Modelagem de padrões de sucesso no IMDb (1960-2024) usando aprendizado de máquina e otimização de hiperparâmetros

Modeling success patterns on IMDb (1960-2024) using machine learning and hyperparameter optimization

Robert Richard das Neves Correia dos Santos
Fatec Baixada Santista - Rubens Lara
robert.santos01@fatec.sp.gov.br

Luís Felipe Ruas do Nascimento
Fatec Baixada Santista - Rubens Lara
luis.nascimento20@fatec.sp.gov.br

Victor Barbosa Gonçalves
Fatec Baixada Santista - Rubens Lara
victor.goncalves4@fatec.sp.gov.br

Victor Roma Vianna Ferreira
Fatec Baixada Santista - Rubens Lara
victor.ferreira38@fatec.sp.gov.br

José Augusto Theodósio Pazetti
Fatec Baixada Santista - Rubens Lara
jose.pazetti01@cps.sp.gov.br

Revista Datapoint
eISSN 3086-433X
Faculdade de Tecnologia Rubens Lara – FATEC
Ciência de Dados
Peridicidade: Anual
Vol 01, n. 01, 2025
revistadp@fatecrl.edu.br

Recebido: Jun 2025
Aceito: Set 2025
Publicado: Dez 2025

URL: https://www.fatecrl.edu.br/revista/datapoint/index.php/dp/article/view/5
DOI: https://doi.org/10.5281/zenodo.19240819

RESUMO

Este estudo analisa e modela padrões de sucesso entre os filmes mais bem avaliados no IMDb entre 1960 e 2024, utilizando aprendizado de máquina supervisionado e não supervisionado. O objetivo é compreender os fatores que explicam o desempenho crítico, comparando modelos de regressão e avaliando quanto da variação das notas IMDb pode ser explicada pelos atributos dos filmes. O pipeline foi desenvolvido em Python (Google Colab) e incluiu etapas de pré-processamento com One-Hot Encoding e Standard Scaler (Scikit-learn), redução de dimensionalidade com PCA, e agrupamento com K-Means. Para a modelagem preditiva, aplicaram-se os algoritmos KNN, SVM, Random Forest e XGBoost, com ajuste de hiperparâmetros via Optuna. As visualizações foram geradas com Matplotlib e Seaborn. Os resultados destacam o XGBoost como modelo de melhor desempenho, revelando que indicações a prêmios, duração e número de votos são as variáveis mais associadas a notas elevadas, oferecendo uma visão ampla dos fatores que caracterizam o sucesso cinematográfico ao longo das décadas.

PALAVRAS-CHAVE: Aprendizado de Máquina; IMDb; Otimização de Hiperparâmetros; Sucesso Cinematográfico; Clusterização

ABSTRACT

This study analyzes and models success patterns among the top-rated films on IMDb from 1960 to 2024 using both supervised and unsupervised machine learning approaches. The aim is to understand the factors explaining critical performance by comparing regression models and assessing how much of IMDb rating variability can be explained by film attributes. The pipeline was implemented in Python (Google Colab) and included pre-processing with One-Hot Encoding and Standard Scaler (Scikit-learn), dimensionality reduction via PCA, and clustering with K-Means. For predictive modeling, KNN, SVM, Random Forest, and XGBoost algorithms were applied, with hyperparameter tuning using Optuna. Visualizations were generated through Matplotlib and Seaborn. Results highlight XGBoost as the best-performing model, indicating that award nominations, duration, and number of votes are the strongest predictors of higher ratings, providing comprehensive insights into the factors shaping cinematic success over time.

KEY-WORDS: Machine Learning; IMDb; Hyperparameter Optimization; Cinematic Success; Clustering

INTRODUÇÃO

De acordo com a B_Arco (2023), “o setor audiovisual tem presenciado uma transformação extraordinária ao longo das últimas décadas, impulsionada pelo avanço rápido da tecnologia”.

Nesse cenário, o uso de técnicas de aprendizado de máquina surge como uma ferramenta promissora para compreender os múltiplos fatores que explicam o sucesso cinematográfico.

Este estudo investiga quais características dos filmes, como gênero, país, década, duração e número de votos, influenciam significativamente as notas IMDb e em que medida tais atributos permitem modelar e explicar a variação dessas avaliações. O objetivo central é analisar e comparar o desempenho de diferentes modelos de aprendizado de máquina na identificação e interpretação dos padrões de sucesso cinematográfico, e não apenas prever notas, mas compreender o comportamento das variáveis que mais contribuem para o reconhecimento crítico.

Para alcançar esses objetivos, foi desenvolvido um *pipeline* híbrido de ciência de dados, estruturado em Python (Google Colab) e composto por etapas de pré-processamento (*One-Hot Encoding* e *Standard Scaler*), análise exploratória, redução de dimensionalidade via PCA, agrupamento com *K-Means* e aplicação de modelos supervisionados (KNN, SVM, *Random Forest* e *XGBoost*). O processo incluiu ainda a otimização automática de hiperparâmetros com a biblioteca *Optuna*, garantindo maior precisão e robustez na comparação entre algoritmos.

