7*** faz uma comparação com o score de cada modelo testado e o desvio padrão entre os resultados.

Tabela 7: Comparação entre os modelos

Modelo	NRMSE Médio	Desvio Padrão
Random Forest	-0,696413819	0,121451971
AdaBoost	-0,713045963	0,112615303
Gradient Boosting	-0,742817617	0,122884393
Ridge	-0,758289679	0,089349699
Regressão Linear	-0,786326437	0,097646968
Árvore de Decisão	-0,902873599	0,253239419
Lasso	-1,000047922	0,228977162
SVM	-1,0245275	0,229788707
MLP	-64,89742225	10,82065414

A partir desses resultados foram selecionados os 4 modelos com melhor desempenho: Ridge, Linear Regreession, Random Forest e Gradient Boosting. Os hiperparâmetros desses modelos são otimizados pelo algoritmo Optuna, que sugere valores para eles dentro de intervalos pré-definidos usando a otimização bayesiana, sendo que para cada um deles foram executados 50 testes.

Com os hiperparâmetros ajustados, os modelos foram combinados utilizando a técnica stacking, que os treina separadamente e combina suas previsões individuais a partir de um meta modelo, que nesse caso é o Gradient Boosting, para gerar uma previsão final. Apesar de aumentar a complexidade e o custo computacional, o Stacking Regressor é um bom modelo para executar as previsões já que melhora o desempenho da predição, usando diferentes técnicas de regressão combinadas, como mínimos quadrados, regularização e ensemble. Além disso, também reduz a possibilidade de ocorrer sobreajuste nos dados de treinamento, o overfitting, que seria maior nas predições individuais.

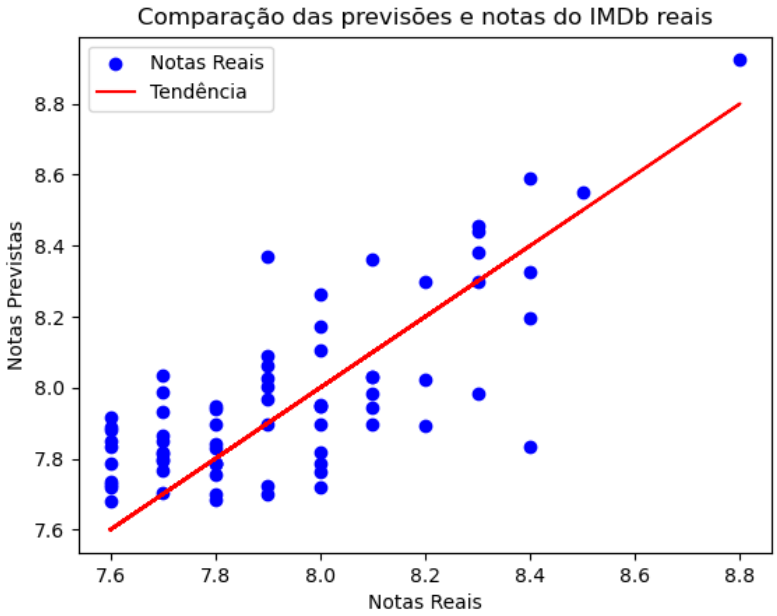
Depois de treinar o modelo Stacking Regressor no conjunto de dados de treino, ele é utilizado para prever a nota do IMDb do conjunto de teste. Na ***Tabela 8***, é possível estabelecer uma comparação entre os valores reais e os valores previstos de 10 filmes desse conjunto.

Tabela 8: Comparação entre valores reais e previstos

Titulo	Valor Real	Valor Previsto
Star Trek Into Darkness	7,7	7,988505522
Kaze tachinu	7,8	7,784875096
Gully Boy	8	7,763132574
The Incredibles	8	7,952399838
Cast Away	7,8	7,830506163
Todo sobre mi madre	7,8	7,938864906
Darbareye Elly	8	7,784875096
Blade Runner 2049	8	7,947370024
Amadeus	8,3	7,983348738
The Insider	7,8	7,89831653

Também é possível compará-los por meio da ***Figura 12*** que 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.

