Universidade Federal do ABC
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Trabalho de Graduação em Engenharia da Informação

# Previsão de vencedores do Óscar: Uma abordagem de aprendizado de máquina

**Vitor Henrique Alves de Oliveira**

**RA: 11099415**

**Santo André - SP, 6 de dezembro de 2023**

Vitor Henrique Alves de Oliveira

# Previsão de vencedores do Óscar: Uma abordagem de aprendizado de máquina

**Trabalho de Graduação** apresentado na disciplina Trabalho de Graduação III, no curso de Engenharia de Informação, como parte dos requisitos necessários para a conclusão do curso.

Universidade Federal do ABC – UFABC

Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas

Graduação em Engenharia de Informação

Orientador: Kenji Nose Filho

Santo André - SP

6 de dezembro de 2023

# Resumo

Este trabalho explora a aplicação de técnicas de aprendizado de máquina na simulação do processo de votação da Academia de Artes e Ciências Cinematográficas no Prêmio Óscar. Utilizando um algoritmo de raspagem para coletar dados da Wikipédia, analisamos informações de premiações notáveis, incluindo os *Directors Guild Awards*, BAFTA (*British Academy of Film and Television Arts*), *Producers Guild Awards*, *Screen Actors Guild Awards*, Globo de Ouro e o Festival de Cannes. A simulação, baseada em um modelo de Floresta Aleatória, demonstrou sucesso na previsão dos vencedores da categoria de Melhor Filme em edições passadas do Óscar. Esse êxito ressalta a convergência entre a análise de dados e a indústria cinematográfica, estabelecendo uma base sólida para futuras pesquisas e investigações nesse campo. O trabalho apresenta uma abordagem abrangente e promissora para a previsão de vencedores do Óscar com base em informações históricas e técnicas de aprendizado de máquina, explorando o potencial da análise de dados na indústria do entretenimento.

**Palavras-chaves**: Aprendizado de máquina, Prêmio Óscar, simulação, algoritmo de raspagem, análise de dados, Floresta Aleatória, indústria cinematográfica, previsão de vencedores.

# Abstract

This study explores the application of Machine Learning techniques in simulating the voting process of the Academy of Motion Picture Arts and Sciences at the Óscars. Using a web scraping algorithm to collect data from Wikipedia, we analyzed information from notable award ceremonies, including the Directors Guild Awards, BAFTA, Producers Guild Awards, Screen Actors Guild Awards, the Golden Globe Awards, and the Cannes Film Festival. The simulation, based on a Random Forest model, demonstrated success in predicting the winners of the Best Picture category in past editions of the Óscars. This success highlights the convergence between data analysis and the film industry, establishing a solid foundation for future research and investigations in this field. The study presents a comprehensive and promising approach to predicting Óscar winners based on historical information and machine learning techniques, exploring the potential of data analysis in the entertainment industry.

**Keywords**: Machine Learning, Óscar Awards, preferential voting, scraping algorithms, data analysis, Random Forest, film industry, prediction of winners.

# Lista de ilustrações

# Lista de tabelas

# Lista de abreviaturas e siglas

| | |
|---|---|
| DGA | *Directors Guild of America Awards* |
| BAFTA | *British Academy Film Awards* |
| PGA | *Producers Guild of America* |
| SAG | *Screen Actors Guild* |
| AUC | *Area Under The Curve* |
| ROC | *Receiver Operating Characteristics* |

# Sumário

# 1 Introdução

Na indústria cinematográfica, o Óscar, oficialmente chamado de Prêmio da Academia (em inglês: *The Academy Awards*), é uma das cerimônias de premiação mais cobiçadas, celebrando anualmente a excelência dos filmes. Entre todas as categorias, o prêmio de Melhor Filme é o mais prestigioso, representando a culminação das preferências artísticas, comerciais e pessoais. A seleção do vencedor, no entanto, é um processo complexo. A Academia utiliza um sistema de votação preferencial, no qual os membros classificam os filmes com base em suas preferências. Este projeto de graduação mergulha nessa complexidade, explorando como a análise de dados e técnicas de aprendizado de máquina podem ser utilizados para prever os vencedores do Óscar na categoria de Melhor Filme.

Inspirados pela natureza multifacetada do processo de seleção, o objetivo é construir um modelo que simule o processo de votação com algoritmo de Floresta Aleatória para capturar as múltiplas preferências dos membros da Academia e prever o resultado do sistema de votação. Coletamos e estruturamos dados históricos de diversos prêmios cinematográficos e aplicamos uma simulação baseada em algoritmos.

Este estudo visa explorar os detalhes do sistema de votação, elaborar uma estrutura de simulação e analisar os resultados para contribuir para a compreensão das tendências da Academia e percepções orientadas por dados. À medida que desvendamos padrões ocultos nas escolhas dos vencedores do Óscar, enriquecemos as discussões sobre as complexidades dos prêmios e as influências que moldam esse prestigioso processo de seleção.

Ao longo deste esforço, esclarecemos como algoritmos e análises podem revelar as preferências da Academia, proporcionando uma visão aprofundada das tendências e padrões subjacentes que influenciam os resultados deste prêmio icônico.

## 1.1 Motivação

A indústria cinematográfica desempenha um papel fundamental na cultura e na sociedade, moldando opiniões, disseminando narrativas e fornecendo entretenimento global. Em seu centro está o Óscar, um pináculo de excelência cinematográfica que cativa cineastas, críticos e público em todo o mundo. Entre suas categorias de maior prestígio, a seleção do prêmio de Melhor Filme reflete as preferências e critérios multifacetados que moldam as escolhas da Academia de Artes e Ciências Cinematográficas.

A busca para prever os vencedores do Óscar tem interessado pesquisadores, analistas e entusiastas do cinema há décadas. No entanto, o processo de seleção é influenciado por fatores multidimensionais que vão além da mera avaliação artística.

Esta pesquisa é motivada pela crescente convergência da análise de dados e da indústria cinematográfica. A era digital permitiu a coleta, organização e análise de grandes quantidades de informações, abrindo oportunidades sem precedentes para desvendar as tendências subjacentes nas seleções dos vencedores do Óscar. A aplicação de técnicas de aprendizado de máquina, como o modelo Floresta Aleatória, nos capacita a modelar o intrincado processo de votação e prever possíveis resultados com base em dados históricos.

Ao obter entendimento sobre os padrões ocultos por trás das escolhas da Academia, este trabalho tem como objetivo lançar luz sobre os intrincados critérios que moldam os prêmios de Melhor Filme. A validação e aplicação de modelos preditivos podem enriquecer a análise cinematográfica, fornecendo informações valiosas tanto para a indústria quanto para a academia. Além disso, a abordagem inovadora deste estudo abre novos caminhos para a análise de dados em contextos artísticos, demonstrando como a tecnologia pode auxiliar na compreensão das preferências humanas.

Ao considerar a intrincada interação entre dados, preferências e escolhas artísticas, este trabalho visa revelar as complexidades do processo de votação do Óscar e contribuir para o discurso contínuo sobre a arte e a ciência por trás da seleção dos melhores filmes do ano.

## 1.2 Objetivos

O objetivo central deste projeto consiste em criar um modelo preditivo, fazendo uso da análise de dados e técnicas de Aprendizado de Máquina, com o propósito de prever os filmes vencedores da categoria de Melhor Filme no Óscar.