Com base nesses procedimentos, o estudo busca não apenas mensurar o desempenho dos modelos por meio de métricas como R^2 e RMSE, mas também identificar os fatores mais determinantes do sucesso cinematográfico no IMDb, oferecendo uma perspectiva quantitativa sobre a evolução da recepção crítica e do público ao longo de mais de seis décadas.

1. FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica deste estudo se divide em duas vertentes principais: a análise de fatores de sucesso cinematográfico e a aplicação de técnicas de aprendizado de máquina no domínio cultural.

1.1 SOBRE O INTERNET MOVIE DATABASE (IMDB)

Segundo Mercado Filho (2022), o *Internet Movie Database* (IMDb) foi criado em 1990 por Col Needham como um banco de dados mantido por fãs e, após migrar para a *web* em 1993, consolidou-se como uma enciclopédia *online* que reúne informações sobre filmes, séries e artistas.

> Além de reunir informações sobre artistas e produções, o site também permite que usuários criem listas e avaliem seus filmes favoritos. Qualquer usuário cadastrado pode dar notas aos títulos. Os votos individuais são agregados e resumidos em uma nota única, exibida com destaque na página principal do título (Melo, 2018 *apud* Coelho; Ferreira; Faustino, 2021, p. 2).

"Com autoridade na indústria de entretenimento, […], a IMDb registra mais de 200 milhões de visitantes mensais na sua página. Além disso, o banco de dados da empresa já ultrapassou a marca de 7.5 milhões de títulos. […]."(Canaltech, [s.d.]).

Em razão de sua credibilidade e amplo volume de informações, o IMDb foi escolhido como fonte de dados deste estudo, por oferecer uma base sólida e representativa para a análise do desempenho e da recepção de produções cinematográficas.

1.2 SOBRE O APRENDIZADO DE MÁQUINA

A segunda vertente revisa o uso de aprendizado de máquina para extrair padrões de grandes conjuntos dos dados utilizados no presente estudo.

O Aprendizado de Máquina (*Machine Learning*), como um subcampo da inteligência artificial (IA), permite que sistemas aprendam autonomamente a partir de dados, sem necessidade de programação direta para tarefas. Essa tecnologia é amplamente aplicada em diversos setores, sendo utilizada em filtros de spam e sistemas de recomendação de *streaming* (Charleaux; Toledo, 2024).

2. PROCEDIMENTOS METODOLÓGICOS

A metodologia deste estudo foi estruturada em um *pipeline* de ciência de dados (uma sequência organizada de etapas automáticas que conduz os dados desde a coleta e preparação até a análise e geração de resultados) executado inteiramente na linguagem de programação

Python, utilizando o ambiente de desenvolvimento Google *Colaboratory* para garantir a reprodutibilidade.

As principais bibliotecas utilizadas foram: Pandas para a manipulação e limpeza do *dataset; Scikit-learn* para as etapas de pré-processamento, modelagem e avaliação; *Matplotlib* e *Seaborn* para a análise exploratória e visualização de resultados; e Optuna para a otimização de hiperparâmetros.

2.1 COLETA E ANÁLISE INICIAL DOS DADOS

O conjunto de dados (*dataset*) utilizado neste estudo foi obtido na plataforma Base dos Dados, a partir do conjunto intitulado "Melhores Filmes por Ano (1960-2024)" (Oliveira, 2024), originalmente proveniente do IMDb. O conjunto de dados conta com 33.600 linhas e 23 colunas, abrangendo variáveis como título do filme, ano de lançamento, gênero, duração, país de origem, diretor, elenco principal, nota IMDb e número de votos, entre outras informações relevantes para análise.

2.2 PRÉ-PROCESSAMENTO

Após a limpeza inicial dos dados, a imputação de valores ausentes foi realizada pelo método *KNN Imputation*. Em seguida, as variáveis categóricas foram codificadas por meio das técnicas *One-Hot Encoding* e *Label Encoding*, que transformam categorias em representações numéricas (Hastie; Tibshirani; Friedman, 2009).

Por fim, as variáveis numéricas foram padronizadas com o *Standard Scaler*, ajustando-as para média zero e desvio padrão unitário. Todas essas etapas de pré-processamento (imputação, codificação e padronização) foram implementadas utilizando-se a biblioteca *Scikit-learn* (Pedregosa *et al*., 2011), de modo a equilibrar a influência das variáveis nos modelos de aprendizado de máquina.

2.3 ANÁLISE EXPLORATÓRIA E PCA

Para identificar padrões e relações entre diferentes décadas e gêneros cinematográficos, foi aplicada a redução de dimensionalidade por meio da *Principal Component Analysis* (PCA).

> A análise de componentes principais, ou PCA, reduz o número de dimensões em grandes conjuntos de dados aos componentes principais que retêm a maior parte das informações originais. Ela faz isso transformando variáveis potencialmente correlacionadas em um conjunto menor de variáveis, chamadas componentes principais (IBM, 2023, s.p.).

Para facilitar a interpretação dos resultados, foram geradas visualizações em 2D e 3D dos componentes principais. Essas visualizações foram criadas utilizando as bibliotecas *Matplotlib* (Hunter, 2007) e *Seaborn* (Waskom, 2021), ambas amplamente empregadas na comunidade *Python* para análise e visualização de dados.

