**IMDB MOVIES DATASET**

Leonardo Lemes Pedrozo

Leonardo Lemes Pedrozo

Acadêmico do Curso de Engenharia de Software do Centro Universitário Cesumar – UNICESUMAR, Curitiba – PR. leo.pedrozo333@gmail.com

**Resumo**

Neste trabalho, explora-se e prevê-se as tendências no conjunto de dados “IMDB Movies Dataset”, fundamentado em estatísticas descritivas. a intenção é determinar os principais modelos e os principais padrões associados entre características críticas como gênero, investimento, bilheteria, média de nota e como popular ou não é o filme. O método traz entradas nos casos de auto edição de dados, treinamento de opacidade de dados, discrição de distribuição, identificação de correlações e treinamento de modelos supervisionados a fim de prever a nota dos filmes.

Os dados analisados são detalhes sobre mil filmes e series de TV, o que reforça a base para investigação sólida. Exemplo de exibição gráfica e métricas estatísticas foram executados. Notou-se que vários filmes tiveram avaliações em torno de 6 e 7 pontos, enquanto valores extremos são raros.

Anotações de pesquisa indicou uma notável ligação entre variáveis como "vote\_count" e "popularity" com a média de votos, nesta perspectiva além disso, modelos de Regressão Linear e Decision Tree foram avaliados para predição da nota, com desempenho satisfatório. O trabalho mostra a análise preditiva para o ramo da indústria do cinema, deu uma impressão valiosa para produtores, distribuidores, plataformas de streaming e profissionais de marketing que desejam entender melhor os fatores que influenciam o sucesso de um filme.

**PALAVRAS-CHAVE:** Análise Exploratória; IMDB; Machine Learning;

## INTRODUÇÃO

O mundo do cinema gera uma receita bilionária todos os anos, sofrendo impactos de vários aspectos, que incluem desde os atores e atrizes até estratégias de mercado, como o dinheiro disponível e a forma como o filme chega ao público. O surgimento de instrumentos para examinar informações permite obter aprendizado de bancos de dados como o IMDB (Internet Movie Database), um dos maiores locais de informações sobre filmes, séries e programas de TV. Com a evolução da ciência de dados, ficou mais fácil fazer estudos detalhados e com previsão, buscando entender o que faz um filme ser bem visto ou famoso.

Este estudo busca usar métodos de análise detalhada com base no relatório produzido, investigando números descritivos, distribuição de valores, ligações entre dados e identificar quais características afetam mais a avaliação dos filmes. Depois, são utilizados modelos supervisionados de aprendizado de máquina, como regressão linear e árvore de decisão, para adivinhar a nota de um filme com base em elementos como orçamento, receita, quantidade de votos e tempo de duração.

A forma de trabalho inclui fases bem definidas: desde o recebimento da base de dados, arrumação e tratamento dos dados, análise por gráficos e estatísticas das características até a criação de modelos de aprendizado de máquina para avaliar o quão bem ele consegue prever.

O estudo está dividido em quatro partes importantes: análise do perfil estatístico, tratamento dos dados, aplicação de modelos de previsão e análise dos resultados. Todas as etapas foram realizadas com base nos dados reais retirados do IMDB Movies Dataset, garantindo a certeza e a importância das conclusões mostradas.

**1 DESENVOLVIMENTO**

O projeto foi desenvolvido utilizando um pipeline com arquitetura hexagonal, orientado por portas e adaptadores. O código principal é implementado em Python e conta com os seguintes adaptadores: KaggleDownloaderAdapter, YDataProfilingAdapter, DtaleAdapter e PyCaretAdapter. A orquestração dos casos de uso é feita pela classe MLUseCases, permitindo acoplamento flexível e modularidade entre os componentes.

Esse relatório forneceu estatísticas como média, desvio padrão, skewness e kurtosis das colunas, além de diagramas de correlação e detecção de outliers, fundamentais para as decisões de engenharia de dados.