# 2 Revisão Bibliográfica

A busca por prever os vencedores do Óscar, especialmente na categoria de Melhor Filme, tem sido objeto de interesse de pesquisadores e analistas na interseção da análise de dados e da indústria cinematográfica. A combinação de elementos subjetivos e objetivos que influenciam as escolhas dos membros da Academia criou um ambiente desafiador para previsões precisas. Nos últimos anos, a aplicação de técnicas de Aprendizado de Máquina e análise de dados tem demonstrado potencial para iluminar as nuances desse processo complexo.

Uma pesquisa seminal nessa área foi conduzida por Krauss et al. (2008), que desenvolveram um modelo baseado em análise de sentimentos de críticas de filmes para prever os vencedores do Óscar. Eles destacaram a importância de incorporar elementos subjetivos, como a opinião crítica, além de fatores objetivos, na previsão de vencedores.

Outro estudo relevante é o trabalho de Zhang et al. (2016), que aplicaram algoritmos de Aprendizado de Máquina, incluindo Redes Neurais Artificiais, para prever os vencedores do Óscar em várias categorias. Eles enfatizaram a importância de considerar a diversidade de características dos filmes, como elenco, diretor e gênero, ao desenvolver modelos de previsão.

Em relação ao sistema de votação preferencial, metodologias específicas de modelagem foram exploradas. Mattei, Forshee e Goldsmith (2012) desenvolveram um modelo de votação preferencial que considerava as interações entre os filmes e a complexidade das preferências dos eleitores. Sua abordagem destacou a necessidade de capturar as nuances do processo de seleção.

No contexto específico de algoritmos de aprendizado de máquina, a Floresta Aleatória tem se destacado. Breiman (2001) apresentou essa técnica como uma combinação de múltiplas árvores de decisão para melhorar a precisão e robustez das previsões. A abordagem tem sido aplicada em várias áreas e, mais recentemente, na previsão de vencedores de premiações, como o estudo de Jewalikar (2016) sobre a previsão dos vencedores do Grammy.

A abordagem de simulação adotada neste projeto para reproduzir o processo de votação preferencial assemelha-se à pesquisa de Laver e Sergenti (2011), que utilizaram simulações baseadas em agentes para modelar o processo de votação em competições artísticas. Sua pesquisa enfatizou a capacidade das simulações em refletir a complexidade das escolhas humanas.

# 3 Metodologia

Nesta seção, descrevemos detalhadamente o processo de coleta de dados, destacando a utilização da Wikipédia como fonte primária e o uso de algoritmos de raspagem de dados para a coleta das informações. Também apresentamos todas as premiações que foram incluídas na avaliação, bem como as diferentes coletas de dados realizadas.

## 3.1 Coleta de dados

A coleta de dados foi realizada principalmente a partir da Wikipédia, e tem como base o trabalho desenvolvido por (TLALKA, 2019), utilizando técnicas de raspagem para extrair informações relevantes dos artigos dedicados a cada premiação. As premiações avaliadas incluem o *Directors Guild Awards* (DGA), BAFTA (*British Academy of Film and Television Arts*), *Producers Guild Awards* (PGA), *Screen Actors Guild Awards* (SAG), Globo de Ouro, Festival de Cannes e, é claro, o próprio Óscar.

Durante o processo de coleta, foram realizadas diferentes abordagens para obter informações variadas. Uma coleta de dados foi focada nos indicados e vencedores do Óscar, proporcionando um panorama dos filmes reconhecidos pela Academia.

A Tabela 1 exibe uma breve amostra dos resultados obtidos por meio da raspagem de dados para filmes que foram indicados ao prêmio de Melhor Filme no Óscar. A coluna *Year* apresenta os anos de lançamento de cada filme, enquanto a coluna *Film* exibe os títulos correspondentes. A coluna *Wiki* contém as referências da Wikipédia associadas a cada filme. Por fim, a coluna *Winner* indica, por meio dos valores verdadeiros (*True*) ou falsos (*False*), se um determinado filme foi vencedor do prêmio de Melhor Filme.

Tabela 1 – Dados coletados via raspagem de dados, com os indicados e vencedores do prêmio de melhor filme do Óscar.

| *Year* | *Film* | Wiki | *Winner* |
|---|---|---|---|
| 1927 | Wings (1927 film) | </wiki/Wings_(1927_film)> | True |
| 1927 | 7th Heaven (1927 film) | </wiki/7th_Heaven_(1927_film)> | False |
| 1927 | The Racket (1928 film) | </wiki/The_Racket_(1928_film)> | False |
| 1928 | The Broadway Melody | </wiki/The_Broadway_Melody> | True |
| 1928 | Alibi (1929 film) | </wiki/Alibi_(1929_film)> | False |
| 1928 | Hollywood Revue | </wiki/Hollywood_Revue> | False |
| 1928 | In Old Arizona | </wiki/In_Old_Arizona> | False |
| 1929 | Disraeli (1929 film) | </wiki/Disraeli> | False |
| 1931 | Grand Hotel (1932 film) | </wiki/Grand_Hotel_(1932_film)> | True |
| 1934 | Viva Villa! | </wiki/Viva_Villa!> | False |
| 1928 | The Patriot (1928 film) | </wiki/The_Patriot_(1928_film)> | False |

Realizamos uma segunda coleta de dados, a Tabela 2 que inclui as mesmas informações da tabela anterior, com a adição do campo *Nominations* e o título abreviado (*Film Text*). Esse valor numérico representa o total de indicações que um determinado filme recebeu em diversas categorias do Óscar.

Tabela 2 – Dados coletados via raspagem de dados, com o número de indicações de cada filme no Óscar.

| Year | Film | Wiki | Nominations | Film Text |
|---|---|---|---|---|
| 2006 | Babel (film) | </wiki/Babel_(film)> | 7 | Babel |
| 2018 | Bao (film) | </wiki/Bao_(film)> | 1 | Bao |
| 1951 | Benjy (film) | </wiki/Benjy_(film)> | 1 | Benjy |
| 1957 | Les Girls | </wiki/Les_Girls> | 3 | Les Girls |
| 2013 | Mr Hublot | </wiki/Mr_Hublot> | 1 | Mr Hublot |
| 1958 | Mon Oncle | </wiki/Mon_Oncle> | 1 | My Uncle |

Adicionalmente, conduzimos coletas de dados específicas para cada um dos prestigiados prêmios paralelos, tais como o DGA, BAFTA, PGA, SGA, Globo de Ouro e o Festival de Cannes. Essas coletas proporcionaram informações detalhadas sobre os indicados e vencedores em cada uma dessas premiações, compartilhando semelhanças notáveis com a primeira raspagem realizada. Essa abordagem enriqueceu significativamente nossa base de dados, acrescentando uma camada diversificada de informações. Um exemplo ilustrativo, contendo uma pequena amostragem do DGA, é apresentado na Tabela 3.