2.4 CLUSTERIZAÇÃO

Em seguida, aplicou-se o algoritmo *K-Means* com o intuito de agrupar filmes com características semelhantes, permitindo observar formações naturais de grupos, como *clusters* de filmes de ação dos anos 1990 em contraste com dramas contemporâneos.

> O *k-means* é um algoritmo que treina um modelo para agrupar objetos semelhantes. Para isso, ele mapeia cada observação no conjunto de dados de entrada para um ponto no espaço de "n" dimensões (em que "n" é o número de atributos da observação). (Amazon Web Services, [s.d.], n.p.).

2.5 COMPARAÇÃO DE MODELOS

Para avaliar e comparar o desempenho de diferentes abordagens de aprendizado supervisionado na estimativa da nota IMDb, os dados foram primeiramente segmentados em conjuntos de treino (70%) e teste (30%).

Foram treinados e avaliados quatro algoritmos de aprendizado supervisionado distintos: *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), *Random Forest* e o *XGBoost*. O objetivo do treinamento é ajustar os parâmetros internos do modelo até que ele aprenda a mapear corretamente os dados rotulados às suas saídas correspondentes (Yokoyama, 2020).

A comparação de desempenho dos modelos foi realizada por meio das métricas R^2 (coeficiente de determinação) e RMSE (*Root Mean Squared Error*) que medem, respectivamente, o quanto da variabilidade da variável resposta é explicada pelo modelo e o desvio padrão dos erros de previsão, calculando a raiz quadrada da média dos erros quadráticos (Dubiella, 2024).

2.6 OTIMIZAÇÃO COM OPTUNA

Realizou-se uma busca automática pelos melhores hiperparâmetros de cada modelo, ajustando elementos como profundidade das árvores, número de estimadores e taxa de aprendizado. Esse processo, conhecido como otimização de hiperparâmetros, tem o objetivo de melhorar o desempenho dos algoritmos por meio da seleção das combinações que produzem os resultados mais precisos.

Os hiperparâmetros são definições feitas antes do treinamento que controlam a arquitetura e o aprendizado do modelo. Eles não são aprendidos pelo algoritmo e influenciam diretamente sua capacidade de generalização. Como ajustá-los é complexo, ferramentas como o *Optuna* automatizam a busca pelas melhores combinações para melhorar o desempenho do modelo (Pinheiro, 2023).

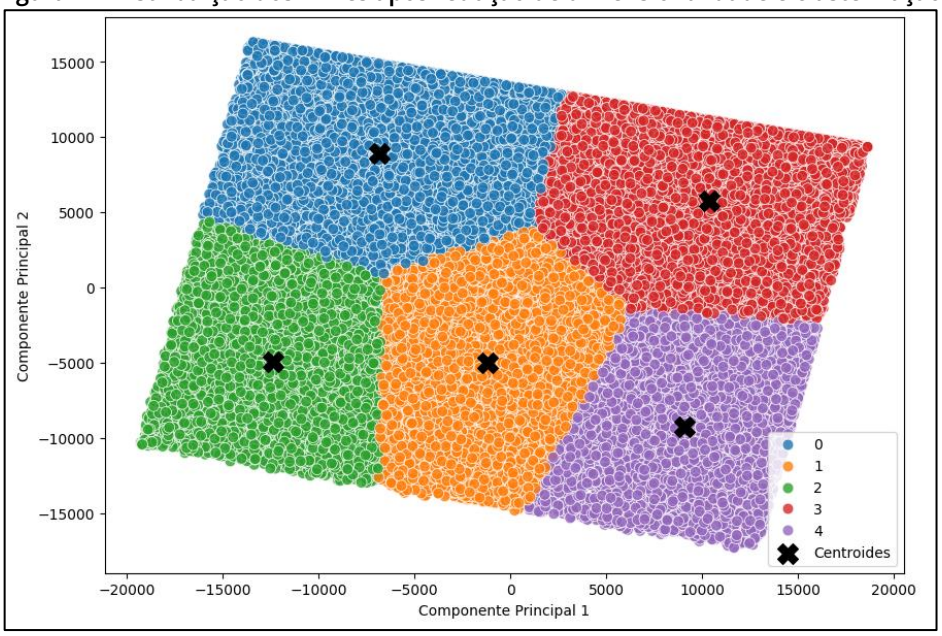
3. RESULTADOS E DISCUSSÃO

A metodologia proposta permitiu identificar padrões consistentes nos dados e avaliar o desempenho dos modelos aplicados.

A redução de dimensionalidade via PCA mostrou-se eficaz para representar a estrutura dos filmes com baixo número de componentes e alta variância explicada. O *K-Means*, aplicado sobre os componentes principais, revelou agrupamentos naturais de produções, sugerindo a existência de perfis distintos de sucesso ao longo das décadas.

A Figura 1 apresenta a projeção 2D da PCA, onde os dois primeiros componentes explicam 91,53% da variância. Os cinco *clusters* formados destacam diferentes combinações de gênero, década e país, ilustrando a diversidade de padrões no *dataset*.