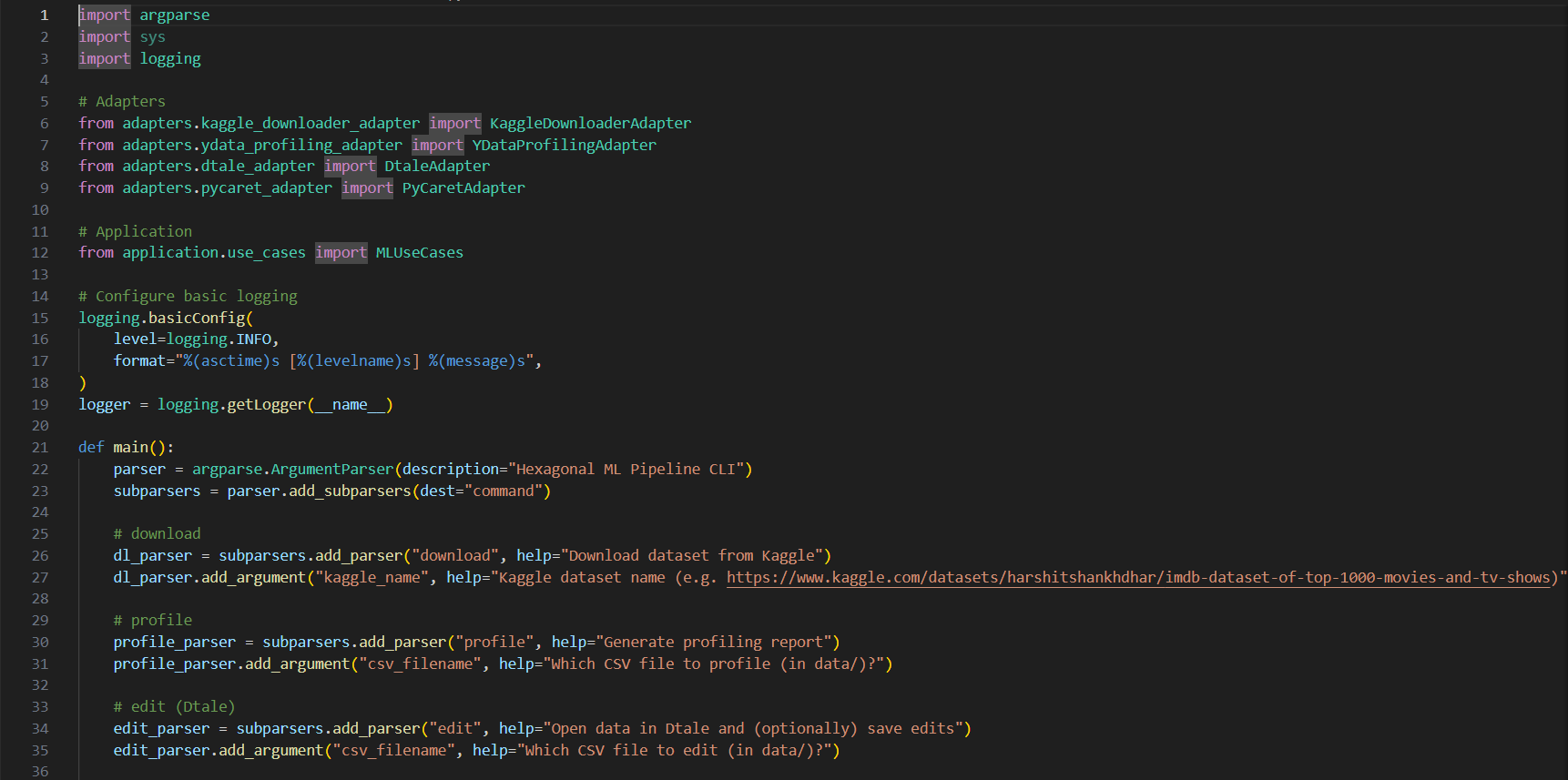
Com os dados carregados, o relatório indicou a presença de valores nulos em budget, revenue e runtime, que foram tratados com remoção de registros. Utilizou-se StandardScaler para normalizar colunas numéricas e codificação one-hot para colunas categóricas.