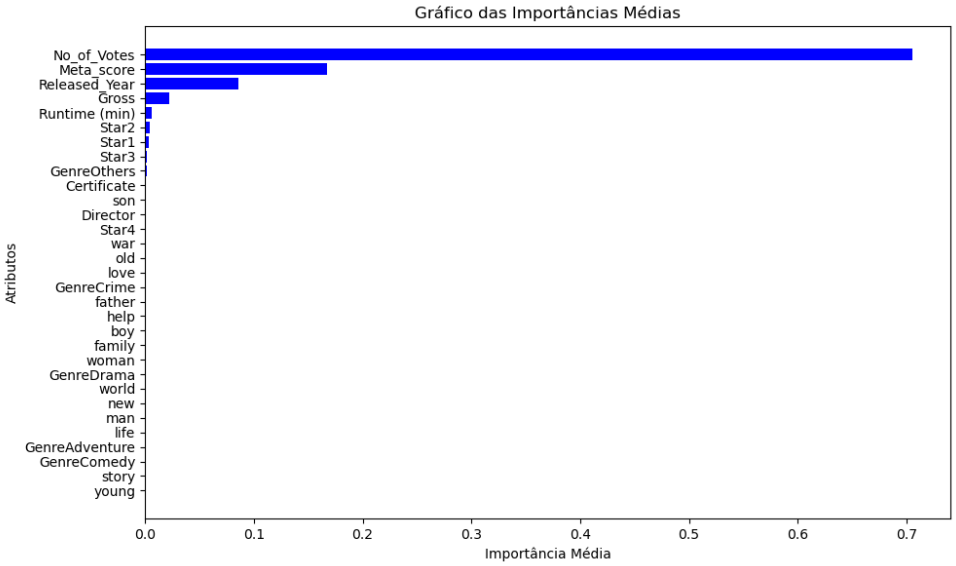
Figura 12: Gráfico de comparação entre os valores reais e previstos



Para avaliar mais precisamente o desempenho do modelo foi feita uma avaliação utilizando as métricas Mean Squared Error (Erro quadrático médio) e Mean Absolute Error (Mean Absolute Error), sendo que a primeira possui uma penalização quadrática dos erros e a segunda mostra o erro absoluto na mesma escala das notas. O erro quadrático médio entre os valores reais e previstos das notas dos filmes do conjunto de teste foi de aproximadamente 0.03, já o erro absoluto médio teve um resultado aproximado de 0.15.

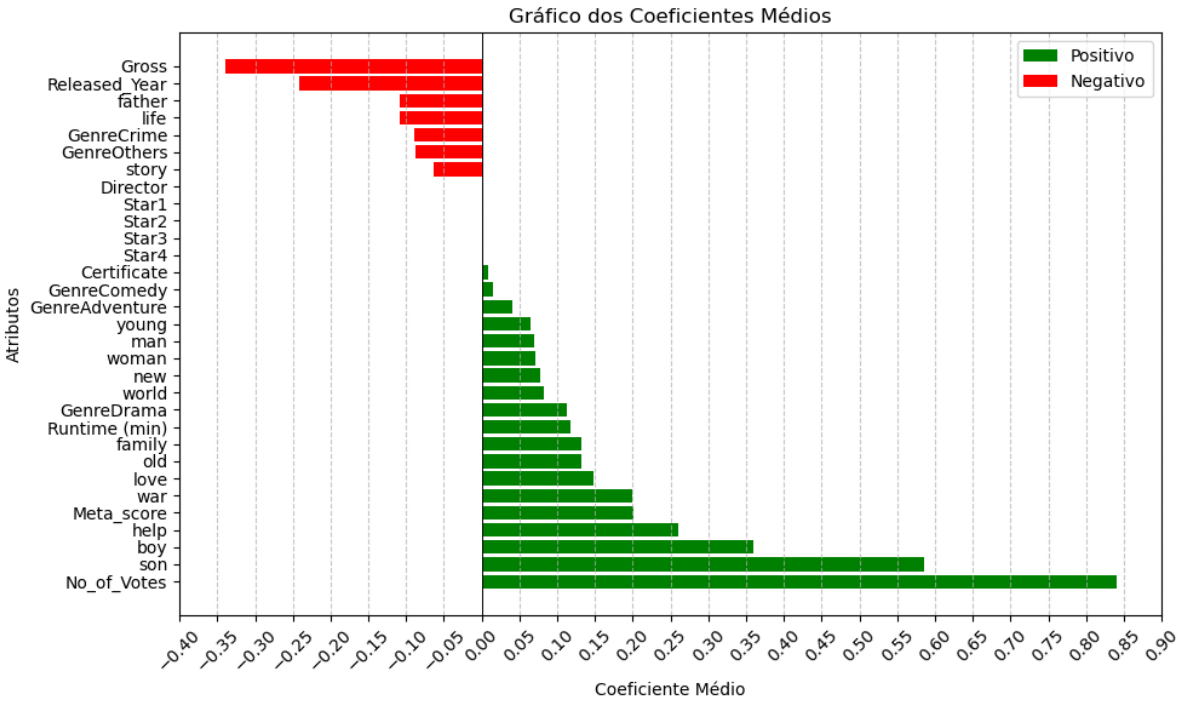
Vale ressaltar que além de serem utilizados para executar predições, os modelos de machine learning também podem ser utilizados para determinar a importância de cada atributo em relação ao atributo alvo e para obter coeficientes que relacionem eles. A partir da média das importâncias obtidas pelos modelos Gradient Boosting e Random Forest, foi plotado a ***Figura 13*** para compará-las entre os atributos.