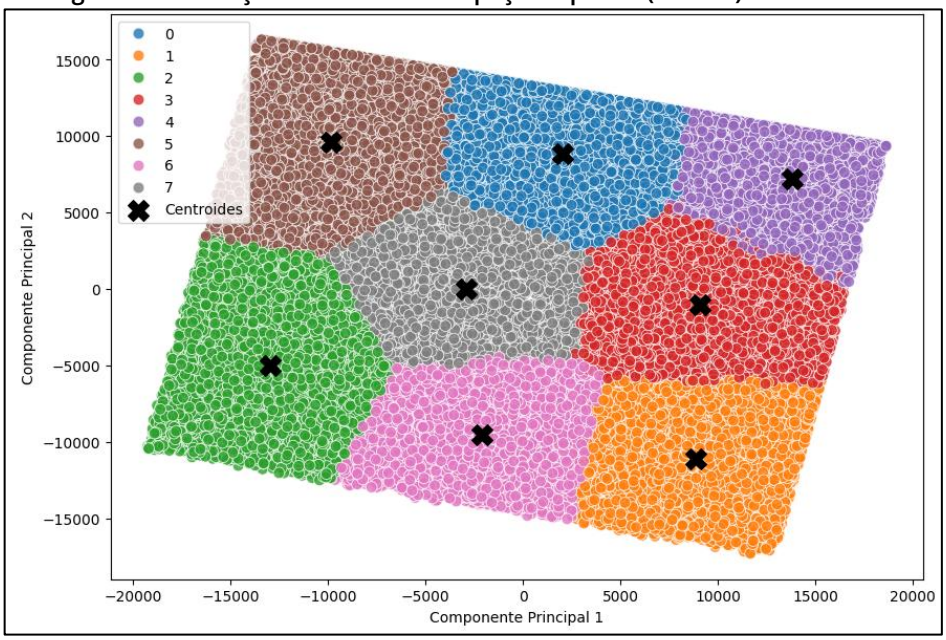
Figura 1 – Distribuição dos filmes após redução de dimensionalidade e clusterização



Fonte: Elaborado no Google Colab pelos Autores (2025)

O gráfico da Figura 2, construída sobre seis componentes principais (100% da variância explicada), mostra uma clusterização mais detalhada, com oito grupos que capturam nuances adicionais e subdivisões entre estilos e períodos cinematográficos.

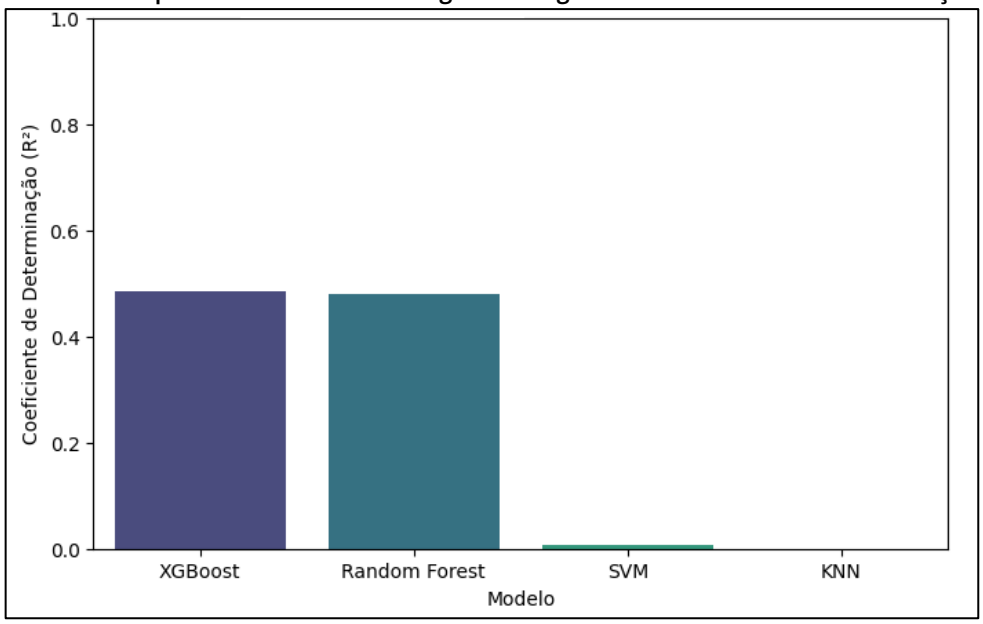
Figura 2 – Formação de *clusters* em espaço ampliado (PCA 6D) com *K-Means*



Fonte: Elaborado no Google Colab pelos Autores (2025)

Na comparação entre modelos de regressão, como mostrado na Figura 3, o *XGBoost* obteve o melhor desempenho com $R^2 \approx 0{,}50$, seguido pelo *Random Forest* ($R^2 \approx 0{,}46$). Ambos superaram amplamente o SVM e o KNN, que apresentaram baixo poder explicativo.