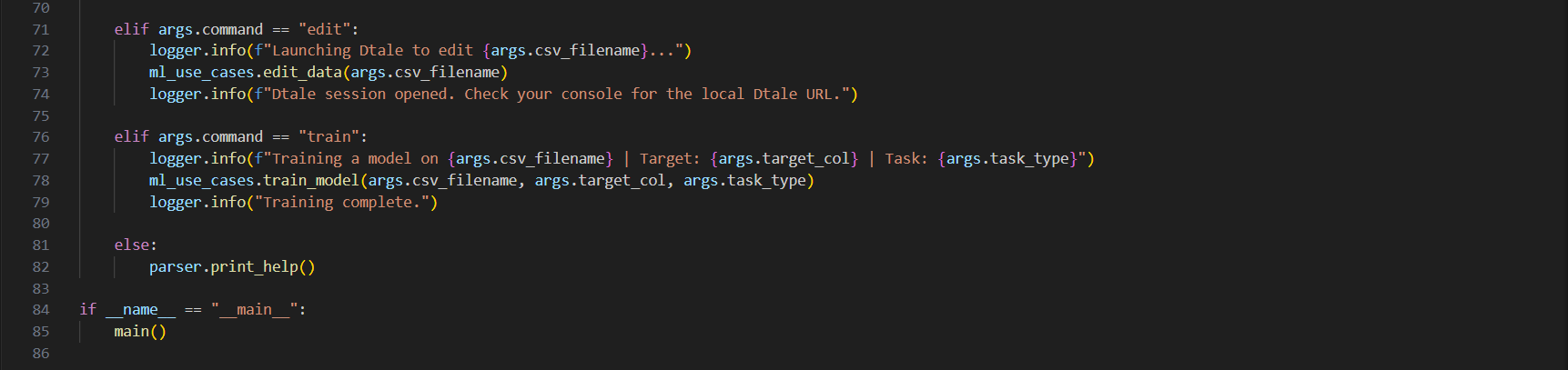
Os dados foram divididos automaticamente em treino e teste. Avaliaram-se múltiplos algoritmos, sendo a Regressão Linear e a Árvore de Decisão os escolhidos por seu desempenho e interpretabilidade.

A árvore de decisão teve melhor desempenho. A importância dos atributos destacou vote\_count, popularity e revenue como os fatores mais impactantes.

A integração entre os componentes do pipeline permitiu a automação completa do fluxo de análise, desde a coleta até a modelagem. O uso do YData Profiling agilizou a compreensão do dataset, e o PyCaret facilitou os testes com múltiplos algoritmos. O pipeline hexagonal contribuiu para a organização e extensibilidade do projeto.