Tabela 3 – Dados coletados via raspagem de dados, com os indicados e vencedores do DGA.

| Year | Film | Wiki | Winner |
|---|---|---|---|
| 1948 | A Letter to Three Wives | </wiki/A_Letter_to_Three_Wives> | True |
| 1948 | Red River (1948 film) | </wiki/Red_River_(1948_film)> | False |
| 1948 | The Snake Pit | </wiki/The_Snake_Pit> | False |
| 1948 | The Search | </wiki/The_Search> | False |
| 1949 | The Third Man | </wiki/The_Third_Man> | False |
| 1949 | Champion (1949 film) | </wiki/Champion_(1949_film)> | False |

Os dados coletados contêm informações a partir da década de 20 e vai até o ano de 2022 e as informações obtidas foram salvos em formato CSV (*Comma-separated values*). Nos Anexos A e B apresentamos os códigos utilizados para realizar a coleta dos dados do Óscar e DGA, respectivamente.

## 3.2 Pré-processamento e organização dos dados

Após a coleta de dados, o próximo passo envolveu o pré-processamento e a organização das informações. Os dados coletados foram estruturados em uma base de dados que continha detalhes como o ano de lançamento do filme, seu nome, indicados e vencedores

do Óscar, número de nomeações e resultados em outras premiações como o DGA, BAFTA, PGA, SGA, Globo de Ouro e o Festival de Cannes.

Essa estruturação, que serve como entrada para o modelo, se utiliza dos dados coletados, para organizá-los em uma grande base de dados, que contem os anos, nomes, número de nomeações para o Óscar e outras colunas com valores binários para as nomeações e prêmios que o filme possui (1 para vencedor ou nomeado, 0 para não-vencedor ou não-nomeado). Dentre as premiações considerados podemos citar: nomeação de melhor filme na categoria drama do Globo de Ouro, vencedor de melhor filme na categoria drama do Globo de Ouro, nomeação de melhor filme na categoria comédia do Globo de Ouro, vencedor de melhor filme na categoria comédia do Globo de Ouro, nomeação de melhor filme do PGA, vencedor de melhor filme do PGA, nomeação de melhor filme do BAFTA, vencedor de melhor filme do BAFTA, nomeação de melhor filme do DGA, vencedor de melhor filme do DGA, nomeação de melhor elenco de filme do SAG, vencedor de melhor elenco de filme do SAG, nomeação de melhor filme do CANNES, e vencedor de melhor filme do CANNES. Estas *features*, juntamente com o número de nomeações para o Óscar, serviram como entrada para o nosso classificador, como é possível observar no código apresentado no Anexo C.

A Tabela 4 apresenta uma breve visualização da base de dados final utilizada para o treinamento do modelo, ilustrando apenas algumas das *features* utilizadas como entrada.

Tabela 4 – Exemplo de alguns dos dados extraídos e utilizados para o treinamento do classificador.

| *Year* | *Film* | *Nominations* | *Nom GG Drama* |
|---|---|---|---|
| 1927 | Wings (1927 film) | 2 | 0 |
| 1927 | 7th Heaven (1927 film) | 5 | 0 |
| 1927 | The Racket (1928 film) | 7 | 0 |
| 1928 | The Broadway Melody | 3 | 0 |
| 1928 | Alibi (1929 film) | 7 | 0 |
| 1928 | Hollywood Revue | 7 | 0 |
| 1928 | In Old Arizona | 5 | 0 |

## 3.3 Aplicação do algoritmo de Floresta Aleatória na simulação

A escolha do algoritmo de Floresta Aleatória baseou-se na sua eficácia comprovada em lidar com conjuntos de dados complexos e na capacidade de modelar relações não-lineares. A Floresta Aleatória é uma técnica que combina múltiplas árvores de decisão para alcançar maior robustez e precisão, como feito por (PARKER, 2020), que, juntamente com AutoML, previu o vencedor da categoria de Melhor Filme do Óscar de 2020.

Ao contrário de abordagens anteriores, que serviram de referência para este trabalho, buscamos desenvolver um algoritmo capaz de modelar o problema e avaliar seu desempenho

com base em métricas de precisão. Esse enfoque demandou uma atenção especial na filtragem dos dados de entrada do modelo, assim como na criação de condicionais exclusivas para os anos que estavam sendo previstos. Por exemplo, um filme do ano em análise poderia ser nomeado ao Óscar, mas não poderia constar como vencedor, a fim de evitar a contaminação do modelo com os resultados que ele procurava antecipar. Além disso, optamos por prever um intervalo de tempo, em vez de um único ano, como realizado em outros estudos.

A técnica de Floresta Aleatória é uma poderosa ferramenta no campo de aprendizado de máquina, especialmente eficaz em problemas de classificação e regressão. Baseada em um conjunto de árvores de decisão, a Floresta Aleatória se destaca pela sua capacidade de capturar padrões complexos em conjuntos de dados heterogêneos, como é o caso das preferências cinematográficas dos membros da Academia de Artes e Ciências Cinematográficas.

O funcionamento da Floresta Aleatória é razoavelmente simples, mas extremamente eficaz. Em vez de depender de uma única árvore de decisão, a Floresta Aleatória cria múltiplas árvores independentes durante o treinamento. Cada uma dessas árvores é treinada com uma amostra aleatória dos dados disponíveis, e em cada nó de divisão, um subconjunto aleatório de características (*features*) é considerado para fazer a melhor divisão possível (BREIMAN, 2001). Na Figura 1 é apresentada um diagrama ilustrativo simplificado do funcionamento de uma Floresta Aleatória.

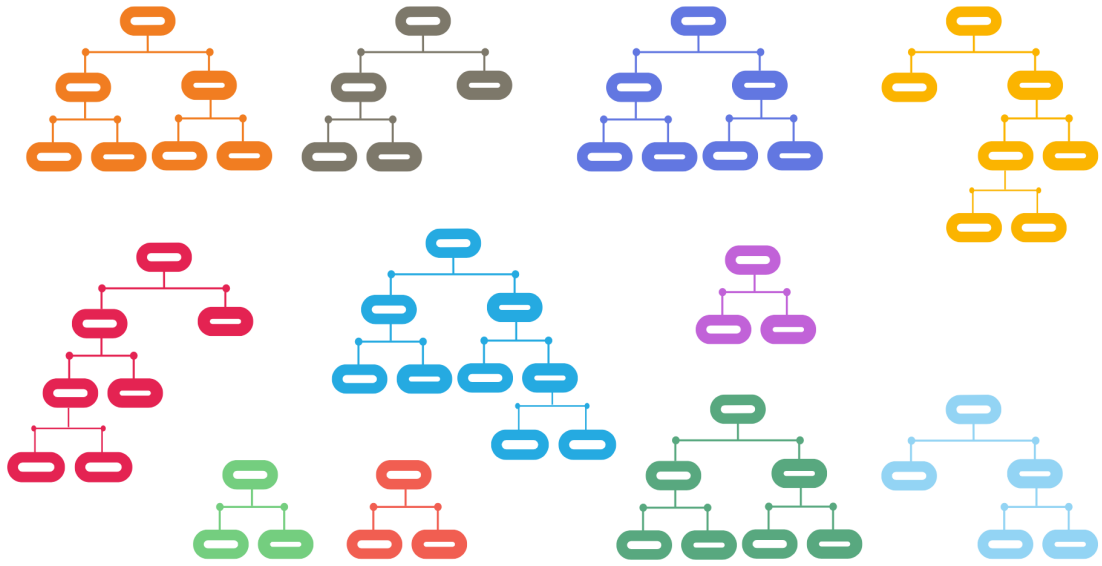

Figura 1 – Diagrama ilustrativo simplificado do funcionamento de uma Floresta Aleatória. Figura extraída de: (BREIMAN, 2001)

A diversidade entre as árvores individuais é fundamental para o sucesso da Flo-

resta Aleatória. Ao introduzir aleatoriedade na seleção de características e nos dados de treinamento para cada árvore, o modelo é capaz de capturar uma variedade mais ampla de padrões e relações nos dados. Isso reduz a tendência de *overfitting*, onde o modelo se adapta excessivamente aos dados de treinamento específicos e falha em generalizar para novos dados. Isso acaba tornando-a especialmente útil para lidar com a complexidade das preferências cinematográficas.