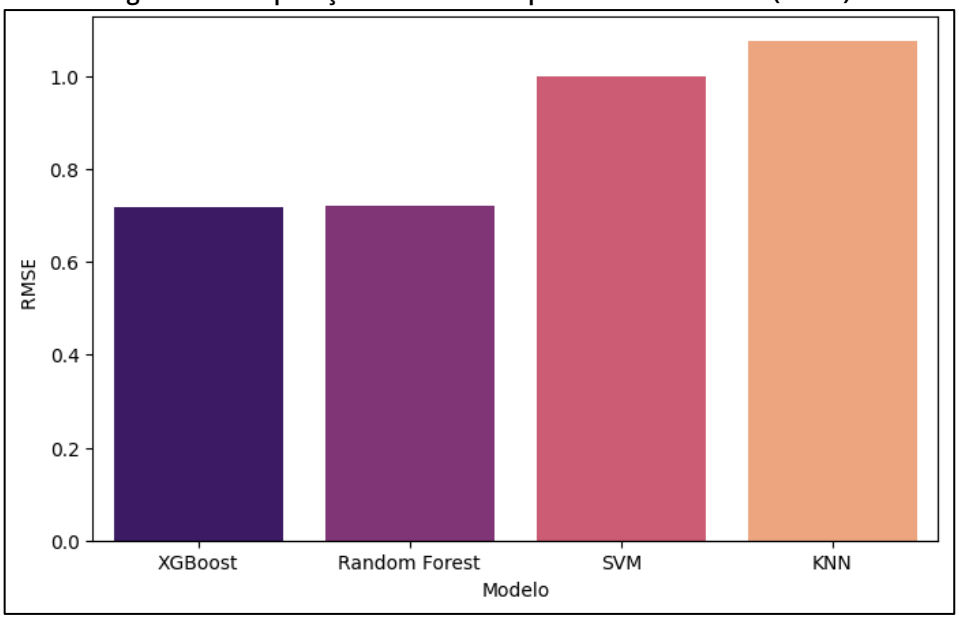
Figura 3 – Desempenho dos modelos de regressão segundo o coeficiente de determinação (R^2)



Fonte: Elaborado no Google Colab pelos Autores (2025)

O erro médio (Figura 4), medido pelo RMSE, confirmou esses resultados: *XGBoost* e *Random Forest* obtiveram menores valores (≈ 0,72), enquanto SVM e KNN registraram erros próximos a 1,0 e 1,1, respectivamente.

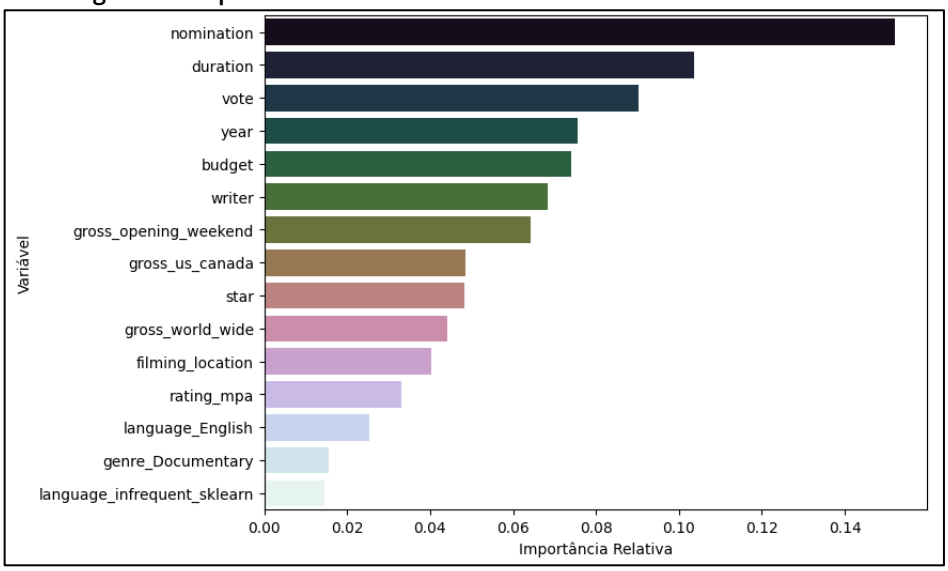
Figura 4 – Comparação dos modelos quanto ao erro médio (RMSE)



Fonte: Elaborado no Google Colab pelos Autores (2025)

A análise de importância das variáveis da Figura 5, referente ao modelo *Random Forest*, mostrou que indicações a prêmios, duração do filme e número de votos foram os fatores mais relevantes para a previsão das notas. Essas variáveis se destacam por representarem tanto o reconhecimento técnico quanto o engajamento do público.