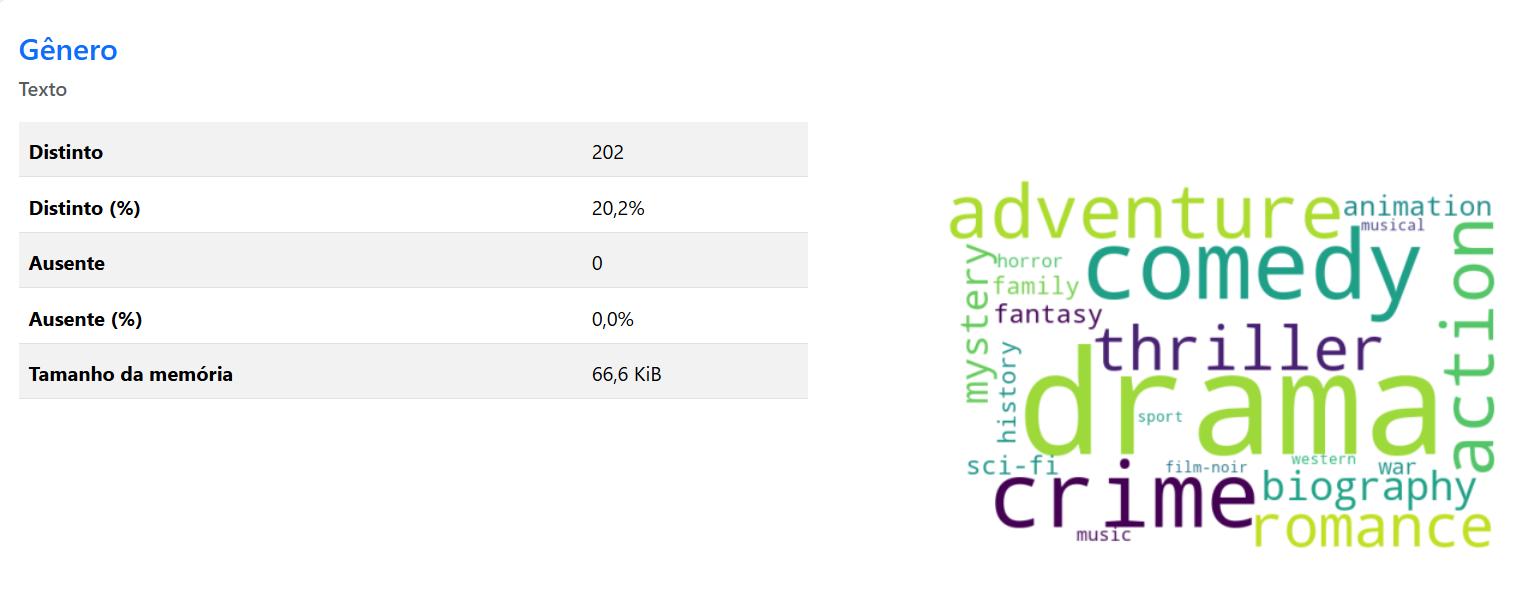
**Figuras, Gráficos**





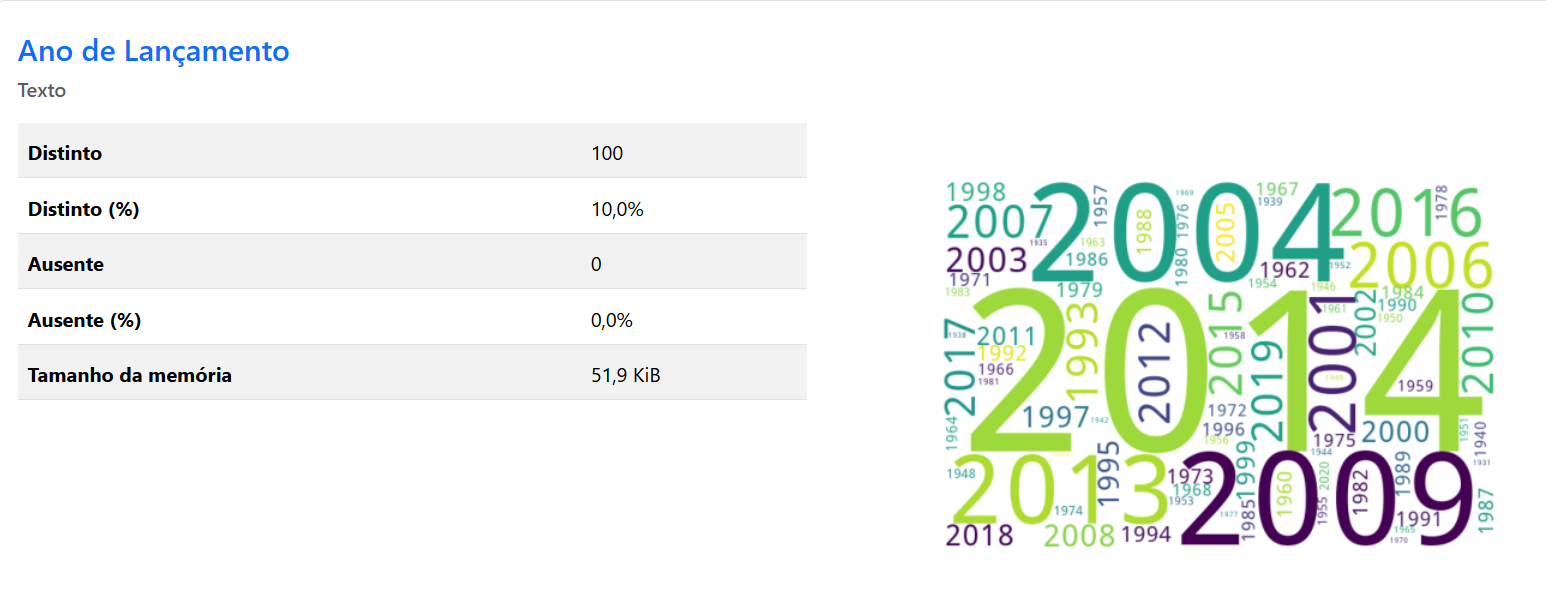
**Figura 1** – Código

**Fonte:** Visual Studio Code



**Figura 2 –** Gênero de filmes

**Fonte:** Kaggle



**Figura 3 – Ano de** lançamento

**Fonte:** Kaggle



**Figura 4 – Título da