Durante a fase de previsão, cada árvore na Floresta Aleatória emite uma previsão individual. Para problemas de classificação, a previsão final é determinada por votação majoritária entre todas as árvores; para problemas de regressão, a previsão final é a média das previsões de todas as árvores. Esse processo de agregação de previsões reduz o impacto de eventuais erros individuais de árvores específicas, resultando em previsões mais estáveis e confiáveis.

Além disso, a Floresta Aleatória oferece uma série de parâmetros ajustáveis que permitem otimizar o desempenho do modelo para o problema em questão. Por exemplo, é possível controlar o número de árvores na floresta, a profundidade máxima das árvores individuais e a quantidade de características consideradas em cada divisão. Esses ajustes finos permitem adaptar a Floresta Aleatória para diferentes conjuntos de dados e requisitos de desempenho. O número de árvores de decisão (*n-estimators*) e a profundidade máxima de cada árvore (*max-depth*) foram ajustados para otimizar o desempenho do modelo. Essa busca permitiu encontrar configurações que proporcionaram o melhor desempenho para o problema em questão.

É importante observar que a análise de dados e a simulação não substituem as nuances e complexidades humanas envolvidas no processo de votação. Além disso, esta pesquisa se baseia em dados históricos e suposições de padrões de votação que podem variar ao longo do tempo.

# 4 Resultados e Discussão

Neste Capítulo, apresentamos os resultados obtidos com o algoritmo de Floresta Aleatória, bem como uma discussão desses resultados.

## 4.1 Resultados da simulação

Podemos dizer que a simulação foi conduzida com êxito, prevendo com uma certa acurácia os vencedores esperados na categoria de Melhor Filme do Óscar para cada ano, abrangendo filmes lançados de 1996 a 2022.

Para prever o ano de 1996, foram utilizados no treinamento os dados históricos desde a década de 20 até 1995. Para prever o ano de 1997, o ano de 1996 foi incluído na base de dados de treinamento e assim por diante, até prevermos o ganhador do prêmio de melhor filme de 2022.

Na Tabela 5, apresentamos os filmes vencedores pelo nosso modelo comparados aos vencedores reais ao longo desses anos. Para facilitar a visualização dos resultados destacamos, em negrito, as previsões que foram realizadas corretamente.

Foram analisadas um total de 27 edições do Óscar, desde o ano de 1996 até 2022. O modelo de Floresta Aleatória conseguiu prever corretamente os vencedores do Óscar em 16 das 27 edições analisadas. Isso significa que em mais da metade das edições, o modelo acertou o filme que seria premiado como Melhor Filme. Essas vitórias previstas incluíram filmes notáveis, como *Titanic*, *The Lord of the Rings: The Return of the King*, *The Departed*, *No Country for Old Men*, e *The King's Speech*.

Nas outras 11 edições, o modelo não conseguiu prever o vencedor corretamente. Nestes casos, o filme que o modelo previu como vencedor não correspondeu ao filme que a Academia de Artes e Ciências Cinematográficas premiou. Alguns exemplos notáveis de filmes que não foram previstos corretamente incluem *American Beauty*, *Gladiator*, *A Beautiful Mind*, *The Shape of Water*, e *Nomadland*.

Ao comparar os resultados da simulação com a realidade, surgem questões intrigantes sobre as variáveis que podem determinar a seleção do Melhor Filme. Fatores como popularidade, direção, atuação e narrativa podem desempenhar papéis cruciais nas preferências dos membros da Academia, embora não sejam abordados nesse trabalho. A identificação de discrepâncias entre previsões e resultados reais também pode sugerir mudanças nas preferências ao longo do tempo, refletindo tendências da indústria cinematográfica e da sociedade.

Tabela 5 – Vencedores Previstos e Reais do Óscar de Melhor Filme, de 1996 à 2022.

| Estreia | Vencedores Previstos | Vencedores Reais |
|---|---|---|
| **1996** | **The English Patient** | **The English Patient** |
| **1997** | **Titanic** | **Titanic** |
| 1998 | Saving Private Ryan | Shakespeare in Love |
| **1999** | **American Beauty** | **American Beauty** |
| 2000 | Crouching Tiger, Hidden Dragon | Gladiator |
| **2001** | **A Beautiful Mind** | **A Beautiful Mind** |
| **2002** | **Chicago** | **Chicago** |
| **2003** | **TLoTR: The Return of the King** | **TLoTR: The Return of the King** |
| 2004 | The Aviator | Million Dollar Baby |
| 2005 | Brokeback Mountain | Crash |
| **2006** | **The Departed** | **The Departed** |
| **2007** | **No Country for Old Men** | **No Country for Old Men** |
| **2008** | **Slumdog Millionaire** | **Slumdog Millionaire** |
| **2009** | **The Hurt Locker** | **The Hurt Locker** |
| **2010** | **The King's Speech** | **The King's Speech** |
| **2011** | **The Artist** | **The Artist** |
| **2012** | **Argo** | **Argo** |
| 2013 | Gravity | 12 Years a Slave |
| **2014** | **Birdman** | **Birdman** |
| 2015 | The Revenant | Spotlight |
| 2016 | La La Land | Moonlight |
| **2017** | **The Shape of Water** | **The Shape of Water** |
| 2018 | Roma | Green Book |
| 2019 | 1917 | Parasite |
| 2020 | The Trial of the Chicago 7 | Nomadland |
| 2021 | The Power of the Dog | CODA |
| **2022** | **Everything Everywhere** | **Everything Everywhere** |

É importante reconhecer que a simulação, embora baseada em dados históricos e algoritmos avançados, não substitui a riqueza das escolhas humanas e a subjetividade inerente à seleção de vencedores do Óscar. A aplicação da simulação a edições passadas não garante previsões precisas para edições futuras, considerando a evolução das preferências e contextos. No entanto, a simulação proporciona uma plataforma inovadora para explorar tendências e dinâmicas subjacentes.

É possível realizar análises mais detalhadas e notar um comportamento interessante emergindo. Observa-se que até o ano de 2012, o modelo acertou 13 dos 17 vencedores, apresentando uma taxa de acerto superior a 76%. No entanto, considerando as 10 edições subsequentes, a acurácia do modelo cai para cerca de 33% (3 de 10), evidenciando uma queda acentuada. É notável que, nesse segundo período, as preferências da academia sofreram alterações significativas.