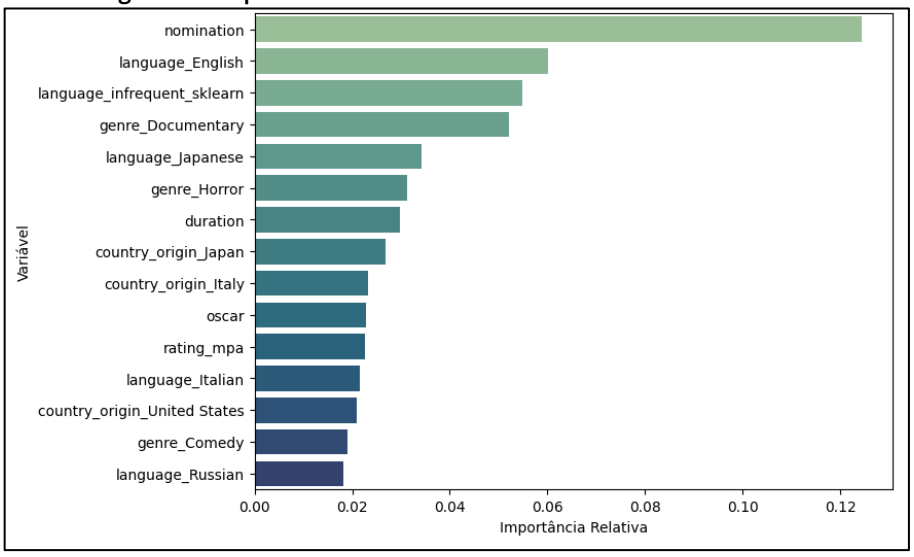
Figura 5 – Importância relativa das variáveis no modelo Random Forest



Fonte: Elaborado no Google Colab pelos Autores (2025)

Na Figura 6, observa-se o comportamento do modelo *XGBoost*, que apresentou um padrão semelhante, mas com nuances adicionais. Além das indicações a prêmios, o algoritmo destacou variáveis relacionadas ao idioma e ao gênero, sugerindo que a recepção crítica varia conforme o idioma de produção e o tipo de narrativa.

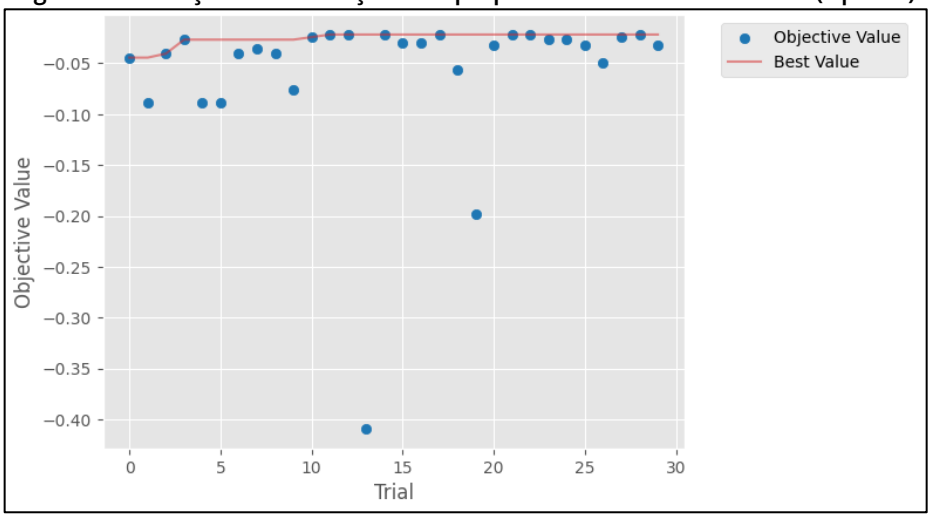
Figura 6 – Importância relativa das variáveis no modelo *XGBoost*



Fonte: Elaborado no Google Colab pelos Autores (2025)

Referente ao processo de otimização do KNN, como ilustrado na Figura 7, nota-se um baixo desempenho do modelo, com valores de R^2 negativos mesmo após 30 tentativas. Isso indica que o algoritmo não conseguiu capturar relações significativas entre as variáveis, sendo inferior a uma simples média na previsão das notas IMDb.

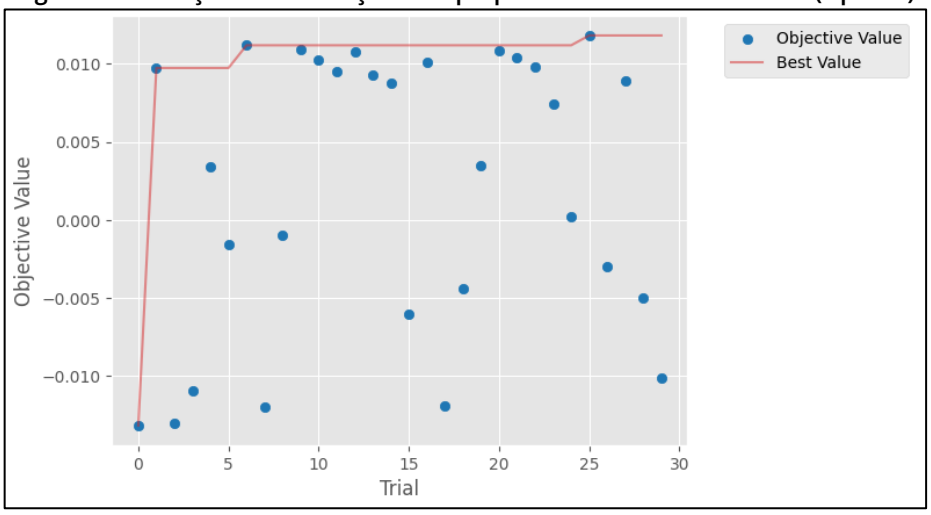
Figura 7 – Evolução da otimização de hiperparâmetros no modelo KNN (Optuna)



Fonte: Elaborado no Google Colab pelos Autores (2025)

Já na Figura 8, que apresenta a otimização do SVM, observa-se comportamento semelhante. Apesar de maior variação entre as tentativas, o melhor valor de R^2 encontrado foi próximo de zero, confirmando que o modelo não se ajusta adequadamente a esse tipo de dado multivariado e heterogêneo.

