UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Pedro Milvar Santos Vieira

Previsão do vencedor do Oscar de Melhor Filme: Uma abordagem com algoritmos de classificação

Uberlândia, Brasil

2025

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Pedro Milvar Santos Vieira

Previsão do vencedor do Oscar de Melhor Filme: Uma abordagem com algoritmos de classificação

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Paulo Henrique Ribeiro Gabriel

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2025

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Faculdade de Computação
Av. João Naves de Ávila, nº 2121, Bloco 1A - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4144 - http://www.portal.facom.ufu.br/ facom@ufu.br

ATA DE DEFESA - GRADUAÇÃO

Curso de Graduação em:	Sistemas de Informação: Bacharelado				
Defesa de:	Trabalho de Conclusão de Curso 2 (FACOM31802)				
Data:	13/05/2025	Hora de início:	14:00	Hora de encerramento:	15:00
Matrícula do Discente:	11921BSI207				
Nome do Discente:	Pedro Milvar Santos Vieira				
Título do Trabalho:	Previsão do vencedor do Oscar de Melhor Filme: Uma abordagem com algoritmos de classificação				
A carga horária curricular foi cumprida integralmente?	(X) Sim ( ) Não				

Reuniu-se remotamente através da plataforma MS Teams, da Universidade Federal de Uberlândia, a Banca Examinadora, designada pelo Colegiado do Curso de Graduação em Sistemas de Informação, assim composta: Professores: Dr. Bruno Augusto Nassif Travençolo - FACOM/UFU; Dr. Rodrigo Sanches Miani - FACOM/UFU; e Dr. Paulo Henrique Ribeiro Gabriel - FACOM/UFU, orientador do candidato.

Iniciando os trabalhos, o presidente da mesa, Dr. Paulo Henrique Ribeiro Gabriel, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao discente a palavra, para a exposição do seu trabalho. A duração da apresentação do discente e o tempo de arguição e resposta foram conforme as normas do curso.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

(X) Aprovado Nota 95 (Somente números inteiros)

OU

( ) Aprovado(a) sem nota.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Bruno Augusto Nassif Travençolo**, **Professor(a) do Magistério Superior**, em 13/05/2025, às 14:59, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Paulo Henrique Ribeiro Gabriel**, **Professor(a) do Magistério Superior**, em 13/05/2025, às 14:59, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rodrigo Sanches Miani**, **Professor(a) do Magistério Superior**, em 13/05/2025, às 15:00, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6332824** e o código CRC **C93F8D4E**.

Referência: Processo nº 23117.031552/2025-14 SEI nº 6332824

À minha mãe Romilda, exemplo de força e determinação, que sempre me mostrou que, embora o caminho nem sempre seja fácil, a recompensa é grandiosa quando seguimos com fé.A minha irmã Andressa, exemplo de vida e conselheira para todas as horas, que me ensina que intensidade e determinação são os caminhos mais seguros para alcançar o sucesso.A minha namorada Maraynne, companheira que escolhi para a vida, e que a cada dia reafirma minha decisão, ao demonstrar que a vida a dois é mais leve, e que apoio e carinho são sempre o melhor caminho. Este trabalho é tão meu quanto de vocês.

Agradecimentos

Agradeço primeiramente à minha mãe, Romilda, pelo amor incondicional, pelo exemplo de força, e por sempre acreditar em meu potencial mesmo nos momentos mais desafiadores. Sua dedicação e ensinamentos foram fundamentais para que eu chegasse até aqui.

À minha irmã, Andressa, pelo exemplo de determinação e pelas palavras de motivação que me impulsionaram em momentos de dúvida. Sua trajetória e seus conselhos sempre foram fonte de inspiração para mim.

À minha namorada, Maraynne, meu eterno agradecimento pelo apoio constante, pela paciência, pela parceria diária, pelo amor e por me incentivar a ser a minha melhor versão. Seu carinho e presença tornaram essa caminhada mais leve e significativa.

Ao meu orientador, Professor Dr. Paulo Henrique Ribeiro Gabriel, pela orientação atenciosa, pelo conhecimento compartilhado e pela confiança depositada em meu trabalho. Sua dedicação e comprometimento foram essenciais para o desenvolvimento deste projeto.

A banca examinadora, prof. Dr. Bruno Augusto Nassif Travençolo e prof. Dr. Rodrigo Sanches Miani não só por disponibilizarem de seu tempo para comtemplar minha apresentação mas também pelos ensinamentos dados durante minha graduação.

Aos meus amigos, João Vítor, Gabriel Andretta e Guilherme Dias (Manbo) que mesmo à distância ou na correria do dia a dia, sempre estiveram presentes com palavras de incentivo, apoio e compreensão, agradeço por fazerem parte dessa jornada.

E não posso deixar de agradecer as várias pessoas que sempre me ajudaram e apoiaram de alguma forma para chegar aonde chegueis

A todos que, de alguma forma, contribuíram para a realização deste trabalho, deixo aqui a minha mais sincera gratidão.

"A coisa mais incompreensível sobre o mundo é que ele é compreensível." Einstein (2024)

Resumo

Este trabalho apresenta uma abordagem quantitativa para prever o vencedor do Oscar de Melhor Filme utilizando técnicas de aprendizado de máquina supervisionado. A partir da coleta de dados históricos de premiações relevantes como BAFTA, DGA, PGA e SAG, entre os anos de 2004 e 2023, foram construídas variáveis indicadoras da performance dos filmes nas etapas anteriores da temporada de premiações. Com base nesse conjunto de dados estruturado, foram aplicados os algoritmos de Regressão Logística, Árvore de Decisão e Floresta Aleatória, os quais foram avaliados por métricas como acurácia, precisão, *recall* e *F1-score*. Além disso, foram realizadas análises estatísticas como correlação, teste Qui-Quadrado e coocorrência para identificar padrões de associação entre as premiações. A aplicação prática dos modelos para prever o Oscar 2025 demonstrou resultados consistentes, com destaque para a Floresta Aleatória, que apresentou o melhor desempenho. Os resultados indicam que, mesmo em eventos de natureza subjetiva como o Oscar, é possível identificar padrões históricos com potencial preditivo, ampliando o uso da ciência de dados em contextos culturais.

Palavras-chave: aprendizado de máquina, Oscar, premiações, classificação, previsão.

Abstract

This work presents a quantitative approach to predict the Academy Award for Best Picture winner using supervised machine learning techniques. Based on historical data from major awards such as BAFTA, DGA, PGA, and SAG, covering the years from 2004 to 2023, binary variables were created to indicate each film's performance during the awards season. Using this structured dataset, Logistic Regression, Decision Tree, and Random Forest algorithms were applied and evaluated through accuracy, precision, recall, and F1-score metrics. In addition, statistical analyses such as correlation, Chi-Square test, and co-occurrence were conducted to identify association patterns between awards. The practical application of the models for predicting the 2025 Oscars showed consistent results, with Random Forest achieving the best performance. The findings suggest that even in subjective contexts such as film awards, it is possible to identify predictive historical patterns, expanding the role of data science in cultural domains.

Keywords: *machine learning, Oscars, film awards, classification, prediction.*

Lista de ilustrações

Figura 1 – Exemplo visual da estrutura de uma Árvore de Decisão, com raiz, nós internos e folhas. 19
Figura 2 – Representação gráfica da estrutura de uma Floresta Aleatória, composta por múltiplas Árvores de Decisão independentes. 20
Figura 3 – Matriz de confusão ilustrando TP, FP, FN e TN. 21
Figura 4 – Correlação de *Pearson* entre premiações e o Oscar 28
Figura 5 – Correlação de *Spearman* entre premiações e o Oscar 29
Figura 6 – Taxa de coocorrência entre as principais premiações e o Oscar de Melhor Filme. 31
Figura 7 – Matriz de Confusão - Regressão Logística 35
Figura 8 – Matriz de Confusão - Árvore de Decisão 35
Figura 9 – Matriz de Confusão - Floresta Aleatória 36
Figura 10 – Boxplot das probabilidades de vitória por modelo 38
Figura 11 – Distribuição das probabilidades de vitória por modelo 38
Figura 12 – Top 10 filmes com maior chance de vitória segundo cada modelo 39
Figura 13 – Probabilidade de Vitória - Regressão Logística - Oscar 2025 41
Figura 14 – Probabilidade de Vitória - Árvore de Decisão - Oscar 2025 41
Figura 15 – Probabilidade de Vitória - Floresta Aleatória - Oscar 2025 42
Figura 16 – Probabilidade Média de Vitória - Oscar 2025 42

Lista de tabelas

Tabela 1 – Resultados do Teste Qui-Quadrado entre cada premiação e o Oscar . . 30
Tabela 2 – Métricas de Avaliação dos Modelos Preditivos 36
Tabela 3 – Atributos utilizados nos modelos preditivos: filmes indicados ao Oscar 2025 (entradas), premiações prévias recebidas (atributos) variando entre 0 (não vitória) e 1 (vitória). 40
Tabela 4 – Previsão de Vencedor do Oscar 2025 por Modelo 40

Lista de abreviaturas e siglas

AD	Árvore de Decisão - *Decision Tree*
FA	Floresta Aleatória - *Random Florest*
KNN	K-Vizinhos Mais Próximos – *K-Nearest Neighbors*
RL	Regressão Logística - *Logistic regression*
SVM	Máquina de Vetores de Suporte – *Support Vector Machine*
AI	Inteligência Artificial – *Artificial Intelligence*
ML	Aprendizado de Máquina – *Machine Learning*
PCA	Análise de Componentes Principais – *Principal Component Analysis*
FP	Falso Positivo
FN	Falso Negativo
TP	Verdadeiro Positivo
TN	Verdadeiro Negativo
BAFTA	*British Academy of Film and Television Arts*
DGA	*Directors Guild of America*
NBR	*National Board of Review*
PGA	*Producers Guild of America*
SAG	*Screen Actors Guild*

Sumário

1 INTRODUÇÃO 14
1.1 Organização do Trabalho 15

2 FUNDAMENTAÇÃO TEÓRICA 16
2.1 Aprendizado de Máquina 16
2.1.1 Aprendizado Supervisionado 16
2.1.2 Aprendizado Não Supervisionado 16
2.2 Modelos de Classificação 17
2.2.1 Regressão Logística 17
2.2.2 Árvore de Decisão 18
2.2.3 Floresta Aleatória 19
2.3 Avaliação dos Modelos Preditivos 20
2.4 Premiações Cinematográficas 22
2.5 Trabalhos Relacionados 23

3 DESENVOLVIMENTO E RESULTADOS 25
3.1 Linguagem e Bibliotecas Utilizadas 25
3.1.1 Web Scraping e Pré-Processamento de Dados 26
3.1.2 Tratamento e Consolidação dos Dados 26
3.2 Análise Exploratória dos Dados 27
3.2.1 Análise de Correlação 27
3.2.1.1 Correlação de *Pearson* 28
3.2.1.2 Correlação de *Spearman* 28
3.2.2 Análise de Associação com o Oscar: Teste Qui-Quadrado 29
3.2.3 Análise de Coocorrência de Premiações 30
3.3 Construção dos Modelos Preditivos 32
3.3.1 Definição do Problema 32
3.3.2 Separação dos Dados 32
3.3.3 Modelos Aplicados 32
3.3.4 Ferramentas e Bibliotecas 32
3.3.5 Treinamento dos Modelos 32
3.3.5.1 Treinamento da Regressão Logística 33
3.3.5.2 Treinamento da Árvore de Decisão 33
3.3.5.3 Treinamento da Floresta Aleatória 34
3.3.6 Matrizes de Confusão por Modelo 34
3.3.7 Métricas de Avaliação dos Modelos Preditivos 36

3.3.8 Distribuição das Probabilidades de Vitória 37
3.3.9 Top 10 Filmes com Maior Probabilidade de Vitória 39
3.4 Avaliação dos Modelos com Dados de 2025 39
3.4.1 Previsão de Vencedor do Oscar 2025 39
3.4.2 Análise Gráfica das Predições 40
3.4.3 Discussão dos Resultados 42
3.4.4 Limitações Observadas 43
3.5 Considerações Finais 44

4 CONCLUSÃO 45

REFERÊNCIAS 47

1 Introdução

A aplicação de técnicas de aprendizado de máquina tem crescido significativamente nos últimos anos, incluindo em áreas que envolvem aspectos culturais, subjetivos e simbólicos. Um exemplo notável é o interesse acadêmico e prático na previsão de resultados de premiações cinematográficas, especialmente o Oscar de Melhor Filme – a mais prestigiada distinção concedida pela Academia de Artes e Ciências Cinematográficas. Dada a complexidade do processo de votação, que envolve múltiplos critérios objetivos e subjetivos, prever o vencedor dessa categoria representa um desafio relevante para a ciência de dados.