Desde então, a Academia de Artes e Ciências Cinematográficas tem enfrentado

diversas críticas e polêmicas, especialmente relacionadas à sua composição majoritariamente formada por homens brancos e conservadores. Essas críticas levaram a academia a buscar mudanças significativas, como mencionado por Horn (2014). Um exemplo dessas mudanças é a decisão de premiar um filme sobre a escravidão, *12 Years a Slave*, visando se desvincular da imagem preconcebida que foi construída ao longo dos anos.

## 4.2 Avaliação do desempenho do modelo

A avaliação do desempenho do modelo desempenha um papel crucial em qualquer projeto de aprendizado de máquina. No contexto deste estudo, é fundamental determinar o quão bem o modelo de Floresta Aleatória pode prever os vencedores do Óscar na categoria de Melhor Filme. Para medir esse desempenho, empregamos várias métricas de avaliação que nos fornecem informações valiosas sobre a qualidade das previsões e a eficácia do modelo, como a Acurácia, a Área Sob a Curva da Característica de Operação do Receptor (AUC-ROC) e a Matriz de Confusão Geral

- Acurácia: A acurácia é uma métrica geral que mede a proporção de todas as previsões corretas feitas pelo modelo. É uma medida geral de quão bem o modelo está se saindo.
- AUC-ROC (Área Sob a Curva da Característica de Operação do Receptor): A AUC-ROC é uma métrica que avalia a capacidade do modelo de distinguir entre as classes (vencedor ou não vencedor do Óscar). Quanto mais próxima de 1 for a pontuação AUC-ROC, melhor o modelo está em fazer previsões precisas.
- Matriz de Confusão Geral: A matriz de confusão fornece uma visão detalhada do desempenho do modelo, incluindo os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Isso ajuda a entender onde o modelo está acertando e onde está cometendo erros.

Cada uma dessas métricas desempenha um papel único na avaliação do modelo. A AUC-ROC nos ajuda a entender quão bem o modelo é capaz de classificar os vencedores do Óscar. A acurácia fornece uma medida geral do desempenho, enquanto a matriz de confusão nos ajuda a entender o comportamento do modelo em detalhes.

Utilizando essas métricas em conjunto, podemos avaliar de forma objetiva o desempenho do modelo de Floresta Aleatória na previsão dos vencedores do Óscar.

### 4.2.1 Acurácia

A acurácia é uma métrica fundamental que nos permite avaliar o quão bem o modelo se saiu em prever os vencedores reais do Óscar ao longo dos anos. Essa métrica

revela a proporção de previsões corretas em relação ao total de previsões feitas pelo modelo.

A análise da acurácia ano a ano nos fornece informações valiosas sobre o desempenho variável do modelo. Alguns anos alcançam uma acurácia perfeita, atingindo 1,00, o que significa que o modelo fez previsões precisas nesses casos. No entanto, em outros anos, a acurácia é mais baixa, atingindo valores como 0,60 ou 0,78, o que indica que o modelo não teve o mesmo sucesso nas previsões nesses períodos, como pode ser visto pelo gráfico da Figura 2.

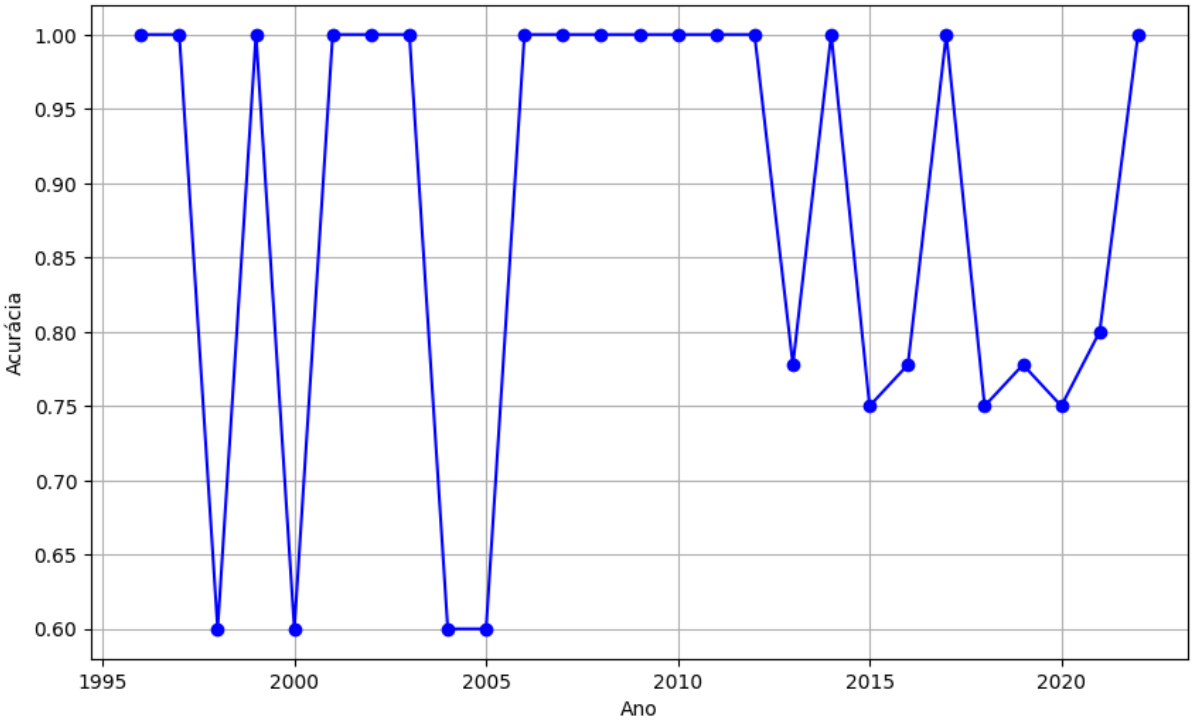


Figura 2 – Acurácia ao longo dos anos

A acurácia geral do modelo, que leva em consideração o desempenho ao longo de todos os anos analisados, é uma métrica que nos fornece uma visão global do quão bem o modelo se sai. A acurácia geral ficou em torno de 0,87, sugerindo que o modelo possui um desempenho muito bom na previsão dos vencedores do Óscar.

### 4.2.2 AUC-ROC (*Receiver Operating Characteristics-Area Under The Curve*)

Obtivemos um desempenho notável com um AUC-ROC de 0,90, indicando que nosso modelo de simulação é altamente eficaz em distinguir os vencedores reais do Óscar dos outros filmes. Quanto mais próximo de 1, melhor o desempenho.

A curva ROC mostra como nosso modelo varia seu desempenho com diferentes limites de classificação. Quanto mais a curva se afasta da linha de referência, melhor o modelo. Nossa curva está bem distante, indicando uma forte capacidade de separar as classes, como pode ser visto pela curva da Figura 3.