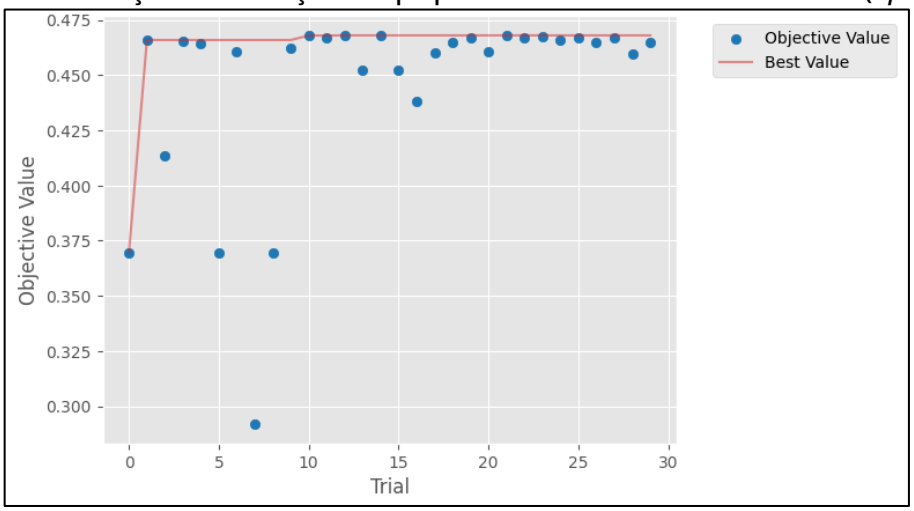
Figura 8 – Evolução da otimização de hiperparâmetros no modelo SVM (Optuna)



Fonte: Feito no Google Colab pelos Autores (2025)

A Figura 9 mostra o desempenho do *Random Forest* durante o processo de otimização. O modelo apresentou rápida convergência e estabilidade, alcançando $R^2 \approx 0{,}46$, o que indica bom equilíbrio entre complexidade e generalização. Ainda que algumas configurações tenham produzido resultados mais baixos, o otimizador encontrou combinações de parâmetros altamente eficazes.

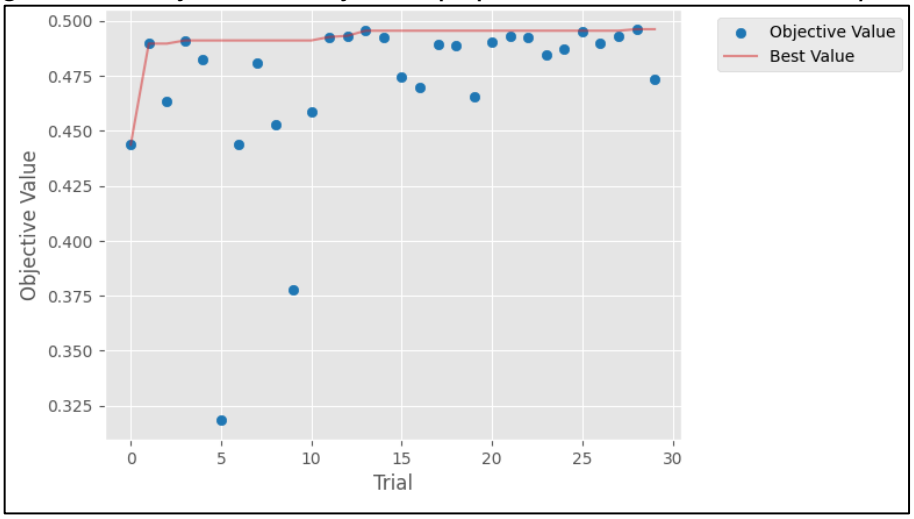
Figura 9 – Evolução da otimização de hiperparâmetros no modelo *Random Forest* (*Optuna*)



Fonte: Feito no Google Colab pelos Autores (2025)

Por fim, a Figura 10 apresenta o comportamento do *XGBoost*. O modelo atingiu o melhor resultado entre todos ($R^2 \approx 0{,}50$), com convergência rápida e consistente já nas primeiras 15 tentativas. Isso demonstra a robustez do algoritmo, cuja estrutura permite identificar padrões sutis e interações não lineares entre as variáveis de forma mais eficiente.

Figura 10 – Evolução da otimização de hiperparâmetros no modelo *XGBoost* (*Optuna*)



Fonte: Elaborado no Google Colab pelos Autores (2025)

De modo geral, os resultados evidenciam que, embora todos os modelos tenham sido otimizados, os baseados em árvores, especialmente o *XGBoost*, apresentaram maior capacidade de generalização e explicação das notas IMDb, consolidando-se como as abordagens mais adequadas para modelar os padrões de sucesso cinematográfico.