Segundo estudos de aprendizado supervisionado (HASTIE; TIBSHIRANI; FRIEDMAN, 2009), algoritmos como Regressão Logística, Árvore de Decisão e Floresta Aleatória têm se mostrado eficazes na detecção de padrões em bases históricas. No contexto cinematográfico, a utilização desses modelos permite explorar relações entre o desempenho de um filme em premiações anteriores e sua probabilidade de vencer o Oscar, oferecendo uma abordagem quantitativa para um fenômeno tradicionalmente interpretado de forma qualitativa.

A estrutura das temporadas de premiações apresenta uma sequência lógica que influencia as campanhas e a visibilidade dos filmes, conforme discutido por Villaça (2017). Premiações como o BAFTA, DGA *(Directors Guild of America)*, PGA *(Producers Guild of America)*, SAG *(Screen Actors Guild)* e outras atuam como indicadores antecipados do Oscar. A análise de dados históricos dessas premiações, portanto, pode fornecer insights preditivos valiosos, principalmente quando modelados com o suporte de algoritmos de aprendizado supervisionado.

Este estudo propõe o desenvolvimento de um sistema preditivo capaz de estimar a probabilidade de um filme vencer o Oscar de Melhor Filme com base em dados históricos das premiações que o antecedem. Para isso, foram utilizadas as bibliotecas `scikit-learn`, `pandas` e ferramentas de visualização estatística em *Python*, compondo um pipeline completo desde a coleta dos dados (via *web scraping*) até a modelagem, validação e análise dos resultados.

A metodologia adotada segue os princípios do processo KDD (*Knowledge Discovery in Databases*), conforme proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996). Esse processo inclui as etapas de seleção dos dados, pré-processamento, transformação, mineração de dados e avaliação/interpretação dos resultados, sendo amplamente utilizado para extração de conhecimento em bases estruturadas.

Foram consideradas as edições do Oscar entre 2004 e 2023, construindo-se uma base de dados estruturada com variáveis binárias indicativas de vitória em cada premiação

relevante. Em seguida, foram aplicadas análises exploratórias, como correlação de *Pearson* e *Spearman*, análise de coocorrência e o teste de independência qui-quadrado, com o objetivo de validar estatisticamente as relações entre as premiações e o Oscar. Por fim, os modelos de classificação foram avaliados com base em métricas como acurácia, precisão, *recall* e *F1-score*, além da matriz de confusão.

Com os modelos avaliados e validados em dados históricos, procedeu-se então à aplicação prática do sistema preditivo nos dados referentes à temporada de premiações de 2025. Essa etapa teve como objetivo testar a capacidade dos algoritmos em generalizar padrões aprendidos no passado para prever, com base nos vencedores das premiações antecedentes, qual filme teria a maior probabilidade de conquistar o Oscar de Melhor Filme em 2025. Os resultados obtidos permitiram comparações entre os modelos e uma análise crítica das previsões geradas, consolidando as contribuições do trabalho tanto em termos técnicos quanto em sua relevância aplicada.

A principal contribuição deste trabalho reside na utilização de aprendizado de máquina como ferramenta de apoio à análise cultural, demonstrando que, apesar do caráter subjetivo do Oscar, existem padrões históricos que podem ser explorados com rigor técnico. Os resultados obtidos reforçam a aplicabilidade da ciência de dados em domínios não convencionais, como o entretenimento, e incentivam futuras pesquisas interdisciplinares que combinem estatística, cultura e inteligência artificial.

1.1 Organização do Trabalho

Este trabalho está estruturado em quatro capítulos.

O Capítulo 1 apresenta a introdução do tema, contextualizando a problemática, os objetivos da pesquisa e a justificativa do estudo, além de uma visão geral da metodologia aplicada.

O Capítulo 2 reúne a fundamentação teórica necessária, abordando os conceitos de aprendizado de máquina (Seção 2.1), os principais modelos de classificação (Seção 2.2), as métricas de avaliação dos modelos preditivos (Seção 2.3), as premiações cinematográficas consideradas (Seção 2.4) e os trabalhos relacionados (Seção 2.5).

O Capítulo 3 descreve o desenvolvimento prático do projeto e os resultados obtidos. São apresentados a linguagem e bibliotecas utilizadas (Seção 3.1), o processo de coleta e pré-processamento dos dados (Seção 3.1.1), a análise exploratória dos dados (Seção 3.2), a construção dos modelos preditivos (Seção 3.3) e a avaliação das previsões para o Oscar de 2025 (Seção 3.4).

Por fim, o Capítulo 4 apresenta as conclusões do estudo, destacando as principais contribuições, limitações enfrentadas e sugestões para trabalhos futuros.

2 Fundamentação Teórica

Este capítulo apresenta os conceitos teóricos que embasam o desenvolvimento do sistema preditivo de vencedores do Oscar. São abordados os fundamentos de aprendizado de máquina, técnicas supervisionadas, algoritmos de classificação, além de um panorama sobre as principais premiações cinematográficas utilizadas como variáveis no modelo.

2.1 Aprendizado de Máquina

O aprendizado de máquina (*machine learning*) é um subcampo da inteligência artificial que permite a extração de padrões complexos a partir de dados, sem necessidade de programação explícita para cada tarefa. Segundo a *Amazon Web Services* (AWS, 2025a), o aprendizado de máquina pode ser supervisionado ou não supervisionado, e é utilizado para treinar modelos a partir de dados históricos e realizar previsões ou classificações. Dentre os principais algoritmos de classificação, destaca-se a Regressão Logística (AWS, 2025b), utilizada para prever categorias binárias, e a Floresta Aleatória, um conjunto de árvores de decisão ideal para conjuntos de dados mais complexos (IBM, 2025g).

2.1.1 Aprendizado Supervisionado

O aprendizado supervisionado consiste em treinar algoritmos com um conjunto de dados rotulado. O objetivo é aprender uma função que relacione entradas e saídas, e que possa generalizar bem para novos dados (AWS, 2025a). A biblioteca `scikit-learn` oferece implementações eficientes de algoritmos supervisionados, como Regressão Logística e Floresta Aleatória (PEDREGOSA et al., 2011).

2.1.2 Aprendizado Não Supervisionado

O aprendizado não supervisionado, por sua vez, trabalha com dados não rotulados, ou seja, sem uma variável-alvo claramente definida. Conforme a *Amazon Web Services* (AWS, 2025a), esse tipo de aprendizado busca encontrar padrões ocultos ou estruturas nos dados, como agrupamentos, associações e redução de dimensionalidade. Técnicas comuns incluem algoritmos de clusterização, como *K-Means*, e análise de componentes principais (PCA). Embora não tenha sido utilizado diretamente neste trabalho, o aprendizado não supervisionado é amplamente aplicado em áreas como segmentação de público e detecção de anomalias.

2.2 Modelos de Classificação

Modelos de classificação são uma das principais abordagens em aprendizado de máquina supervisionado e têm como objetivo atribuir uma ou mais classes a uma observação com base em suas características. Em outras palavras, tratam-se de algoritmos capazes de prever rótulos categóricos, como "*spam*" ou "não *spam*", "doente" ou "saudável", "sim" ou "não". Esses modelos são amplamente utilizados em aplicações práticas, como diagnóstico médico, reconhecimento de voz, detecção de fraudes e, no caso deste trabalho, predição de vencedores de premiações culturais.

O funcionamento de um modelo de classificação baseia-se na análise de um conjunto de dados rotulado, ou seja, no qual já se conhece a classe correta para cada exemplo. A partir disso, o modelo aprende padrões e relações entre as variáveis de entrada (também chamadas de atributos ou preditores) e a saída (classe), permitindo generalizar essas relações para novas instâncias.

De acordo com a IBM (2025e), os principais tipos de classificadores incluem modelos lineares, como a regressão logística; algoritmos baseados em árvores, como a Árvore de Decisão e a Floresta Aleatória; além de métodos mais complexos, como Máquinas de Vetores de Suporte (SVM) e redes neurais. Cada técnica possui características distintas e pode ser mais adequada para certos tipos de problemas, dependendo da estrutura dos dados, da presença de ruído e do objetivo da análise.

A escolha do modelo ideal depende também de critérios como interpretabilidade, desempenho computacional e capacidade de generalização. Em contextos onde a transparência é fundamental, modelos mais simples e interpretáveis, como a regressão logística, são frequentemente preferidos. Já em cenários com grandes volumes de dados e interações não lineares, modelos mais sofisticados, como Florestas Aleatórias, podem apresentar melhor desempenho.

A seguir, serão descritos os algoritmos de classificação utilizados neste trabalho, com ênfase em seus fundamentos teóricos e aplicação na predição do Oscar de Melhor Filme.

2.2.1 Regressão Logística

A regressão logística é um algoritmo amplamente utilizado para modelagem de classificações binárias, ou seja, quando o resultado desejado é representado por dois estados possíveis, como "sim" ou "não". Segundo a IBM (2025f), essa técnica é baseada na regressão linear, mas em vez de prever valores contínuos, ela estima a probabilidade de ocorrência de um evento com base em uma função logística, também conhecida como função sigmoide.

Essa função transforma uma combinação linear dos preditores em uma probabilidade entre 0 e 1:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}} \tag{2.1}$$

Nessa equação, $p(x)$ representa a probabilidade estimada do evento ocorrer. O termo β_0 é o intercepto do modelo e os coeficientes $\beta_1, \ldots, \beta_k$ indicam a influência de cada variável $x_1, \ldots, x_k$ sobre a variável-alvo. Esses coeficientes são ajustados pelo método da máxima verossimilhança, que busca os parâmetros que maximizam a chance de observar os dados disponíveis.

A principal vantagem do modelo logístico é sua capacidade de interpretação: é possível compreender a direção e magnitude dos efeitos de cada variável, o que torna a técnica adequada em contextos onde a transparência do modelo é importante.

Neste trabalho, a regressão logística foi utilizada para estimar a probabilidade de um filme vencer o Oscar de Melhor Filme, com base em indicadores binários de vitórias em premiações anteriores. O modelo foi treinado com dados históricos de 2004 a 2023 e testado na previsão do Oscar de 2025, apresentando desempenho satisfatório com boa precisão e equilíbrio entre as classes.

2.2.2 Árvore de Decisão

A Árvore de Decisão é um algoritmo de classificação que utiliza uma estrutura hierárquica semelhante a uma árvore para tomar decisões com base em perguntas sucessivas sobre os atributos de entrada. A ideia central é dividir o espaço de dados em regiões progressivamente mais homogêneas com relação à variável de saída, de modo a facilitar a classificação.

Segundo IBM (2025h), a árvore é construída a partir de um conjunto de dados de treinamento, utilizando um critério de divisão como entropia ou índice de *Gini*. Cada nó da árvore representa uma decisão baseada em um atributo, e cada ramo indica o resultado dessa decisão. O processo de divisão continua até que os dados em um nó sejam suficientemente homogêneos ou algum critério de parada seja atendido.

A Figura 1 ilustra a estrutura típica de uma Árvore de Decisão, composta pela raiz, nós internos e folhas.

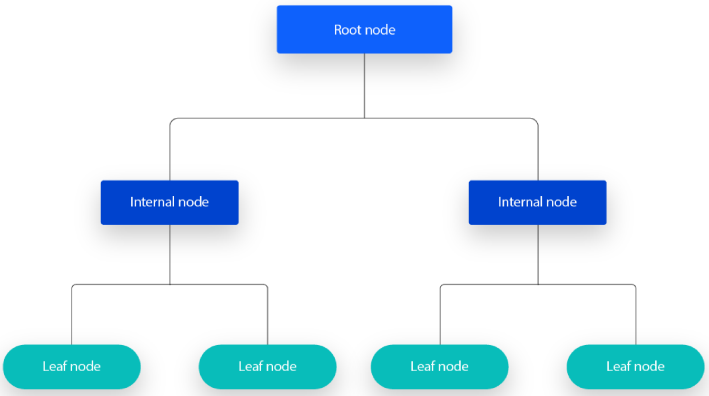

Figura 1 – Exemplo visual da estrutura de uma Árvore de Decisão, com raiz, nós internos e folhas.

Fonte: Adaptado de IBM (2025h).

Uma das principais vantagens da Árvore de Decisão é sua interpretabilidade. É fácil visualizar o caminho que o modelo segue para chegar a uma determinada previsão, o que torna essa técnica bastante útil em contextos onde a transparência da decisão é importante.