Figura 13: Importâncias dos atributos



Já os modelos Ridge e Linear Regression fornecem coeficientes dos quais foram obtidas médias para comparar a relação entre cada atributo e o atributo alvo com a ***Figura 14***.

Figura 14: Coeficientes dos atributos



A partir desses resultados, é possível observar que o número de votos tem a maior importância para predição da nota do IMDb, além de possuir uma forte correlação direta com ela, ou seja, quanto maio o número de votos, maior será essa nota. As outras características mais importantes são ano de lançamento, a média ponderada das críticas, o faturamento, os dois atores principais e o tempo de duração. Os demais atributos possuem importância aproximadamente igual a 0.

Quanto aos coeficientes obtidos, é possível observar que quanto mais recente o filme, sua nota será moderadamente menor, o que pode ser justificado pela maior exigência e análise de mais características na atribuição da nota. Também é possível inferir que a utilização de palavras como ‘father’, ‘life’ e ‘story’ não sejam tão cativantes ou estejam atrelados a filmes com avaliação um pouco menor, assim como o gênero Crime.

Por outro lado, os diretores, atores e atrizes não causam um efeito muito significativo a nota, apesar de terem uma importância superior do que a maioria dos atributos. Já média das críticas e a importância das palavras como ‘old’, ‘war’, ‘help’, ‘love’, ‘boy’ e ‘son’ no overview estão associadas a um efeito positivo na nota do IMDb, sobretudo a importância da palavra ‘son’ obtida pelo algoritmo TD-IDF.

Para executar predições sobre a nota do IMDb de novos filmes, é necessário organizas suas características em um data frame e executar a função

tratar_novos_dados(), criada para aplicar os tratamento necessários para que sejam alterados os tipos de algumas colunas, para que as colunas de categorias de gêneros sejam criadas, para vetorizar a importância das palavras que já foram extraídas pelo método TF-IDF e realizar a padronização e codificação dos dados para implementar o modelo e obter as novas predições.

Ao executar a previsão da nota de um novo filme com as características descritas na ***Tabela 9***, a nota do IMBD resultante foi de aproximadamente 9.0.

Tabela 9: Novos dados

Series_Title	Released_Year	Certificate	Runtime	Genre	Overview	Meta_score	Director	Star1	Star2	Star3	Star4	No_of_Vot	Gross
The Shawshank Redemption	1994	A	142 min	Drama	Two imprisoned men bond over a number of years, finding solace and eventual redemption through acts of common decency.	80	Frank Darabont	Tim Robbins	Morgan Freeman	Bob Gunton	William Sadler	2343110	28,341,469

REFERÊNCIAS BIBLIOGRÁFICAS

MATPLOTLIB. **Matplotlib: Visualization with Python**. Disponível em: <https://matplotlib.org/>. Acesso em: 03 de jul. de 2024.

SEABORN. **Seaborn: Statistical Data Visualization**. Disponível em: <https://seaborn.pydata.org/>. Acesso em: 03 de jul. de 2024.

SCIKIT-LEARN. **Scikit-learn: Machine Learning in Python**. Disponível em: <https://scikit-learn.org/stable/index.html>. Acesso em: 04 de jul. de 2024.

PINGOUIN. **Pingouin: Statistical Package in Python**. Disponível em: <https://pingouin-stats.org/build/html/index.html>. Acesso em: 06 de jul. de 2024.

NADISPERSA. **Inferência Estatística**. Medium, 2024. Disponível em: <https://nadispersa.medium.com/infer%C3%AAncia-estat%C3%ADstica-59e67d0d70a8>. Acesso em: 06 de jul. de 2024.