4. CONSIDERAÇÕES FINAIS

Este estudo analisou os padrões de sucesso cinematográfico no IMDb (1960–2024) utilizando um pipeline híbrido de aprendizado de máquina. A aplicação do PCA e do *K-Means* revelou agrupamentos consistentes entre filmes de diferentes gêneros, décadas e origens.

Na comparação entre modelos, após a otimização de hiperparâmetros com *Optuna*, o *XGBoost* apresentou o melhor desempenho ($R^2 \approx 0{,}50$; RMSE $\approx 0{,}72$), seguido pelo *Random Forest*, enquanto KNN e SVM mostraram baixo poder explicativo.

A análise das variáveis indicou que indicações a prêmios, duração, número de votos, idioma e gênero foram determinantes para explicar as notas IMDb. Conclui-se que o modelo proposto é eficiente para compreender e modelar os padrões de sucesso cinematográfico, confirmando a influência conjunta do reconhecimento crítico e do engajamento do público na definição do sucesso no IMDb.

REFERÊNCIAS

AMAZON WEB SERVICES. **Como funciona o *clustering* do k-means**. [S.l.]: Amazon Web Services, [s.d.]. Disponível em: https://docs.aws.amazon.com/pt_br/sagemaker/latest/dg/algo-kmeans-tech-notes.html.Acesso em: 13 nov. 2025.

B_ARCO. **A evolução da tecnologia no campo audiovisual**: impactos e oportunidades. 1 dez. 2023. Disponível em: https://barco.art.br/evolucao-da-tecnologia-no-campo-audiovisual/. Acesso em: 15 nov. 2025.

CANALTECH. **Tudo sobre IMDb**. [S.l.], [s.d.]. Disponível em: https://canaltech.com.br/empresa/imdb/. Acesso em: 10 nov. 2025.

CHARLEAUX, Lupa; TOLEDO, Victor. **O que é *Machine Learning*?** Tecnoblog, out. 2024. Disponível em: https://tecnoblog.net/responde/machine-learning-o-que-e-como-funciona-e-quais-sao-os-tipos-de-aprendizado-de-maquina/. Acesso em: 10 nov. 2025.

COELHO, Isabela da Silva Dias; FERREIRA, Marcella Meirelles; FAUSTINO, Marcus Vinícius. ***Machine Learning* no Mundo Cinematográfico**. In: UEADSL 2021.1: SUBMISSÃO DE TRABALHOS PARA O ANFITEATRO (GRADUAÇÃO E PÓS), 2021, [S.l.]. Anais [...]. [S.l.]: TextoLivre, 2021. Disponível em: https://textolivre.pro.br/mod/data/view.php?d=18&rid=533. Acesso em: 10 nov. 2025.

DUBIELLA, Larissa. **Métricas de avaliação para modelos de regressão**. Alura Artigos, 03 nov. 2024. Disponível em: https://www.alura.com.br/artigos/metricas-de-regressao. Acesso em: 09 nov. 2025.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning:** data mining, inference, and prediction. 2. ed. New York: Springer, 2009.

HUNTER, John D. Matplotlib: *A 2D graphics environment*. ***Computing in Science & Engineering***, v. 9, n. 3, p. 90-95, 2007.

IBM. **O que é a análise de componentes principais (PCA)?** [S.l.]: IBM, 2023. Disponível em: https://www.ibm.com/br-pt/think/topics/principal-component-analysis. Acesso em: 11 nov. 2025.

MERCADO FILHO, Alejandro Sigfrido. ***Rotten Tomatoes* e IMDb:** como funcionam os sites de críticas? Mega Curioso, 16 nov. 2022. Disponível em: https://www.megacurioso.com.br/artes-cultura/123509-rotten-tomatoes-e-imdb-como-funcionam-os-sites-de-criticas.htm. Acesso em: 10 nov. 2025.

OLIVEIRA, Vinícius G. de. ***IMDb Movies* (1960-2024) - *Top Rated***. [S.l.]: Kaggle, 2024. Dataset. Disponível em: https://www.kaggle.com/datasets/vinciusgdeoliveira/imdb-movies-1960-2024-top-rated. Acesso em: 10 nov. 2025.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. ***Journal of Machine Learning Research***, v. 12, p. 2825-2830, 2011.

PINHEIRO, João Manoel Herrera. Um estudo sobre Algoritmos de *Boosting* e a Otimização de Hiperparâmetros Utilizando Optuna. 2023. 147 p. **Monografia** (Engenharia Mecatrônica) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2023.

WASKOM, Michael L. *Seaborn: statistical data visualization.* ***Journal of Open Source Software***, v. 6, n. 60, p. 3021, 2021.

YOKOYAMA, Naoki. **Modelos de *Machine Learning***. *Medium*, 30 out. 2020. Disponível em: https://naokiyokoyama.medium.com/modelos-de-machine-learning-bcb3f8ed1513. Acesso em: 15 nov. 2025.