Neste trabalho, a Árvore de Decisão foi utilizada como um dos modelos para prever o vencedor do Oscar de Melhor Filme com base nas premiações anteriores. Apesar de sua simplicidade e interpretabilidade, o modelo demonstrou sensibilidade ao desbalanceamento dos dados, apresentando previsões extremas (como 0% ou 100%) e tendência ao sobreajuste, o que limitou seu desempenho em relação aos demais algoritmos testados.

2.2.3 Floresta Aleatória

A Floresta Aleatória (*Random Forest*) é uma técnica de aprendizado de máquina baseada em conjunto (*ensemble learning*), que combina múltiplas Árvores de Decisão para obter previsões mais robustas e acuradas. Em vez de construir uma única árvore, a Floresta Aleatória constrói várias árvores em subconjuntos aleatórios dos dados e combina suas previsões por votação (classificação) ou média (regressão).

Em acordo com IBM (2025c), a Floresta Aleatória introduz dois elementos-chave: o *bootstrap* (amostragem com reposição dos dados) e a seleção aleatória de atributos em cada divisão. Esses mecanismos promovem diversidade entre as árvores e reduzem o risco de sobreajuste, o que melhora significativamente a capacidade de generalização do modelo.

A estrutura geral desse algoritmo é ilustrada na Figura 2, onde observa-se a formação de múltiplas Árvores de Decisão treinadas em diferentes subconjuntos dos dados.

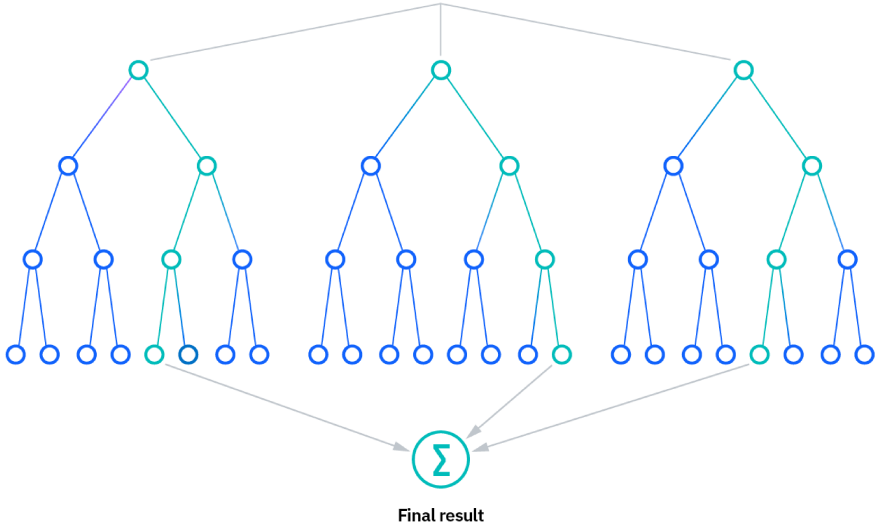

Figura 2 – Representação gráfica da estrutura de uma Floresta Aleatória, composta por múltiplas Árvores de Decisão independentes.

Fonte: Adaptado de IBM (2025c).

Além de oferecer excelente desempenho em muitos cenários, a Floresta Aleatória também fornece medidas de importância das variáveis, ajudando a identificar quais atributos mais influenciam as previsões.

Neste trabalho, a Floresta Aleatória foi o modelo que apresentou o melhor desempenho na predição do vencedor do Oscar de Melhor Filme. Sua robustez frente a ruídos e variações nos dados, combinada com uma boa capacidade de generalização, permitiu resultados consistentes tanto nas métricas avaliadas quanto na análise qualitativa das probabilidades atribuídas aos filmes concorrentes.

2.3 Avaliação dos Modelos Preditivos

A avaliação dos modelos preditivos foi realizada com base em métricas consagradas da literatura de aprendizado de máquina e estatística: Acurácia, Precisão, *Recall* e *F1-score*. Essas métricas oferecem diferentes perspectivas sobre o desempenho de classificadores binários, permitindo uma análise mais abrangente dos resultados. A seguir, são apresentadas suas definições com base em Filho (2020b), Filho (2020a):

- **Acurácia**: indica a proporção total de acertos – tanto verdadeiros positivos quanto verdadeiros negativos – em relação ao total de previsões realizadas. É uma métrica apropriada quando as classes estão balanceadas.
- **Precisão**: mede a proporção de acertos entre todas as previsões positivas feitas pelo modelo. É especialmente relevante em cenários onde falsos positivos são custosos ou indesejados.

- ***Recall***: avalia a capacidade do modelo em identificar corretamente os exemplos positivos reais. É útil quando se deseja minimizar falsos negativos.
- ***F1-score***: representa a média harmônica entre Precisão e *Recall*. É ideal quando há desequilíbrio entre as classes e é necessário equilibrar ambos os aspectos.

Essas métricas são particularmente relevantes neste trabalho, dado que a quantidade de filmes vencedores do Oscar é substancialmente menor do que a de não vencedores. O uso conjunto dessas medidas auxilia na identificação de modelos com maior poder de generalização.

Adicionalmente, utilizou-se a **matriz de confusão**, ilustrada na Figura 3, que oferece uma representação visual e quantitativa do desempenho dos modelos ao comparar as previsões com os resultados reais. A matriz é composta por quatro categorias:

- **Verdadeiro Positivo (TP)**: o modelo previu corretamente que o filme venceria o Oscar.
- **Falso Positivo (FP)**: o modelo previu que o filme venceria, mas ele não venceu.
- **Verdadeiro Negativo (TN)**: o modelo previu corretamente que o filme não venceria.
- **Falso Negativo (FN)**: o modelo previu que o filme não venceria, mas ele venceu.

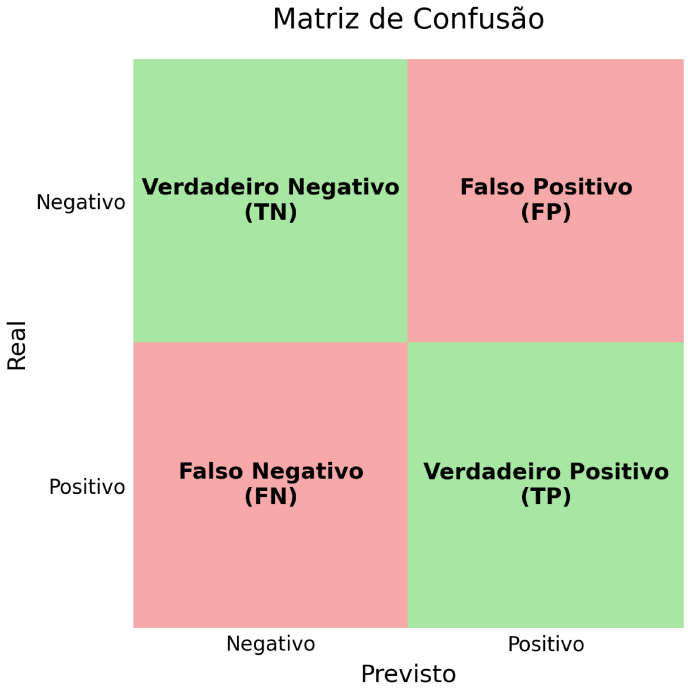

Figura 3 – Matriz de confusão ilustrando TP, FP, FN e TN.

Fonte: Elaborado pelo autor (2025).

2.4 Premiações Cinematográficas

Existe uma lógica cronológica na temporada de premiações que antecedem o Oscar (VILLAÇA, 2017). A temporada se inicia com o *Gotham Awards* e se estende até o Oscar, incluindo premiações como o *National Board of Review*, BAFTA, *Satellite Awards*, Globo de Ouro, e os prêmios das guildas (SAG, DGA, PGA), cujos membros muitas vezes coincidem com os votantes da Academia. Essa sequência serve como indicativo da receptividade dos filmes e justifica o uso dessas premiações como variáveis preditoras.

Cada uma das principais premiações utilizadas neste trabalho tem relevância específica dentro da indústria cinematográfica:

- **PGA (*Producers Guild of America*)**: premiação organizada pelos produtores de *Hollywood*, é considerada uma das principais precursoras do Oscar de Melhor Filme (PGA, 2025).
- **DGA (*Directors Guild of America*)**: premia diretores de cinema e televisão; o vencedor do DGA costuma vencer o Oscar de Melhor Direção (DGA, 2025).
- **SAG (*Screen Actors Guild*)**: premia elencos e atuações individuais. Por ter grande número de votantes também presentes na Academia, seu resultado impacta as categorias de atuação (LEGRAMANDI, 2025).
- **BAFTA (*British Academy of Film and Television Arts*)**: é o prêmio mais importante do cinema britânico. Também influencia o Oscar, especialmente nas categorias técnicas e internacionais (BAFTA, 2025).
- **Globo de Ouro (*Golden Globes*)**: premiação da crítica internacional com categorias separadas por gênero (drama e comédia/musical); é um dos primeiros grandes termômetros da temporada (G1, 2025b).
- ***Gotham Awards***: premiação voltada ao cinema independente. Costuma abrir a temporada e destacar obras artísticas de baixo orçamento (VILLAÇA, 2017).
- ***National Board of Review* (NBR)**: tradicional associação de críticos norte-americanos. Sua lista anual de melhores filmes influencia campanhas rumo ao Oscar (NBR, 2025).
- ***Satellite Awards***: organizado pela *International Press Academy*, contempla diversas categorias do cinema e televisão e serve como indicador complementar da temporada (G1, 2025a).

2.5 Trabalhos Relacionados

Diversos estudos vêm explorando a aplicação de técnicas de aprendizado de máquina em contextos culturais, como premiações cinematográficas, com o objetivo de identificar padrões recorrentes e prever resultados. Este trabalho se insere nesse campo, buscando contribuir com uma abordagem metodológica robusta e interpretável. A seguir, são apresentados alguns trabalhos relevantes da literatura que influenciaram o desenvolvimento desta pesquisa.

Oliveira (2023) analisou a influência das avaliações de críticos e usuários da plataforma *Metacritic* sobre os resultados das premiações do Oscar e do Globo de Ouro em 2023. Foram testados algoritmos como *Naive Bayes*, KNN e Floresta Aleatória, utilizando as notas médias como variáveis preditoras. O estudo concluiu que, apesar de alguma correlação aparente, as avaliações numéricas isoladas não foram suficientes para prever os vencedores, sugerindo a presença de fatores subjetivos ou externos que afetam os resultados.

Corrêa (2017) desenvolveu uma análise dos sentimentos expressos por usuários da rede social *Twitter* acerca dos filmes indicados ao Oscar de Melhor Filme em 2017. O estudo realizou coleta sistemática dos dados publicados no Twitter no período compreendido entre a divulgação dos indicados e a cerimônia de premiação. Utilizando técnicas de aprendizado supervisionado, especialmente o algoritmo *Naive Bayes* multinomial, Corrêa classificou os *tweets* quanto à polaridade (positiva, negativa ou neutra). Apesar dos resultados demonstrarem a viabilidade da abordagem para identificar preferências populares e realizar predições acerca do possível vencedor, não foram encontradas correlações estatísticas significativas entre os sentimentos expressos pelos usuários do *Twitter* e os resultados oficiais do Oscar daquele ano. O estudo ressaltou também uma predominância de manifestações positivas em detrimento das negativas nos *tweets* analisados, além de destacar a expressiva quantidade de comentários relacionados à premiação.

Franck e Wilson (2021) propuseram uma abordagem híbrida para prever os vencedores do Oscar, combinando regressão logística condicional com métodos bayesianos subjetivos. Essa combinação permitiu incorporar opiniões especializadas e informações qualitativas – como percepções da mídia – ao modelo estatístico, resultando em previsões mais alinhadas com os desfechos reais. O estudo é relevante porque consegue transformar aspectos subjetivos dos eventos culturais em algo que pode ser analisado com métodos estatísticos e científicos.

Já Lee et al. (2018) investigaram a previsão do sucesso de bilheteria de filmes utilizando algoritmos supervisionados como Regressão Logística, SVM, KNN, *Naive Bayes* e Floresta Aleatória. A pesquisa utilizou dados sobre orçamento, elenco, festivais e avaliações para alimentar os modelos e obteve bons índices de acurácia. Embora o foco tenha

sido o desempenho comercial e não premiações, o estudo reforça a aplicabilidade de técnicas supervisionadas em domínios culturais com múltiplas variáveis.