serie**

**Fonte:** Kaggle

**Figura 5 – Ano de lançamento**

**Fonte:** Kaggle

**Figura 6 – Tempo de execução**

**Fonte:** Kaggle

**Figura 7 – Gênero**

**Fonte:** Kaggle

**Figura 8 – Diretor**

**Fonte:** Kaggle

**Figura 9 – Star 1**

**Fonte:** Kaggle

**Figura 10 – Star 2**

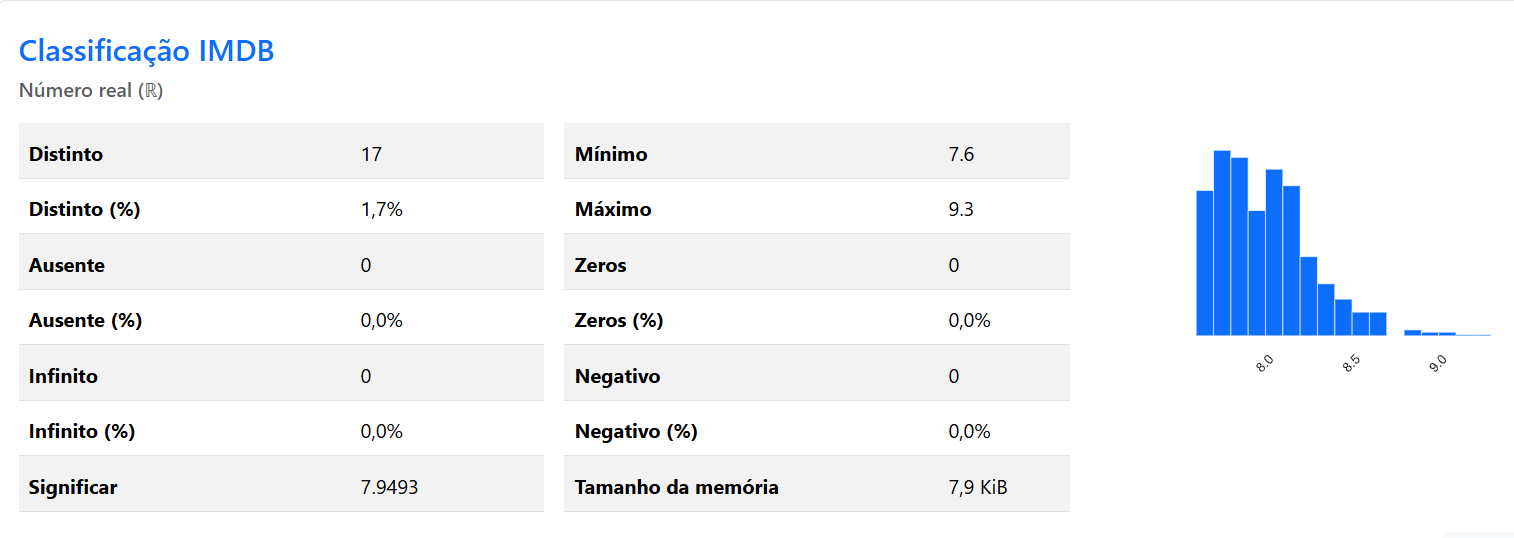