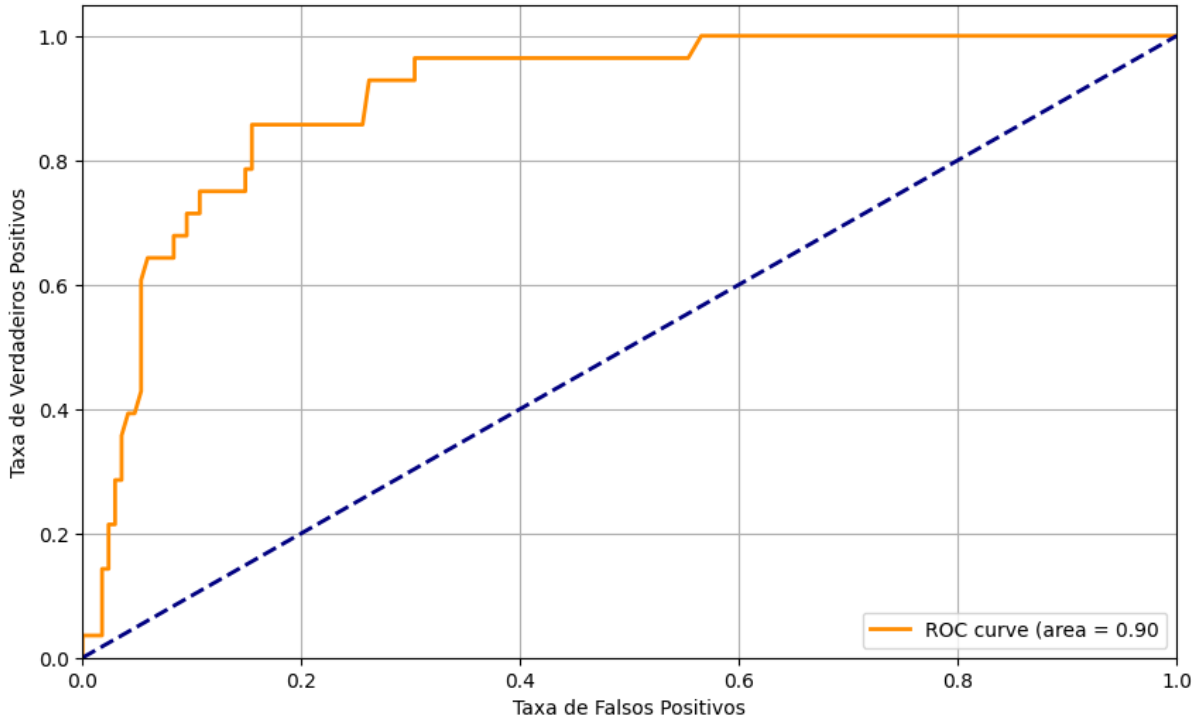


Figura 3 – Curva ROC Geral

### 4.2.3 Matriz de Confusão

Ao avaliar a performance de um modelo de aprendizado de máquina, é crucial analisar a matriz de confusão. Esta matriz apresenta informações sobre como o modelo se comporta em termos de classificações corretas e incorretas. No nosso projeto, a matriz de confusão geral, que combina os resultados de todos os anos, é apresentada pela Figura 4.

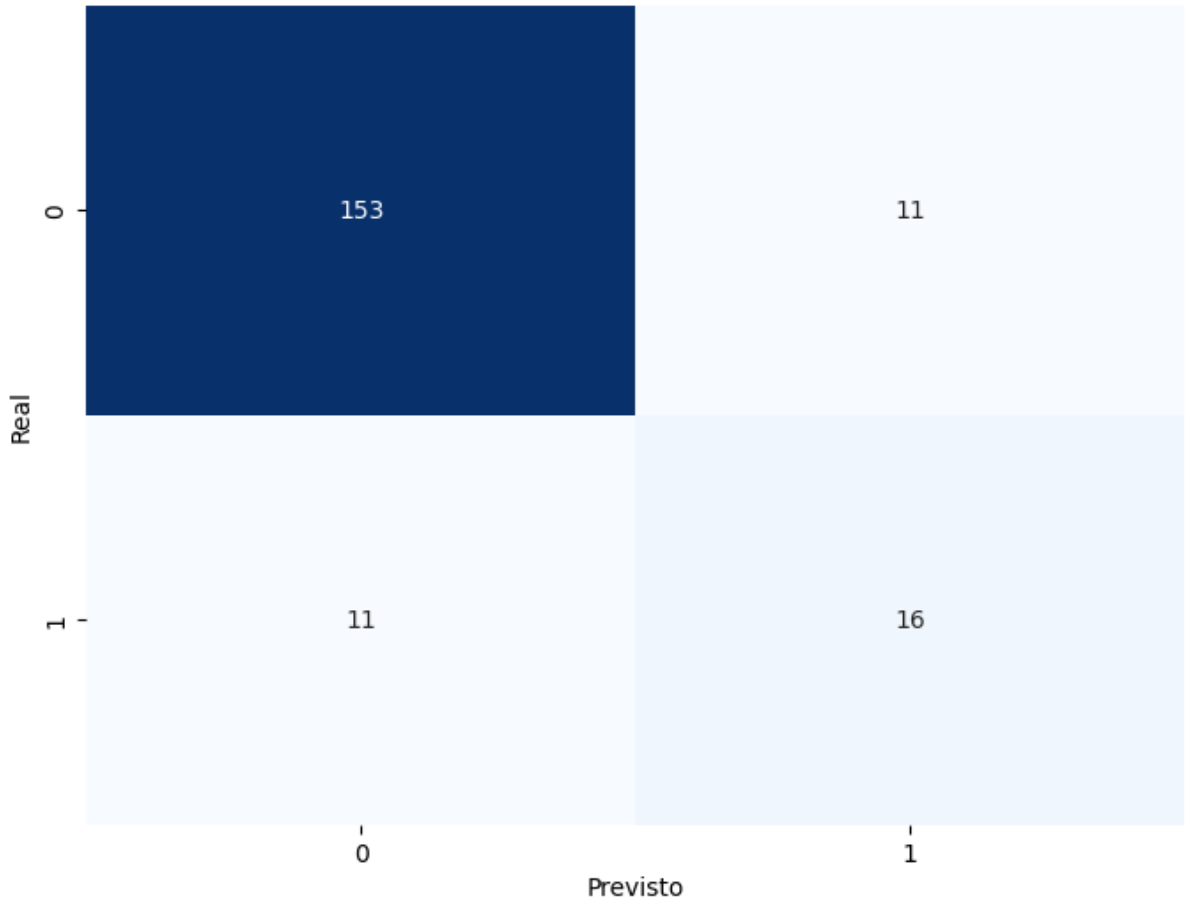


Figura 4 – Matriz de Confusão Geral

Os valores apresentados nesta matriz representam:

153 (Verdadeiros Negativos - VN): Representa quantas vezes o modelo previu corretamente que um filme não ganharia o Óscar e o filme realmente não ganhou.

11 (Falsos Positivos - FP): Refere-se ao número de vezes em que o modelo previu incorretamente que um filme ganharia o Óscar, mas o filme não ganhou.

11 (Falsos Negativos - FN): Indica quantas vezes o modelo previu incorretamente que um filme não ganharia o Óscar, mas o filme acabou ganhando.

16 (Verdadeiros Positivos - VP): Representa o número de vezes em que o modelo previu corretamente que um filme ganharia o Óscar e o filme realmente ganhou.