Masih e Ihsan (2019) também exploraram a relação entre o Oscar e outras indústrias cinematográficas ao desenvolver um modelo de previsão do sucesso de filmes de *Bollywood* com base em dados associados ao *Academy Awards*. Utilizando algoritmos de aprendizado de máquina como Regressão Logística, SVM, KNN e Árvores de Decisão, o estudo analisou como a influência de prêmios internacionais pode impactar o desempenho de filmes em contextos locais. A abordagem reforça a ideia de que padrões preditivos transcendem fronteiras culturais e que modelos supervisionados podem ser adaptados para diferentes realidades cinematográficas.

Esses trabalhos demonstram a diversidade de abordagens possíveis na aplicação de aprendizado de máquina ao cinema e às premiações culturais. A presente pesquisa contribui com esse campo ao propor um sistema preditivo focado na previsão do Oscar de Melhor Filme, utilizando dados estruturados de premiações relevantes e avaliando comparativamente diferentes algoritmos supervisionados.

3 Desenvolvimento e Resultados

Este capítulo apresenta todas as etapas práticas da construção do sistema preditivo do Oscar de Melhor Filme, integrando o desenvolvimento técnico com os resultados obtidos. O objetivo foi construir um *pipeline* desde a aquisição e preparação dos dados até a aplicação dos algoritmos e análise das previsões.

O repositório com todos os códigos-fonte, bases de dados utilizadas e materiais complementares deste trabalho está disponível em *GitHub*.

3.1 Linguagem e Bibliotecas Utilizadas

A linguagem principal foi o *Python* (versão 3.13), escolhida pela sua ampla adoção em ciência de dados e pela vasta disponibilidade de bibliotecas especializadas. Entre as principais bibliotecas utilizadas, destacam-se:

- ***Scikit-learn***: biblioteca robusta de aprendizado de máquina em *Python*, amplamente utilizada para tarefas de classificação, regressão e *clustering*. Possui algoritmos otimizados, ferramentas de validação cruzada e métricas de avaliação (SCIKIT-LEARN DEVELOPERS, 2025).
- ***Pandas***: biblioteca essencial para manipulação de dados tabulares. Permite leitura de arquivos CSV, tratamento de dados faltantes, agrupamentos, agregações e transformações eficientes (THE PANDAS DEVELOPMENT TEAM, 2025).
- ***NumPy***: fornece estruturas e funções para trabalhar com *arrays* multidimensionais e operações matemáticas de alto desempenho (NUMPY DEVELOPERS, 2025).
- ***Matplotlib*** e ***Seaborn***: bibliotecas de visualização de dados. Enquanto o *Matplotlib* é mais versátil e personalizável, o *Seaborn* facilita a criação de gráficos estatísticos com visual elegante (MATPLOTLIB COMMUNITY, 2025; WASKOM, 2025).
- ***BeautifulSoup*** e ***Requests***: utilizadas para coleta automatizada de dados via *web scraping*. *Requests* realiza as requisições HTTP, enquanto o *BeautifulSoup* faz o *parse* do HTML para extrair as informações desejadas (RICHARDSON, 2025; REITZ, 2025).
- ***Joblib***: biblioteca utilizada para serialização e armazenamento eficiente de objetos *Python*, como modelos treinados de *machine learning* (JOBLIB, 2025).

Essas bibliotecas contribuíram significativamente para a automação das etapas do pipeline, desde a aquisição de dados até a visualização dos resultados preditivos.

3.1.1 Web Scraping e Pré-Processamento de Dados

A coleta dos dados foi realizada por meio da técnica de *web scraping*, conforme descrito por Graciano e Ramalho (2023), utilizando as bibliotecas `requests` e `BeautifulSoup`, ambas amplamente empregadas na linguagem *Python* para a extração de informações em páginas web. As páginas-alvo incluíram artigos da Wikipédia e seções específicas do IMDb, que disponibilizam tabelas organizadas com os vencedores de premiações cinematográficas relevantes.

- Para cada premiação, foram extraídos os filmes vencedores no período de 2004 a 2023.
- Cada entrada foi vinculada ao respectivo ano de premiação e ao nome do filme.
- A partir dessas informações, foram criadas variáveis binárias indicando se o filme venceu determinada premiação naquele ano.

Após a coleta, foi realizada uma etapa de pré-processamento dos dados com base nas boas práticas recomendadas por Batista (2003), que envolveu a normalização de textos, padronização dos nomes dos filmes, tratamento de valores ausentes e identificação de inconsistências ou ruídos nos registros.

3.1.2 Tratamento e Consolidação dos Dados

Os dados brutos extraídos das fontes online apresentaram algumas inconsistências, como:

- Diferenças na grafia de nomes dos filmes.
- Premiações com múltiplos vencedores (empates).
- Variações na estrutura das tabelas em diferentes anos.

Para lidar com esses problemas, foi aplicada uma série de transformações:

- Normalização dos nomes dos filmes para facilitar a junção dos dados. A normalização envolveu converter todos os títulos dos filmes para letras minúsculas, remover caracteres especiais e padronizar espaços em branco, garantindo que diferentes grafias fossem corretamente reconhecidas como o mesmo filme

- Exclusão de documentários, curtas e animações.

O resultado foi uma tabela unificada com os seguintes campos principais:

- `"Filme"`
- `"Ano"`
- `"Bafta"`
- `"DGA"`
- `"GloboDeOuro"`
- `"Gotham"`
- `"NationalBoard"`
- `"PGA"`
- `"SAG"`
- `"Satellite"`
- `"Oscar"` (variável-alvo)

3.2 Análise Exploratória dos Dados

Antes da modelagem, foi realizada uma análise exploratória com o objetivo de:

- Verificar correlações entre as premiações e o Oscar.
- Entender a distribuição dos dados.
- Identificar padrões temporais e de recorrência.

3.2.1 Análise de Correlação

A análise de correlação é uma etapa essencial na compreensão da relação entre as variáveis preditoras — neste caso, as premiações anteriores — e o resultado do Oscar de Melhor Filme. Neste trabalho, foram utilizadas duas abordagens estatísticas complementares: a correlação de *Pearson* (Figura 4) e a correlação de *Spearman* (Figura 5). Ambas geraram gráficos de associação para auxiliar na visualização dos padrões entre os dados.

3.2.1.1 Correlação de *Pearson*

A correlação de *Pearson* mede a força e a direção da relação linear entre duas variáveis contínuas. Essa medida é sensível a valores extremos (*outliers*). *Outliers* são observações que se desviam significativamente do padrão geral de um conjunto de dados, podendo indicar erros de coleta ou fenômenos genuínos e importantes (IBM, 2025a). e pressupõe que as variáveis tenham distribuição normal e uma relação linear. O coeficiente de *Pearson* varia de -1 a 1, onde 1 indica correlação positiva perfeita, -1 correlação negativa perfeita e 0 nenhuma correlação (MIOT, 2018).

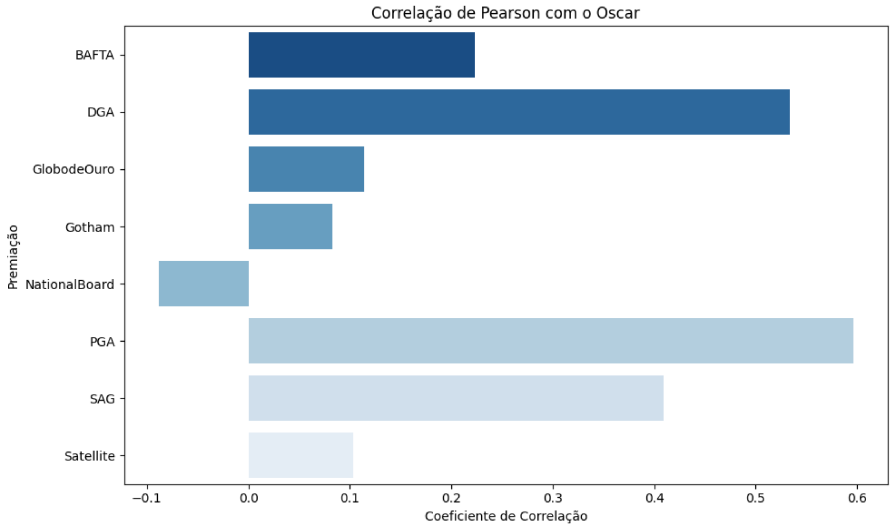

Figura 4 – Correlação de *Pearson* entre premiações e o Oscar

Fonte: Elaborado pelo autor (2025).

3.2.1.2 Correlação de *Spearman*

A correlação de *Spearman*, por sua vez, é uma medida não paramétrica que avalia a associação entre duas variáveis com base em suas classificações (*ranks*). É mais robusta contra *outliers* e não exige que as variáveis apresentem distribuição normal ou relação linear. O coeficiente de *Spearman* também varia de -1 a 1 e é ideal para capturar relações monotônicas, em que os valores sobem ou descem de forma consistente, mas não necessariamente linear (MIOT, 2018).

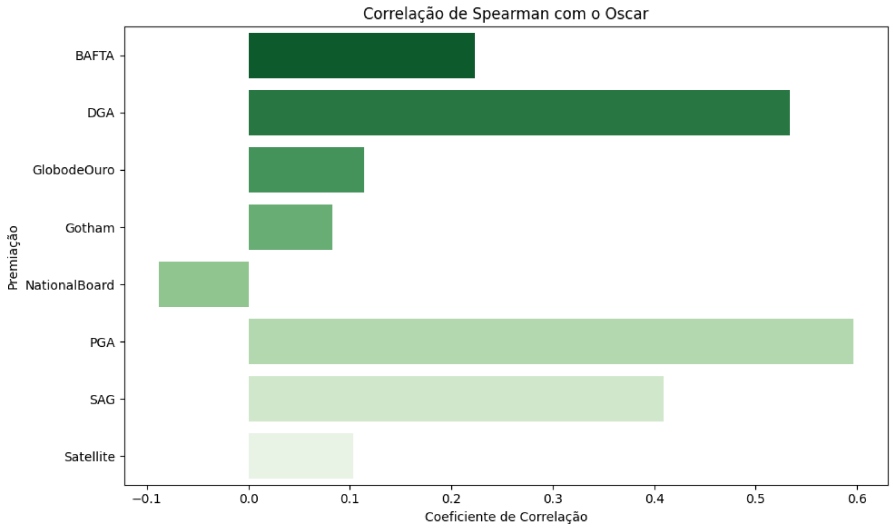

Figura 5 – Correlação de *Spearman* entre premiações e o Oscar

Fonte: Elaborado pelo autor (2025).

Essa etapa foi fundamental para identificar premiações específicas que possuem forte associação com o Oscar, validando sua inclusão como variáveis preditivas essenciais para os modelos utilizados posteriormente.

3.2.2 Análise de Associação com o Oscar: Teste Qui-Quadrado

O teste de independência Qui-Quadrado é uma técnica estatística amplamente utilizada para verificar a existência de associação significativa entre duas variáveis categóricas. Ele compara a frequência observada em uma tabela de contingência com a frequência esperada sob a hipótese de independência. Caso a discrepância entre esses valores seja estatisticamente relevante, rejeita-se a hipótese nula, indicando que há associação entre as variáveis (ANUNCIAÇÃO, 2021).

Neste trabalho, o teste foi aplicado para avaliar a relação entre a vitória nas principais premiações cinematográficas (como BAFTA, DGA, PGA, entre outras) e a conquista do Oscar de Melhor Filme. Para cada premiação, foi construída uma tabela de contingência com os dados históricos de 2004 a 2023, e os valores de Qui-Quadrado e p-valor foram calculados. Resultados com p-valor inferior a 0,05 foram considerados estatisticamente significativos.

Tabela 1 – Resultados do Teste Qui-Quadrado entre cada premiação e o Oscar

Premiação	**Valor Qui-Quadrado**	**p-valor**	**Significância**
BAFTA	4.32	0.038	Significativo
DGA	6.17	0.013	Significativo
PGA	8.49	0.004	Significativo
SAG	2.88	0.089	Não significativo
Globo de Ouro	1.91	0.166	Não significativo
Gotham	0.73	0.392	Não significativo
NBR	5.44	0.019	Significativo
Satellite	0.54	0.462	Não significativo

Fonte: Elaborado pelo autor (2025).

Os resultados indicam que algumas premiações, como PGA, DGA e NBR, apresentam associação estatística significativa com a vitória no Oscar de Melhor Filme, fortalecendo sua inclusão como variáveis preditoras no modelo desenvolvido. Essa etapa complementa as análises de correlação realizadas previamente, oferecendo uma evidência estatística adicional para a seleção de atributos. Posteriormente, a análise de coocorrência foi utilizada para verificar a frequência com que essas premiações coincidem na prática com os vencedores do Oscar, consolidando ainda mais a relevância dessas premiações como indicadores.