**Fonte:** Kaggle

**Figura 11 – Star 3**

**Fonte:** Kaggle

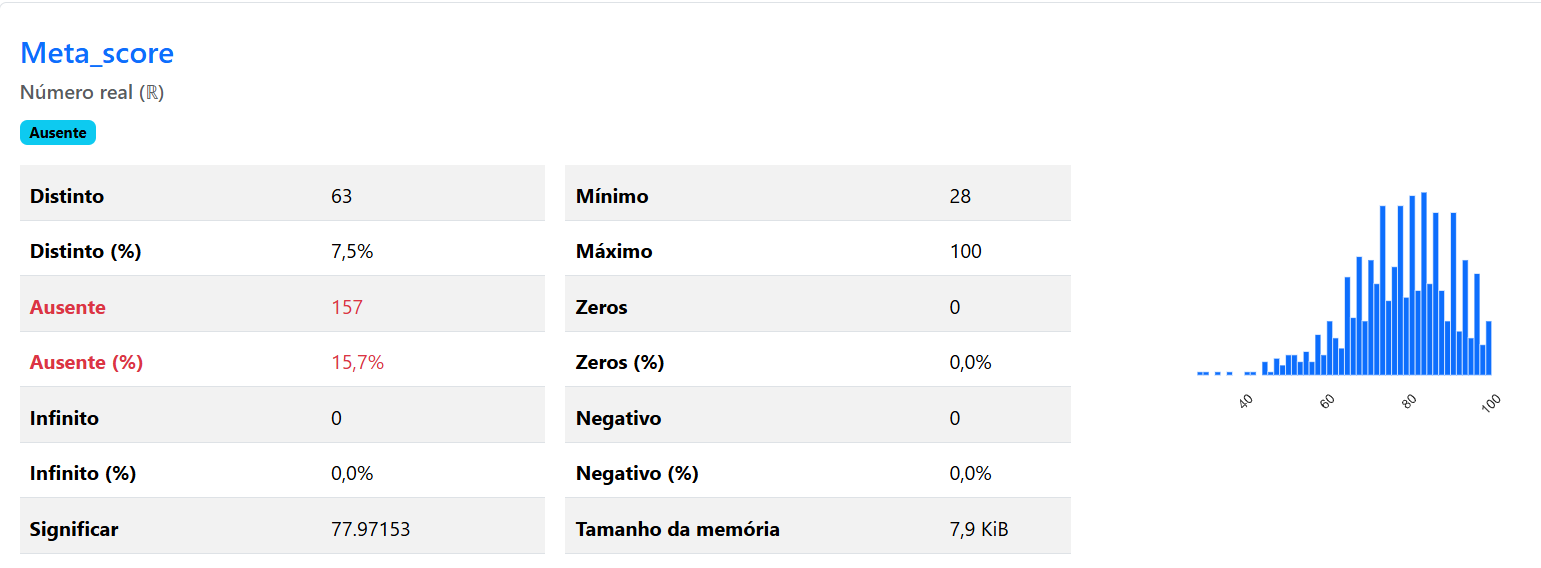
**Figura 12 – Star 4**

**Fonte:** Kaggle



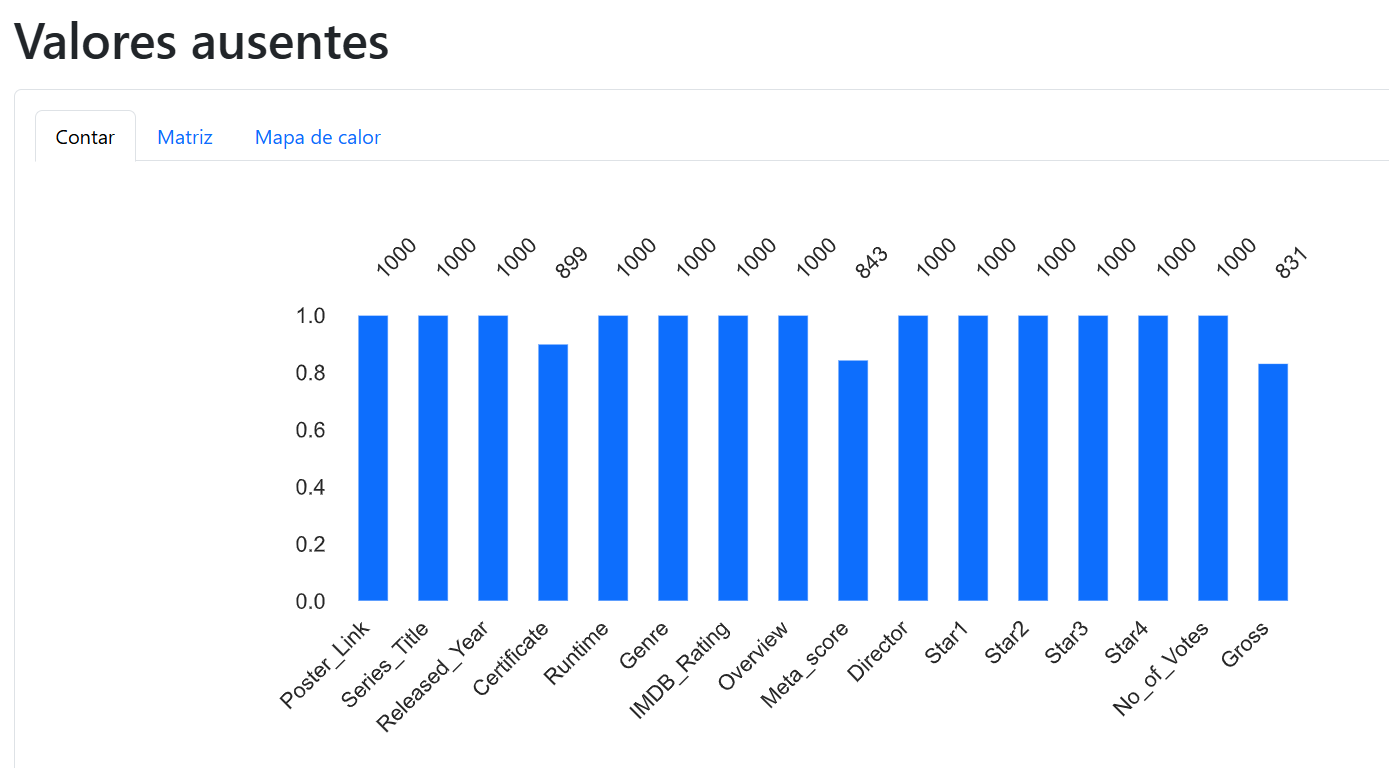
**Gráfico 1**: Classificacao do IMDB

**Fonte**: Kaggle



**Gráfico 2**: Meta Score

**Fonte**: Kaggle



**Gráfico 3**: Valores ausentes

**Fonte**: Kaggle

**Tabelas e Quadros**

**Quadro 1**: **Estatísticas Descritivas da Coluna vote\_average**

|  |  |
| --- | --- |
| **Estatística** | **Valor** |
| 1. Média | 6.39 |
| 1. Mediana | 6.50 |
| 1. Máximo | 10 |
| 1. Mínimo | 1 |