Em uma análise da Matriz de Confusão, é crucial estabelecer claramente o contexto. Em todos os anos de previsão, é esperado um número significativo de Verdadeiros Negativos. Podemos ilustrar isso considerando um cenário com 10 filmes indicados, onde, no mínimo, 8 desses casos seriam classificados como Verdadeiros Negativos. Isso se deve ao fato de que, mesmo entre os 2 filmes restantes, se o modelo errar o vencedor e atribuir a vitória a outro filme que não seja o vencedor real, ele terá previsto corretamente que outros 8 filmes não seriam os vencedores, os quais de fato não foram. Portanto, é fundamental

compreender o comportamento e o contexto nos quais o modelo opera, a fim de evitar decisões equivocadas e interpretações falsas.

É importante destacar que o processo de seleção do Melhor Filme do Óscar é complexo e influenciado por uma série de fatores subjetivos. O modelo de Floresta Aleatória foi projetado para capturar essas complexidades, mas ainda assim, há limitações inerentes à previsão de prêmios em um campo tão diverso e dinâmico.

Os achados deste estudo têm relevância tanto para a área de análise de dados quanto para a indústria cinematográfica. A aplicação de técnicas de Aprendizado de Máquina na previsão de vencedores do Óscar destaca a inovação na interseção da arte e da tecnologia. Além disso, a análise das preferências da Academia oferece informações sobre as tendências da indústria, influenciando estratégias de produção, marketing e distribuição de filmes.

A modelagem da simulação e a análise dos resultados contribuem para uma melhor compreensão das complexidades que permeiam o processo de votação no Prêmio Óscar. Embora não seja uma ferramenta de previsão definitiva, a abordagem enriquece a discussão sobre como as escolhas dos membros da Academia são moldadas por fatores multifacetados.

# 5 Conclusões e Trabalhos Futuros

Neste capítulo, resumimos os principais pontos abordados ao longo deste trabalho, apresentamos um resumo dos resultados obtidos por meio da simulação do processo de votação preferencial e discutimos a relação desses resultados com os objetivos propostos inicialmente.

No contexto do nosso projeto, exploramos a aplicação de técnicas de aprendizado de máquina para prever os vencedores do Óscar na categoria de “Melhor Filme”. Foram analisados dados históricos de 1996 a 2022, considerando diversas características relacionadas a nomeações e prêmios de diferentes associações e academias cinematográficas. A avaliação do modelo foi baseada em métricas como AUC-ROC, Acurácia e Matriz de Confusão.

Os resultados obtidos são promissores, mostrando que é possível construir um modelo capaz de fazer previsões relevantes sobre os vencedores do Óscar. A métrica AUC-ROC geral, que indica a capacidade de distinguir entre vencedores e não vencedores, atingiu um valor de 0,90, o que é considerado um desempenho sólido.

Embora tenhamos obtido resultados encorajadores, existem diversas oportunidades para aprimorar nosso modelo de previsão. Alguns trabalhos futuros incluem:

- Engenharia de Atributos: Explorar e adicionar mais características relevantes para a previsão dos vencedores do Óscar, como dados de crítica, bilheteria e outros prêmios.
- Modelos Avançados: Testar diferentes algoritmos de aprendizado de máquina e técnicas avançadas de modelagem para melhorar o desempenho do modelo.
- Aprendizado Profundo: Investigar a aplicação de redes neurais profundas para capturar relações complexas entre as características e as categorias de vencedores.
- Validação Temporal: Implementar uma validação temporal estrita para simular o cenário de previsão real, onde os dados futuros não estão disponíveis.
- Amostragem Estratificada: Considerar estratégias de amostragem estratificada para lidar com desequilíbrios nos dados, uma vez que o número de vencedores é limitado a cada ano.
- Aprimoramento das Métricas: Trabalhar na melhoria das métricas de precisão, revocação e F1-Score, visando aumentar a assertividade das previsões.

Em resumo, nosso projeto forneceu uma base sólida para a previsão dos vencedores do Óscar na categoria de "Melhor Filme". Contudo, o desenvolvimento e aprimoramento

contínuos do modelo são essenciais para aumentar sua eficácia e capacidade de adaptação às mudanças ao longo do tempo. Este trabalho representa um ponto de partida promissor para futuras pesquisas e desenvolvimentos na área de previsão de vencedores em competições de cinema.

# Referências

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, p. 5–32, 2001. Citado 2 vezes nas páginas 3 e 8.

HORN, J. Oscars: '12 years a slave' puts spotlight on hollywood's approach to race. *Los Angeles Times*, 2014. Citado na página 13.

JEWALIKAR, V. Predicting the grammys with data. *LinkedIn*, 2016. Citado na página 3.

KRAUSS, J. et al. Predicting movie success and academy awards through sentiment and social network analysis. 2008. Citado na página 3.

LAVER, M.; SERGENTI, E. *Party competition: An agent-based model*. [S.l.]: Princeton University Press, 2011. v. 18. Citado na página 3.

MATTEI, N.; FORSHEE, J.; GOLDSMITH, J. An empirical study of voting rules and manipulation with large datasets. *Proceedings of COMSOC*, Citeseer, v. 59, 2012. Citado na página 3.

PARKER, N. *Predicting the Oscars with Machine Learning*. 2020. Disponível em: <https://github.com/njparker1993/oscars_predictions>. Citado na página 7.

TLALKA, J. *Predict Oscars 2019 with Data Science*. 2019. Disponível em: <https://github.com/Buzdygan/movie_analysis>. Citado na página 5.

ZHANG, Y. et al. Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine. *Simulation*, SAGE Publications Sage UK: London, England, v. 92, n. 9, p. 861–871, 2016. Citado na página 3.

# Anexos

# ANEXO A – Algoritmo de raspagem do Melhor Filme do Óscar

```
import requests as rq
import re
import datetime
import traceback
from collections import Counter
from bs4 import BeautifulSoup
import pandas as pd
import requests
import os
import codecs
import lxml
from google.colab import drive
drive.mount('/content/drive')

# URL da página da Wikipedia
url = 'https://en.wikipedia.org/wiki/Academy_Award_for_Best_Picture'

# Função para extrair as informações de cada filme
def extract_film_info(row, current_year):
    columns = row.find_all('td')

    if len(columns) == 2:
        film_col = columns[0]
        if row.get('style') == 'background:#FAEB86':
            winner = True
        else:
            winner = False
        try:
            a = film_col.find('i').find('a')
            return (current_year, a.get('title'), a.get('href'), winner)
        except:
            traceback.print_exc()

    return None
```

```
35
36 # Realize a solicitação HTTP e crie um objeto BeautifulSoup
37 response = requests.get(url)
38 oscar_soup = BeautifulSoup(response.text, 'lxml')
39
40 oscar_results = []
41 current_year = 0
42
43 # Encontre todas as tabelas com a classe 'wikitable'
44 for table in oscar_soup.find_all('table', {'class': 'wikitable'})
   :
45     for row in table.find_all('tr')[1:]:
46         columns = row.find_all('td')
47
48         if len(columns) == 1:
49             current_year = int(re.search(r'[\d]{4}', columns[0].
                  text).group(0))
50         elif len(columns) == 2:
51             film_info = extract_film_info(row, current_year)
52             if film_info:
53                 oscar_results.append(film_info)
54
55 # Crie um DataFrame a partir dos resultados e salve em um arquivo
    CSV
56 df = pd.DataFrame(oscar_results, columns=['year', 'film', 'wiki',
    'winner'])
57 df.to_csv('/content/drive/MyDrive/TG/melhorfilme/Dados/osc_bp.csv
   ', index=False)
```