3.2.3 Análise de Coocorrência de Premiações

A análise de coocorrência é uma técnica estatística utilizada para identificar padrões frequentes de eventos que ocorrem simultaneamente em um conjunto de dados. Segundo a IBM (2025b), as regras de coocorrência permitem examinar pares ou conjuntos de itens que aparecem juntos com frequência, oferecendo uma perspectiva valiosa sobre agrupamentos ou comportamentos coletivos.

Neste trabalho, a coocorrência foi aplicada para compreender com que frequência premiações importantes antecedem a vitória no Oscar de Melhor Filme. A métrica utilizada foi a proporção de edições em que um mesmo filme venceu tanto uma determinada premiação quanto o Oscar, fornecendo uma medida intuitiva da sobreposição histórica entre os prêmios.

Os resultados dessa análise estão ilustrados na Figura 6, que apresenta as taxas de coocorrência entre as principais premiações e o Oscar de Melhor Filme.

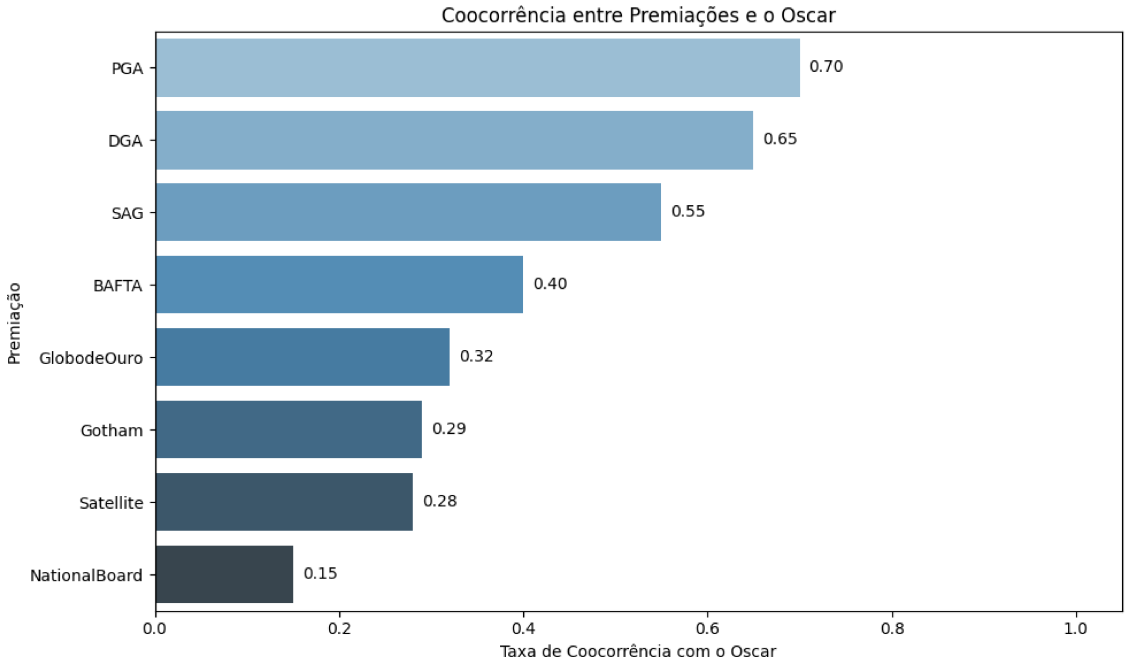

Figura 6 – Taxa de coocorrência entre as principais premiações e o Oscar de Melhor Filme.

Fonte: Elaborado pelo autor (2025).

Como se observa na Figura 6, as premiações com maior taxa de coocorrência com o Oscar foram:

- **PGA (0.70)**
- **DGA (0.65)**
- **SAG (0.55)**

Esse alinhamento pode ser parcialmente explicado pela sobreposição significativa entre os membros votantes dessas instituições e os votantes da Academia, reforçando o valor preditivo dessas premiações para o Oscar.

Por outro lado, premiações como o ***National Board of Review (NBR)***, com taxa de coocorrência de apenas 0,15, apresentaram baixo alinhamento com os vencedores do Oscar. Isso pode ocorrer porque o NBR tende a destacar filmes independentes ou artísticos no início da temporada de premiações, antes do fortalecimento do "*momentum*" das campanhas de estúdio, o que faz com que esses títulos eventualmente percam visibilidade até a cerimônia principal.

Essa abordagem não analisa a causalidade, mas sim a frequência conjunta de eventos, ajudando a identificar grupos de influência — conjuntos de premiações que se reforçam mutuamente na temporada cinematográfica. Tais padrões contribuem para a

seleção de atributos relevantes nos modelos de classificação utilizados na predição do vencedor do Oscar.

3.3 Construção dos Modelos Preditivos

3.3.1 Definição do Problema

O problema foi tratado como uma tarefa de **classificação binária**, em que a variável-alvo indica se um filme venceu (1) ou não venceu (0) o Oscar de Melhor Filme.

3.3.2 Separação dos Dados

Os dados foram divididos da seguinte forma:

- Dados de 2004 a 2023: utilizados para treinamento e validação dos modelos.
- Dados de 2024: utilizados posteriormente como conjunto de teste para prever o Oscar 2025.

3.3.3 Modelos Aplicados

Três modelos de classificação foram treinados e comparados:

1. **Regressão Logística**
2. **Árvore de Decisão**
3. **Floresta Aleatória**

3.3.4 Ferramentas e Bibliotecas

O códigos foram implementados com *Python* na aplicação *Pycharm* e utilizou as bibliotecas `pandas`, `scikit-learn`, `matplotlib`, `seaborn` e `numpy`, em ambiente *Python* 3.13, para realizar todas as tarefas do *pipeline*, desde a preparação dos dados até as visualizações.

3.3.5 Treinamento dos Modelos

O treinamento dos modelos de classificação consistiu em três etapas principais: separação dos dados, ajuste dos parâmetros e avaliação do desempenho. O conjunto de dados históricos, que compreendia informações sobre as premiações de 2004 a 2023, foi dividido entre dados de treinamento e dados de validação, utilizando a técnica de **validação *holdout***, com **80% dos dados destinados ao treino** e **20% reservados**

para validação. Cada modelo seguiu a seguinte lógica geral de treinamento mostrada no Algoritmo 1.

Data: Base de dados com filmes e premiações (2004-2023)
Result: Modelo treinado e pronto para prever o Oscar 2025
Separar dados em conjunto de treino e validação;
Inicializar o modelo escolhido;
Treinar o modelo usando o conjunto de treino;
Avaliar o desempenho no conjunto de validação usando métricas como acurácia e F1-Score;
Salvar o modelo treinado para realizar a previsão;

Algoritmo 1: Processo Geral de Treinamento dos Modelos

3.3.5.1 Treinamento da Regressão Logística

A Regressão Logística foi ajustada para prever a probabilidade de vitória no Oscar a partir de combinações lineares dos atributos (vitórias anteriores). O modelo buscou encontrar os coeficientes que maximizassem a função de verossimilhança dos dados observados.

Data: Conjunto de dados rotulados
Result: Coeficientes ajustados para prever probabilidades
Inicializar coeficientes aleatoriamente;
while *não convergência* **do**
    Atualizar coeficientes utilizando máxima verossimilhança;
end
Calcular as probabilidades estimadas para cada instância;

Algoritmo 2: Treinamento de Regressão Logística

3.3.5.2 Treinamento da Árvore de Decisão

A Árvore de Decisão foi treinada para segmentar o conjunto de dados em grupos homogêneos com base em divisões sucessivas dos atributos. O critério de divisão utilizado foi o índice de *Gini*. O índice de *Gini* é uma medida usada para avaliar a pureza de uma divisão em algoritmos de Árvore de Decisão. Quanto menor o valor de Gini, mais homogêneos são os dados em um nó (IBM, 2025d).

Data: Conjunto de dados rotulados
Result: Árvore construída para classificação
Inicializar a árvore vazia;
while *existirem nós passíveis de divisão* **do**
    Selecionar o atributo que melhor separa os dados (índice de *Gini*);
    Dividir o nó conforme o atributo selecionado;
end

Algoritmo 3: Treinamento da Árvore de Decisão

3.3.5.3 Treinamento da Floresta Aleatória

A Floresta Aleatória construiu múltiplas Árvores de Decisão em subconjuntos aleatórios dos dados para reduzir o sobreajuste e melhorar a generalização.

Data: Conjunto de dados rotulados
Result: Conjunto de árvores votantes
for *cada árvore na floresta* **do**
    Selecionar amostra aleatória com reposição (bootstrap);
    Construir uma Árvore de Decisão com seleção aleatória de atributos;
end
Para novas previsões, combinar o voto de todas as árvores;

Algoritmo 4: Treinamento da Floresta Aleatória

3.3.6 Matrizes de Confusão por Modelo

As matrizes de confusão dos modelos, apresentadas nas Figuras 7, 8 e 9, foram utilizadas para analisar detalhadamente o desempenho dos algoritmos. Elas permitem observar a quantidade de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativo

- **Regressão Logística**: apresentou boa taxa de acertos, com tendência conservadora e poucos falsos positivos.
- **Árvore de Decisão**: teve mais falsos positivos, indicando maior sensibilidade, mas menor precisão.
- **Floresta Aleatória**: desempenho equilibrado, com boa separação das classes e menos erros extremos.

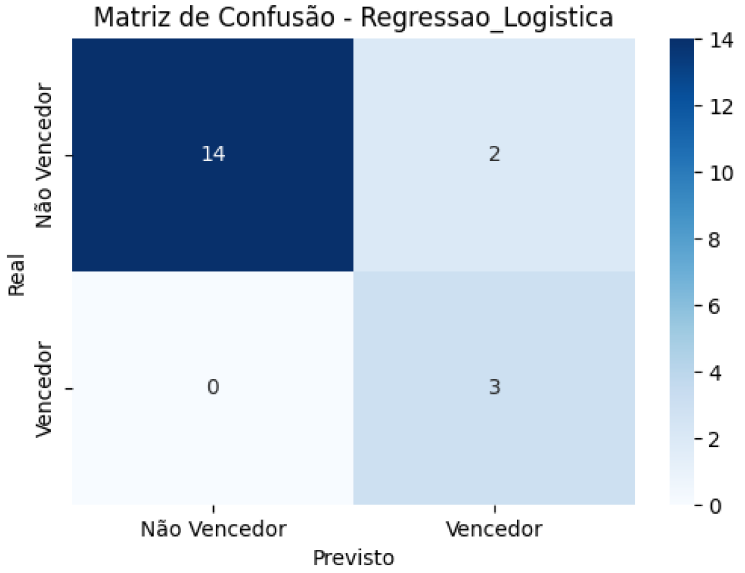

Figura 7 – Matriz de Confusão - Regressão Logística

Fonte: Elaborado pelo autor (2025).

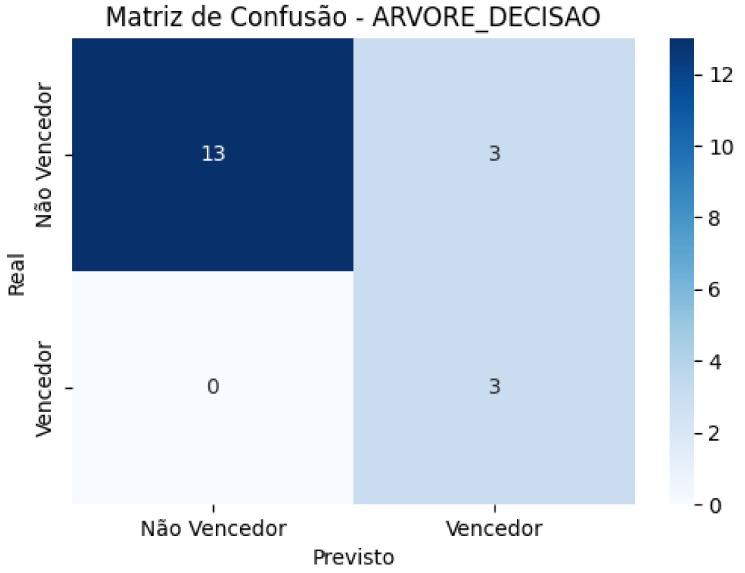

Figura 8 – Matriz de Confusão - Árvore de Decisão

Fonte: Elaborado pelo autor (2025).