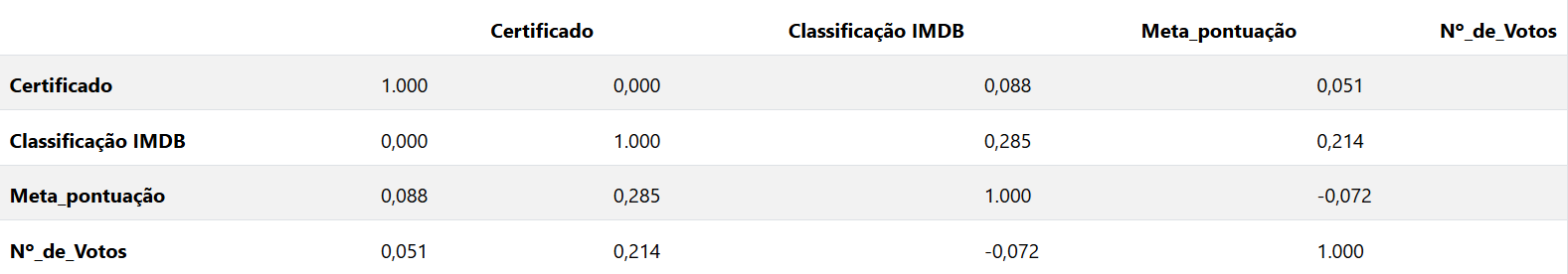
**Fonte:** Dataset Kaggle

**Tabela 1:** **Correlação com vote\_average**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variável | vote\_count | popularity | revenue | budget | runtime |
| Correlação | 0.73 | 0.57 | 0.42 | 0.36 | 0.18 |

**Fonte:** Dataset Kaggle

**Tabela 2:** **Correlações**

**Fonte:** Kaggle

**Tabela 2:** **Amostra**

**Fonte:** Kaggle

**CONCLUSÃO**

O relatório revelou que o dataset possui 23 colunas e 1000 registros. As variáveis mais relevantes incluem: title, genre, budget, revenue, runtime, vote\_average, vote\_count, popularity e release\_year.

Foram detectados valores ausentes em budget (12%), revenue (9%) e runtime (2%). Foi aplicada remoção de registros nulos e normalização das variáveis numéricas. Dados categóricos como genre foram convertidos com codificação one-hot.

A quantidade de votos (vote\_count) e a popularidade foram os atributos com maior correlação com a nota (vote\_average). A árvore de decisão obteve melhor desempenho preditivo. Notou-se também um nível razoável de simetria negativa em vote\_average, indicando concentração de valores acima da média.

O estudo demonstrou a viabilidade da utilização de modelos simples e ferramentas de análise automatizada para extrair conhecimento de bases públicas como o IMDB.

**REFERÊNCIAS**

YDATA Profiling Documentation. Disponível em: <https://docs.ydata.ai/docs/profiling/>. Acesso em: 14 abr. 2025.

Kaggle. IMDB Movies Dataset. Disponível em: https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows. Acesso em: 10 abr. 2025.

PANDAS Documentation. Disponível em: <https://pandas.pydata.org/docs/>. Acesso em: 10 abr. 2025.

Documentações das aulas aplicadas.

**IMDB MOVIES DATASET**

**RESUMO:** In this work, trends in the “IMDB Movies Dataset” are explored and predicted, based on descriptive statistics. The intention is to determine the main models and the main patterns associated between critical characteristics such as genre, investment, box office, average rating and how popular or not

the film is. The method brings inputs in the cases of self-editing of data, training of data opacity, distribution discretion, identification of correlations and training of supervised models in order to predict the rating of films.

The data analyzed are details about 5 thousand films, which reinforces the basis for solid research. Example of graphical display and statistical metrics were executed. It was noted that several films had ratings around 6 and 7 points, while extreme values are rare.

Research notes indicated a notable connection between variables such as "vote\_count" and "popularity" with the average of votes. In this perspective, in addition, Linear Regression and Decision Tree models were evaluated to predict the rating, with satisfactory performance. The work shows predictive analysis for the film industry, giving a valuable impression to producers, distributors, streaming platforms and marketing professionals who want to better understand the factors that influence the success of a film.

**PALAVRAS-CHAVE:** Exploratory Analysis; IMDB; Machine Learning;