Listing A.1 – Parte do algorítmo utilizado para Scrapping, contendo os indicados e vencedores do Óscar de Melhor Filme

# ANEXO B – Algoritmo de raspagem do melhor filme do DGA

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import traceback
import re

# URL da página da Wikipedia
url = 'https://en.wikipedia.org/wiki/Directors_Guild_of_America_Award_for_Outstanding_Directing_%E2%80%93_Feature_Film'

# Faz a solicitação e analisar a página
response = requests.get(url)
soup = BeautifulSoup(response.text, 'lxml')

# Lista para armazenar todos os resultados
all_dga_results = []

# Encontra todas as tabelas com a classe 'wikitable'
tables = soup.find_all('table', {'class': 'wikitable'})

# Itera por todas as tabelas
for table in tables:
    # Lista para armazenar os resultados da tabela atual
    dga_results = []

    # Variável para rastrear o ano da cerimônia do Directors Guild of America Award
    current_year = None
    first_winner_found = False  # Variável para controlar o primeiro vencedor do ano

    # Itera pelas linhas da tabela
    for row in table.find_all('tr'):
        columns = row.find_all('td')

```

```
33          if columns:
34              # Certifique-se de que há pelo menos uma coluna
35              if len(columns) >= 2:
36                  # Verificar o número de colunas para determinar a
                        posição da coluna do filme
37                  if len(columns) == 4:
38                      current_year = int(re.search('[\d]{4}',
                            columns[0].text).group(0))
39                      film_col = columns[2]
40                  else:
41                      film_col = columns[1]
42
43                  # Verifica se o conteúdo da coluna de vencedor
                        indica que o filme é um vencedor
44                  winner = any(keyword in row.text.lower() for
                        keyword in ['won', 'winner', 'victory'])
45
46                  if not first_winner_found and winner:
47                      first_winner_found = True
48                  else:
49                      winner = False
50
51                  try:
52                      # Encontrar o link do filme e adicionar os
                            resultados
53                      a = film_col.find('i').find('a')
54                      dga_results.append((current_year, a['title'],
                            a['href'], winner))
55                  except Exception as e:
56                      print(f"Problema com {row}")
57                      traceback.print_exc()
58
59      # Adiciona os resultados da tabela atual à lista de todos os
            resultados
60      all_dga_results.extend(dga_results)
61
62  # Cria um DataFrame do Pandas com todos os resultados
63  df = pd.DataFrame(all_dga_results, columns=['year', 'film', 'wiki
        ', 'winner'])
64
65  # Especifica o caminho para salvar o arquivo CSV no Google Drive
66  google_drive_path = '/content/drive/MyDrive/TG/melhorfilme/Dados/
```

```
    dgas.csv'
67
68 print(f"Arquivo CSV do Directors Guild of America Award salvo no
    Google Drive em: {google_drive_path}")
```

Listing B.1 – Parte do algorítmo utilizado para Scrapping, contendo os indicados e vencedores do DGA

# ANEXO C – Algoritmo de Floresta Aleatória

```
# Dados verdadeiros e probabilidades previstas para a curva ROC geral
all_y_true = []
all_y_score = []

# Loop de simulação de 1995 a 2022
for i in range(len(anos_treino)):
    ano_treino = anos_treino[i]
    ano_previsao = anos_previsao[i]

    # Filtra os dados de treino até o ano em questão
    data_treino = osc_df[osc_df['year'] <= ano_treino]

    # Remove valores entre parênteses nos títulos dos filmes
    data_treino = data_treino.copy()  # Crie uma cópia do DataFrame para evitar o SettingWithCopyWarning
    data_treino['film'] = data_treino['film'].apply(lambda x: re.sub(r'\([^)]*\)', '', x))

    # Escolhe as características (X) e o rótulo (y) para os dados de treino
    features = ['Nominations', 'nom_gg_drama', 'winner_gg_drama', 'nom_gg_comedy', 'winner_gg_comedy', 'nom_pga', 'winner_pga', 'nom_bafta', 'winner_bafta', 'nom_dga', 'winner_dga', 'nom_sag', 'winner_sag', 'nom_cannes', 'winner_cannes']
    X_treino = data_treino[features]
    y_treino = data_treino['Oscar_win']  # Rótulo indicando se o filme ganhou o Oscar

    # Cria e treina um modelo de Random Forest com dados de treino
    model = RandomForestClassifier(n_estimators=100, random_state=42)
    model.fit(X_treino, y_treino)

    # Filtra os dados do ano de previsão
    data_previsao = osc_df[osc_df['year'] == ano_previsao]
```

```
28
29      # Escolhe as características (X) para os dados de previsão
30      X_previsao = data_previsao[features]
31
32      # Obtem as probabilidades previstas em vez das previsões biná
           rias
33      probabilities_previsao = model.predict_proba(X_previsao)
34
35      # Calcula a AUC-ROC para o ano de previsão
36      y_true = [1 if f == data_previsao['film'].values[0] else 0
           for f in data_previsao['film']]
37      auc_roc = roc_auc_score(y_true, probabilities_previsao[:, 1])
38      auc_rocs.append(auc_roc)
39
40      # Curva ROC geral
41      all_y_true.extend(y_true)
42      all_y_score.extend(probabilities_previsao[:, 1])
43
44      # Encontra o índice do filme com a maior probabilidade
45      indice_filme_vencedor_simulado = probabilities_previsao[:,
           1].argmax()
46      filme_vencedor_simulado = data_previsao.iloc[
           indice_filme_vencedor_simulado]['film']
47      filme_vencedor_simulado = re.sub(r'\([^)]*\)', '',
           filme_vencedor_simulado)  # Remove valores entre parênteses
            nos vencedores simulados
48      vencedores_simulados.append(filme_vencedor_simulado)
49
50      # Calcula o F1-Score, precisão e revocação para o ano de
           previsão
51      y_pred = [1 if i == indice_filme_vencedor_simulado else 0 for
            i in range(len(data_previsao))]
52      f1 = f1_score(y_true, y_pred)
53      precision = precision_score(y_true, y_pred)
54      recall = recall_score(y_true, y_pred)
55
56      f1_scores.append(f1)
57      precisions.append(precision)
58      recalls.append(recall)
59
60      # Calcula a acurácia para o ano de previsão
61      accuracy = accuracy_score(y_true, y_pred)
```

```
62      accuracies.append(accuracy)
63
64      #print(ano_treino)
65      #print(ano_previsao)
66      #print(y_true)
67      #print(y_pred)
68
69      # Calcula a matriz de confusão para o ano de previsão
70      conf_matrix = confusion_matrix(y_true, y_pred)
71      confusion_matrices.append(conf_matrix)
72
73  # Calcula as métricas gerais
74  auc_roc_geral = sum(auc_rocs) / len(auc_rocs)
75  f1_geral = sum(f1_scores) / len(f1_scores)
76  precision_geral = sum(precisions) / len(precisions)
77  recall_geral = sum(recalls) / len(recalls)
78  accuracy_geral = sum(accuracies) / len(accuracies)
79
80  # Calcula a matriz de confusão geral
81  confusion_matrix_geral = sum(confusion_matrices)
```

Listing C.1 – Parte do algorítmo utilizado para treinar o modelo e aplicar a Floresta Aleatória