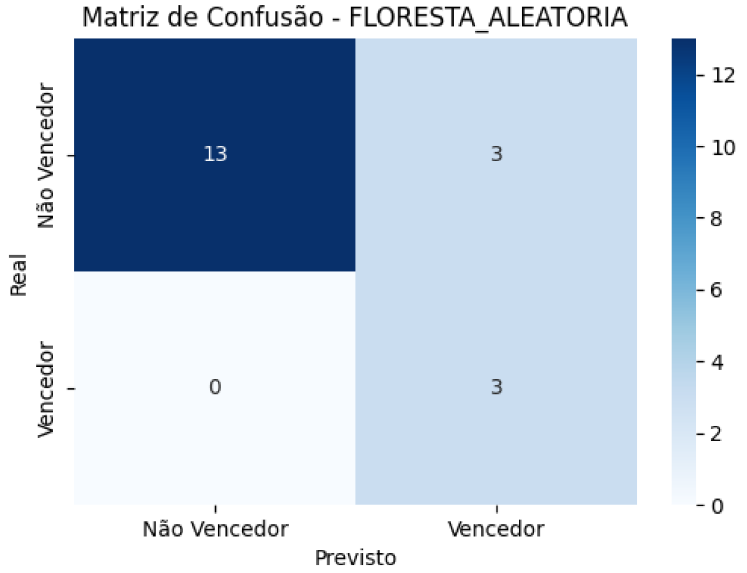

Figura 9 – Matriz de Confusão - Floresta Aleatória

Fonte: Elaborado pelo autor (2025).

Neste contexto específico, modelos que apresentam menor incidência de falsos positivos são preferíveis. Isso ocorre porque minimizar previsões incorretas de vitórias evita expectativas falsas, especialmente relevante em cenários como análises críticas, estratégias de *marketing* ou apostas, nos quais erros dessa natureza poderiam causar impactos negativos significativos.

3.3.7 Métricas de Avaliação dos Modelos Preditivos

As métricas apresentadas na Tabela 2 foram escolhidas com o intuito de avaliar o desempenho dos modelos sob diferentes perspectivas. A seguir, detalha-se brevemente cada uma delas, destacando especificamente a razão pela qual o *recall* atingiu o valor máximo (1,0) no contexto deste estudo:

Tabela 2 – Métricas de Avaliação dos Modelos Preditivos

Modelo	**Acurácia**	**Precisão**	***Recall***	***F1-score***
Regressão Logística	0,8947	0,6000	1,0000	0,7500
Árvore de Decisão	0,8421	0,5000	1,0000	0,6667
Floresta Aleatória	0,8421	0,5000	1,0000	0,6667

Fonte: Elaborado pelo autor (2025).

- **Acurácia**: Indica a proporção geral de acertos (previsões corretas) realizadas pelo modelo em relação ao total de previsões feitas, refletindo a eficiência global na classificação.

- **Precisão (Classe 1)**: Avalia especificamente a capacidade do modelo de prever corretamente a classe positiva (vitória no Oscar), sendo fundamental quando se deseja evitar falsos positivos.
- ***Recall* (Classe 1)**: Mede a habilidade do modelo em identificar corretamente todos os casos reais positivos (verdadeiros vencedores), minimizando falsos negativos. No contexto específico deste trabalho, o valor máximo do *recall* (igual a 1,0) indica que todos os filmes que realmente venceram o Oscar foram identificados corretamente pelo modelo. Isso ocorre pois há somente um vencedor real na premiação, resultando em zero falsos negativos e, consequentemente, o recall atinge seu valor máximo possível.
- ***F1-score* (Classe 1)**: Corresponde à média harmônica entre precisão e recall, oferecendo uma métrica balanceada especialmente útil em situações onde as classes estão desequilibradas, como é o caso deste estudo.

Essas métricas fornecem uma visão abrangente sobre a capacidade preditiva dos modelos, destacando suas forças e fraquezas em diferentes contextos de aplicação. Apesar dos bons resultados obtidos, as limitações identificadas destacam oportunidades importantes para estudos futuros, como a inclusão de variáveis qualitativas relacionadas ao impacto cultural dos filmes, estratégias promocionais, ou ainda a aplicação de modelos mais complexos, como redes neurais profundas.

3.3.8 Distribuição das Probabilidades de Vitória

Foi realizada uma análise estatística da distribuição das probabilidades geradas por cada modelo, conforme ilustrado nas Figuras 10 e 11.

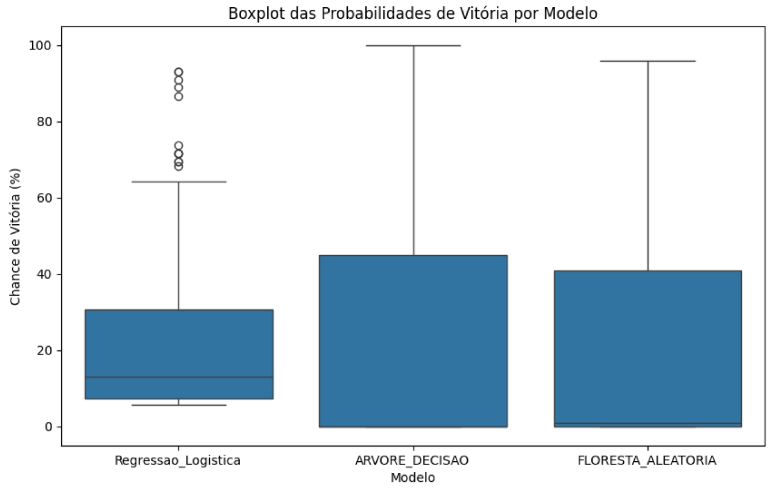

Figura 10 – Boxplot das probabilidades de vitória por modelo

Fonte: Elaborado pelo autor (2025).

- A Regressão Logística apresenta valores medianos com muitos outliers.
- A Árvore de Decisão apresenta previsões extremas (0% ou 100%).
- A Floresta Aleatória gera uma distribuição mais suave, com boa variabilidade.

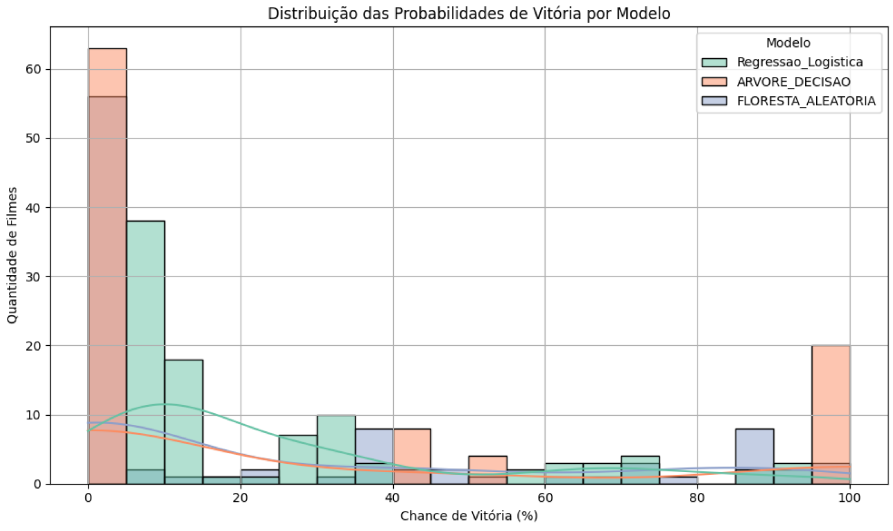

Figura 11 – Distribuição das probabilidades de vitória por modelo

Fonte: Elaborado pelo autor (2025).

As distribuições reforçam que poucos filmes são favoritos absolutos, enquanto a maioria tem chances baixas ou intermediárias.

3.3.9 Top 10 Filmes com Maior Probabilidade de Vitória

A convergência entre os modelos nos principais títulos, conforme ilustrado na Figura 12, evidencia robustez dos algoritmos.

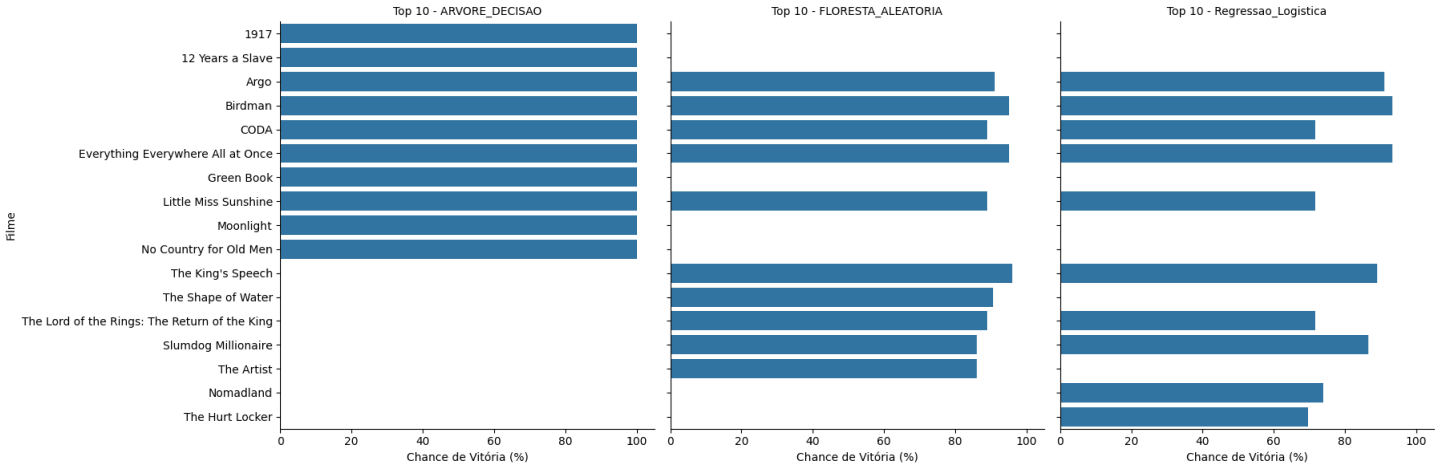

Figura 12 – Top 10 filmes com maior chance de vitória segundo cada modelo

Fonte: Elaborado pelo autor (2025).

Com os modelos devidamente treinados e avaliados em termos de desempenho geral, foi possível aplicá-los ao conjunto mais recente de dados, visando testar sua capacidade preditiva para o Oscar de 2025.

3.4 Avaliação dos Modelos com Dados de 2025

Após o treinamento com os dados históricos de 2004 a 2023, os modelos foram aplicados ao conjunto de teste que representa a temporada do Oscar de 2025, utilizando dados das premiações ocorridas em 2024. A seguir, são apresentados os resultados obtidos.

3.4.1 Previsão de Vencedor do Oscar 2025

A Tabela 3 apresenta os dados de entrada utilizados para a previsão do Oscar 2025, com a indicação binária de vitória (1) ou não vitória (0) dos filmes nas principais premiações da temporada.

Tabela 3 – Atributos utilizados nos modelos preditivos: filmes indicados ao Oscar 2025 (entradas), premiações prévias recebidas (atributos) variando entre 0 (não vitória) e 1 (vitória).

Filme	**BAFTA**	**DGA**	**Globo de Ouro**	***Gotham***	***National Board***	**PGA**	**SAG**	***Satellite***
Anora	0	1	0	0	0	1	0	0
Emilia Pérez	0	0	0	0	0	0	0	0
Um Completo Desconhecido	0	0	0	0	0	0	0	0
Conclave	1	0	0	0	0	0	1	0
Nickel Boys	0	0	0	0	0	0	0	0
Ainda Estou Aqui	0	0	0	0	0	0	0	0
A Substância	0	0	0	0	0	0	0	0
Duna 2	0	0	0	0	0	0	0	0
Wicked	0	0	0	0	1	0	0	0
O Brutalista	0	0	1	0	0	0	0	1

Fonte: Elaborado pelo autor (2025).

Os três modelos foram utilizados para prever as chances de vitória dos filmes elegíveis ao Oscar de Melhor Filme em 2025. A Tabela 4 apresenta as probabilidades estimadas por cada modelo, bem como a classificação binária (Sim/Não) quanto à predição de vitória.

Tabela 4 – Previsão de Vencedor do Oscar 2025 por Modelo

Filme	**RL (%)**	**AD (%)**	**FA (%)**	**RL - V**	**AD - V**	**FA - V**
Anora	69,5	100,0	90,5	Sim	Sim	Sim
Conclave	25,4	0,0	29,6	Não	Não	Não
O Brutalista	9,0	0,0	5,0	Não	Não	Não
Emilia Pérez	9,0	0,0	11,3	Não	Não	Não
Um Completo Desconhecido	8,7	0,0	11,3	Não	Não	Não
Nickel Boys	8,7	0,0	11,3	Não	Não	Não
Ainda Estou Aqui	8,7	0,0	11,3	Não	Não	Não
A Substância	8,7	0,0	11,3	Não	Não	Não
Duna 2	8,7	0,0	11,3	Não	Não	Não
Wicked	6,9	0,0	0,0	Não	Não	Não

Fonte: Elaborado pelo autor (2025).

3.4.2 Análise Gráfica das Predições

As probabilidades atribuídas por cada modelo, bem como a média das previsões para o Oscar de 2025, são apresentadas nas Figuras 13, 14, 15 e 16.

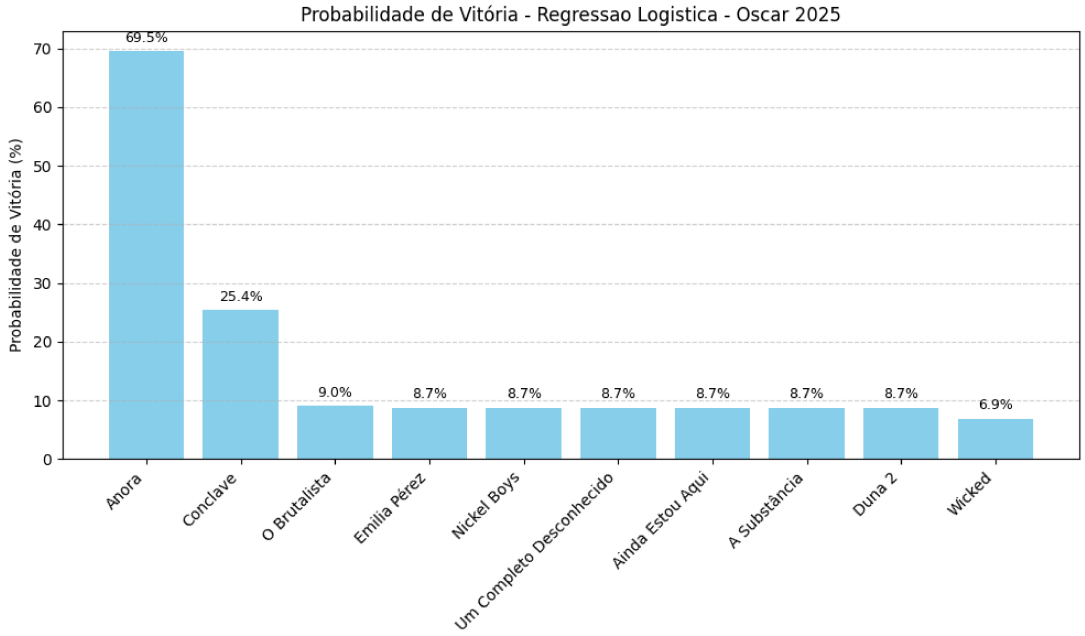

Figura 13 – Probabilidade de Vitória - Regressão Logística - Oscar 2025

Fonte: Elaborado pelo autor (2025).

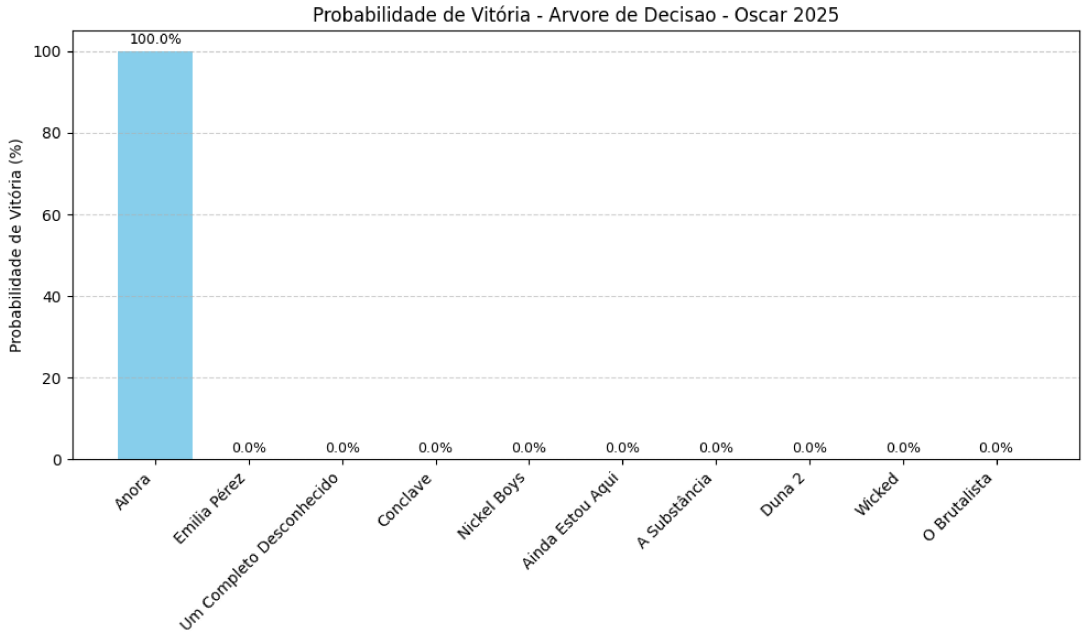

Figura 14 – Probabilidade de Vitória - Árvore de Decisão - Oscar 2025

Fonte: Elaborado pelo autor (2025).

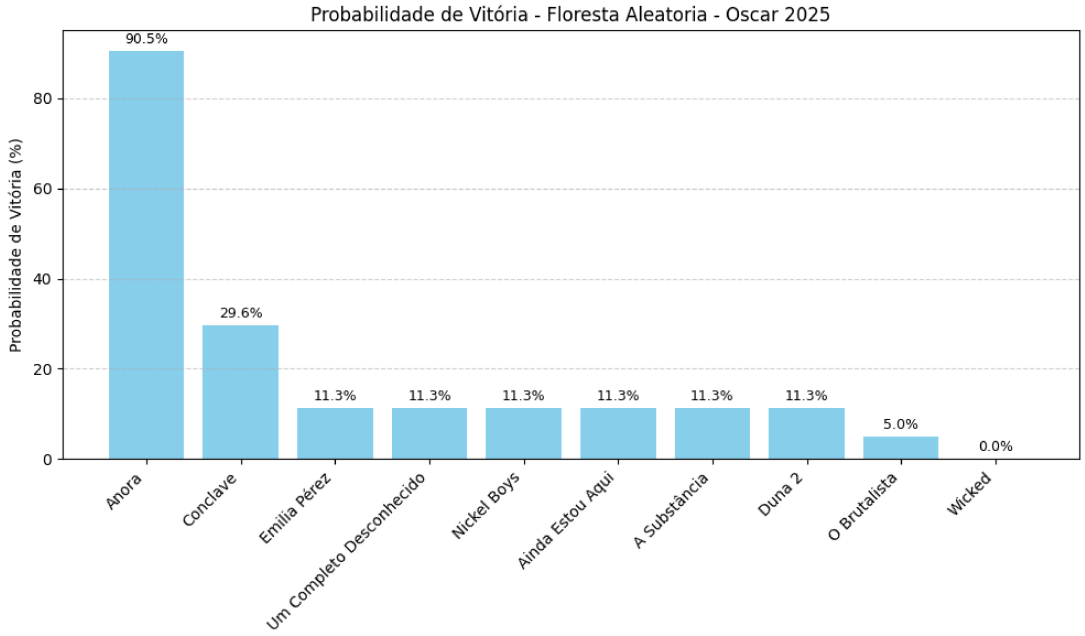

Figura 15 – Probabilidade de Vitória - Floresta Aleatória - Oscar 2025

Fonte: Elaborado pelo autor (2025).

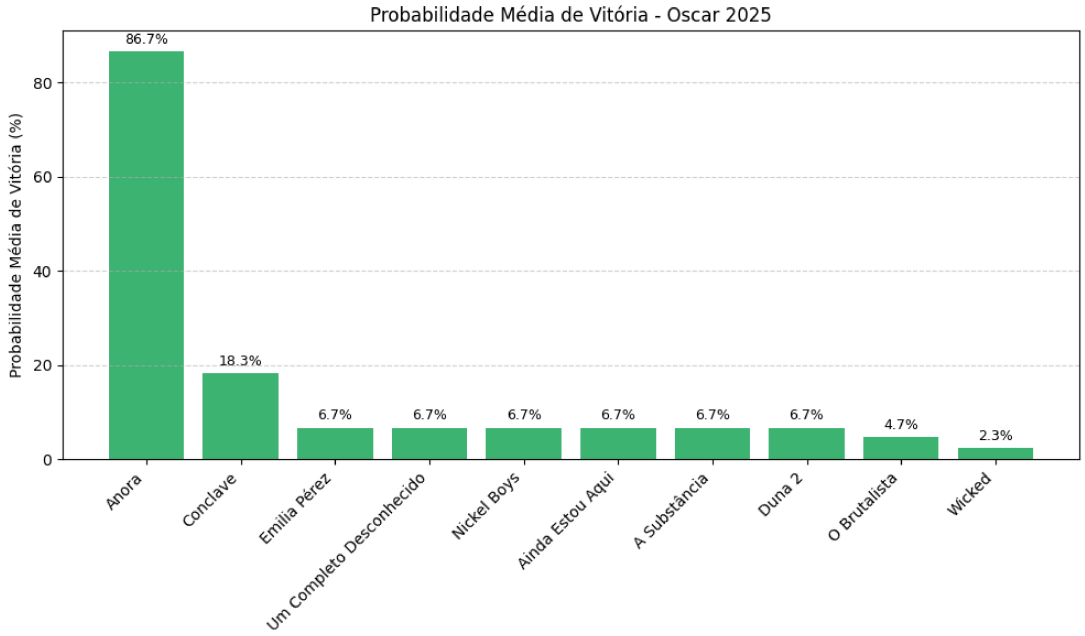

Figura 16 – Probabilidade Média de Vitória - Oscar 2025

Fonte: Elaborado pelo autor (2025).

3.4.3 Discussão dos Resultados

Os três modelos apontaram o filme *Anora* como o mais provável vencedor, com destaque para a Árvore de Decisão, que atribuiu 100% de chance de vitória, e para a Flo-

resta Aleatória, que atribuiu aproximadamente 90,5%. A convergência entre os modelos, mesmo com algumas variações nos percentuais, reforça a confiança na previsão.

Observou-se que a Árvore de Decisão, como esperado, tendeu a extremar os valores, enquanto a Regressão Logística e a Floresta Aleatória distribuíram melhor as probabilidades, considerando incertezas naturais. A Floresta Aleatória se mostrou o modelo mais equilibrado, distribuindo as chances de forma mais realista, sem excessos de otimismo ou pessimismo.

Apesar de a Regressão Logística ter apresentado boa precisão histórica, a Floresta Aleatória demonstrou melhor robustez para este caso específico, oferecendo uma distribuição de probabilidades mais coerente com a natureza incerta das premiações.

A utilização da média das probabilidades entre os modelos visou reduzir vieses individuais e fornecer uma previsão ainda mais robusta, confirmando o protagonismo de *Anora* como o principal favorito ao Oscar de Melhor Filme em 2025.

3.4.4 Limitações Observadas

Apesar do desempenho satisfatório, algumas limitações precisam ser destacadas:

- O número reduzido de instâncias históricas (cerca de 20 por ano) limita a capacidade de generalização.
- A ausência de variáveis subjetivas como "*momentum*"[1] de campanha, campanhas de *marketing* ou *hype* nas redes sociais.
- Algumas premiações, como SAG e Globo de Ouro, não mostraram associação estatística significativa com o Oscar.

Essas limitações destacam oportunidades importantes para estudos futuros, como a inclusão de variáveis qualitativas relacionadas ao impacto cultural dos filmes, estratégias promocionais, ou ainda a aplicação de modelos mais complexos, como redes neurais profundas. Apesar disso, os resultados gerais foram encorajadores e sustentam o potencial do aprendizado de máquina na análise de tendências culturais. A seguir, são apresentadas as considerações finais deste capítulo, consolidando os achados e implicações da abordagem adotada.

[1] No contexto das premiações cinematográficas, "*momentum*"refere-se ao crescimento progressivo da atenção, apoio e entusiasmo que um filme recebe ao longo da temporada de premiações. Esse fenômeno pode ser impulsionado por vitórias anteriores, estratégias de *marketing*, repercussão na mídia e comentários positivos em redes sociais, aumentando suas chances de sucesso em premiações como o Oscar.

3.5 Considerações Finais

Com base nos experimentos realizados, conclui-se que a utilização de algoritmos supervisionados para prever os vencedores do Oscar é uma abordagem viável, especialmente quando ancorada em premiações historicamente relevantes.

Os resultados obtidos em 2025 demonstram que é possível antecipar tendências da Academia com base em dados públicos e estruturados, utilizando aprendizado de máquina como ferramenta de suporte à análise cultural.

4 Conclusão

Este trabalho teve como objetivo principal investigar a viabilidade da aplicação de técnicas de aprendizado de máquina supervisionado na previsão do vencedor do Oscar de Melhor Filme, utilizando dados históricos de premiações anteriores como variáveis preditoras. A proposta buscou ir além da simples construção de um modelo preditivo, promovendo uma reflexão sobre o papel da ciência de dados em contextos culturais complexos e subjetivos.

O processo metodológico adotado foi dividido em múltiplas etapas bem definidas. Inicialmente, foi realizada a coleta automatizada dos dados através de técnicas de web scraping, abrangendo informações de vencedores de premiações relevantes como BAFTA, DGA, PGA, SAG, entre outras, no período de 2004 a 2023. Em seguida, os dados passaram por um rigoroso processo de tratamento, com padronização, limpeza e consolidação em uma base estruturada.

Na etapa de análise exploratória, foram utilizadas técnicas estatísticas como correlação de *Pearson* e *Spearman*, além do teste de independência qui-quadrado, para avaliar a força das associações entre as premiações e o Oscar. Esses testes revelaram que prêmios como o PGA e o DGA possuem elevada coocorrência com o Oscar, sugerindo seu potencial como variáveis preditoras confiáveis.

Com os dados tratados e analisados, foram implementados três algoritmos de classificação supervisionada: Regressão Logística, Árvore de Decisão e Floresta Aleatória. A Floresta Aleatória destacou-se como o modelo com melhor desempenho geral, apresentando maior estabilidade e poder preditivo, especialmente em contextos de baixa amostragem. A convergência entre os modelos quanto à predição do vencedor do Oscar 2025 (filme "*Anora*") reforçou a robustez do sistema desenvolvido.

As visualizações gráficas, como matrizes de confusão, *boxplots* e gráficos de distribuição de probabilidades, complementaram a análise quantitativa, permitindo uma interpretação intuitiva dos resultados. Essas ferramentas facilitaram a comunicação dos achados e contribuíram para tornar o sistema acessível a públicos diversos.

Portanto, conclui-se que o uso de aprendizado de máquina supervisionado pode, sim, ser uma abordagem eficaz para prever eventos culturais como o Oscar, desde que fundamentada em dados históricos relevantes e técnicas analíticas robustas. O trabalho demonstrou que há padrões quantificáveis mesmo em processos subjetivos como premiações artísticas, oferecendo novas possibilidades para estudos futuros.

Apesar dos avanços obtidos, este estudo apresenta limitações que abrem caminho

para investigações futuras. Entre elas, destacam-se:

- A ampliação da base de dados, incorporando um maior número de edições históricas do Oscar e premiações internacionais;
- A inclusão de variáveis qualitativas, como menções em redes sociais, análises da crítica especializada, campanhas promocionais e sentimento do público;
- A aplicação de modelos mais sofisticados, como redes neurais profundas, para avaliar ganhos em capacidade preditiva e sensibilidade a padrões não lineares;
- O desenvolvimento de uma plataforma interativa pública que permita acompanhar, em tempo real, as previsões baseadas na temporada de premiações;
- O uso de técnicas de aprendizado não supervisionado para identificar agrupamentos ou perfis de filmes vencedores, enriquecendo a análise preditiva.

Essas possibilidades reforçam o potencial do cruzamento entre estatística, aprendizado de máquina e análise cultural, e convidam à continuidade de estudos interdisciplinares que explorem o uso de dados em fenômenos sociais complexos.

Referências

AMAZON WEB SERVICES. **Diferenças entre aprendizado supervisionado e não supervisionado**. 2025. Disponível em: <https://aws.amazon.com/pt/compare/the-difference-between-machine-learning-supervised-and-unsupervised/>. Citado na página 16.

_____. **O que é Regressão Logística?** 2025. Disponível em: <https://aws.amazon.com/what-is/logistic-regression/>. Citado na página 16.

ANUNCIAÇÃO, L. Qui quadrado. In: _____. **Conceitos e análises estatísticas com R e JASP**. São Paulo, SP: Nila Press, 2021. cap. 9. Citado na página 29.

BATISTA, G. E. de A. P. A. **Pré-processamento de dados em aprendizado de máquina supervisionado**. Tese (Doutorado) — Universidade de São Paulo, São Carlos, SP, 2003. Disponível em: <https://doi.org/10.11606/T.55.2003.tde-06102003-160219>. Citado na página 26.

BRITISH ACADEMY OF FILM AND TELEVISION ARTS. **About the BAFTA Awards**. 2025. Disponível em: <https://www.bafta.org/about>. Citado na página 22.

CORRêA, I. T. Trabalho de Conclusão de Curso, **Análise dos sentimentos expressos na rede social Twitter em relação aos filmes indicados ao Oscar 2017**. 2017. Disponível em: <https://repositorio.ufu.br/handle/123456789/20133>. Citado na página 23.

DIRECTORS GUILD OF AMERICA. **History of the DGA Awards**. 2025. Disponível em: <https://www.dga.org/Awards/History>. Citado na página 22.

EINSTEIN, A. **Como Vejo o Mundo**. Rio de Janeiro: Nova Fronteira, 2024. Frase adaptada da obra. Citado na página 6.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37, 1996. Disponível em: <https://doi.org/10.1609/aimag.v17i3.1230>. Citado na página 14.

FILHO, M. **O que é acurácia em machine learning?** 2020. Disponível em: <https://mariofilho.com/o-que-e-acuracia-em-machine-learning/>. Citado na página 20.

_____. **Precisão, recall e F1-score em machine learning**. 2020. Disponível em: <https://mariofilho.com/precisao-recall-e-f1-score-em-machine-learning/>. Citado na página 20.

FRANCK, C. T.; WILSON, C. E. Predicting competitions by combining conditional logistic regression and subjective bayes: An Academy Awards case study. **The Annals of Applied Statistics**, v. 15, n. 4, p. 2083–2100, 2021. Disponível em: <https://doi.org/10.1214/21-AOAS1464>. Citado na página 23.

G1. **Fernanda Torres ganha prêmio Satellite Awards**. 2025. Disponível em: <https://g1.globo.com/pop-arte/cinema/noticia/2025/01/26/fernanda-torres-ganha-premio-satellite-awards.ghtml>. Citado na página 22.

_____. **O que é o Globo de Ouro: entenda a importância do prêmio**. 2025. Disponível em: <https://g1.globo.com/pop-arte/cinema/noticia/2025/01/06/o-que-e-o-globo-de-ouro-entenda-a-importancia-do-premio-inedito-de-fernanda-torres.ghtml>. Citado na página 22.

GRACIANO, H. L. dos S.; RAMALHO, R. A. S. ScraperCI: um web scraper para coleta de dados científicos. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 28, p. 1–18, 2023. Citado na página 26.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York, NY: Springer, 2009. Citado na página 14.

IBM. **Análise de valor discrepante**. 2025. Disponível em: <https://www.ibm.com/docs/pt-br/planning-analytics/2.0.0?topic=ai-outlier-analysis>. Acesso em: 14 mai. 2025. Citado na página 28.

_____. **Co-occurrence rules: IBM SPSS Modeler**. 2025. Disponível em: <https://www.ibm.com/docs/pt-br/spss-modeler/18.5.0?topic=techniques-co-occurrence-rules>. Citado na página 30.

_____. **Floresta Aleatória: o que é, como funciona e por que usar**. 2025. Disponível em: <https://www.ibm.com/br-pt/think/topics/random-forest>. Acesso em: 25 abr 2025. Citado 2 vezes nas páginas 19 e 20.

_____. **Gini impurity measure**. 2025. Disponível em: <https://www.ibm.com/docs/pt-br/cognos-analytics/12.1.0?topic=terms-gini-impurity-measure>. Acesso em: 25 abr 2025. Citado na página 33.

_____. **Modelos de Classificação: o que são e como funcionam**. 2025. Disponível em: <https://www.ibm.com/br-pt/think/topics/classification-models>. Acesso em: 25 abr 2025. Citado na página 17.

_____. **O que é regressão logística?** 2025. Disponível em: <https://www.ibm.com/br-pt/think/topics/logistic-regression>. Acesso em: 25 abr 2025. Citado na página 17.

_____. **Random Forest: explicação e aplicação**. 2025. Disponível em: <https://www.ibm.com/br-pt/think/topics/random-forest>. Citado na página 16.

_____. **Árvores de decisão: como funcionam e por que são úteis**. 2025. Disponível em: <https://www.ibm.com/br-pt/think/topics/decision-trees>. Acesso em: 25 abr 2025. Citado 2 vezes nas páginas 18 e 19.

JOBLIB CONTRIBUTORS. **Joblib: Lightweight pipelining with Python**. [S.l.], 2025. Disponível em: <https://joblib.readthedocs.io/>. Citado na página 25.

LEE, K.; PARK, J.; KIM, I.; CHOO, J.; KIM, J. Predicting movie success with machine learning techniques: ways to improve accuracy. **Information Systems Frontiers**, v. 20, p. 577–588, 2018. Disponível em: <https://doi.org/10.1007/s10796-016-9689-z>. Citado na página 23.

LEGRAMANDI, S. O que é o SAG Awards e como ele pode atrapalhar (ou não) Fernanda Torres no Oscar. **Estadão**, 2025. Disponível em: <https://www.terra.com.br/diversao/entre-telas/filmes/o-que-e-o-sag-awards-e-como-ele-pode-atrapalhar-ou-nao-fernanda-torres-no-oscar,0de0e2ed0b85f5394eafdc589997e793ml5isvfq.html>. Citado na página 22.

MASIH, S.; IHSAN, I. Using Academy Awards to predict success of Bollywood movies using machine learning algorithms. **International Journal of Advanced Computer Science and Applications (IJACSA)**, v. 10, n. 2, 2019. Disponível em: <http://dx.doi.org/10.14569/IJACSA.2019.0100257>. Citado na página 24.

MATPLOTLIB COMMUNITY. **Matplotlib: Visualization with Python**. [S.l.], 2025. Disponível em: <https://matplotlib.org/>. Citado na página 25.

MIOT, H. A. Análise de correlação em estudos clínicos e experimentais. **Jornal Vascular Brasileiro**, v. 17, n. 4, p. 275–279, 2018. Disponível em: <https://doi.org/10.1590/1677-5449.174118>. Citado na página 28.

NATIONAL BOARD OF REVIEW. **About the NBR**. 2025. Disponível em: <https://nationalboardofreview.org/about/>. Citado na página 22.

NUMPY DEVELOPERS. **NumPy: The fundamental package for scientific computing with Python**. [S.l.], 2025. Disponível em: <https://numpy.org/>. Citado na página 25.

OLIVEIRA, K. L. S. **Análise da influência de avaliações de críticos e usuários nas premiações do Oscar e do Globo de Ouro em 2023 usando aprendizado de máquina**. Monografia (Trabalho de Conclusão de Curso) — Universidade Federal de Uberlândia, 2023. Disponível em: <https://repositorio.ufu.br/handle/123456789/38672>. Citado na página 23.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, É. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Disponível em: <https://doi.org/10.48550/arXiv.1201.0490>. Citado na página 16.

PRODUCERS GUILD OF AMERICA. **About the PGA Awards**. 2025. Disponível em: <https://www.producersguildawards.com/en/home/about>. Citado na página 22.

REITZ, K. **Requests: HTTP for Humans**. [S.l.], 2025. Disponível em: <https://requests.readthedocs.io/>. Citado na página 25.

RICHARDSON, L. **Beautiful Soup Documentation**. [S.l.], 2025. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Citado na página 25.

SCIKIT-LEARN DEVELOPERS. **Scikit-learn: Machine Learning in Python**. [S.l.], 2025. Disponível em: <https://scikit-learn.org/stable/>. Citado na página 25.

THE PANDAS DEVELOPMENT TEAM. **Pandas: Python Data Analysis Library**. [S.l.], 2025. Disponível em: <https://pandas.pydata.org/>. Citado na página 25.

VILLAÇA, P. **Quanto custa um Oscar e por que o Globo de Ouro é uma piada**. 2017. Disponível em: <https://cinemaemcena.com.br/coluna/ler/2247/79-quanto-custa-um-oscar-e-por-que-o-globo-de-ouro-e-uma-piada>. Citado 2 vezes nas páginas 14 e 22.

WASKOM, M. **Seaborn: Statistical Data Visualization**. [S.l.], 2025. Disponível em: <https://seaborn.pydata.org/>. Citado na página 25.