UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Kauê Lucas Silvério Oliveira

Análise da influência de avaliações de críticos e usuários nas premiações do Oscar e do Globo de Ouro em 2023 usando aprendizado de máquina

Uberlândia, Brasil

2023

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Kauê Lucas Silvério Oliveira

Análise da influência de avaliações de críticos e usuários nas premiações do Oscar e do Globo de Ouro em 2023 usando aprendizado de máquina

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 Ciência da Computação.

Orientador: Prof. Dr. Paulo Henrique Ribeiro Gabriel

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Ciência da Computação

Uberlândia, Brasil

2023

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:	Bacharelado em Ciência da Computação				
Defesa de:	GBC082 - Projeto de Graduação 2				
Data:	22/06/2023	Hora de início:	18:00	Hora de encerramento:	19:15
Matrícula do Discente:	11821BCC007				
Nome do Discente:	Kauê Lucas Silvério Oliveira				
Título do Trabalho:	Análise da influência de avaliações de críticos e usuários nas premiações do Oscar e do Globo de Ouro em 2023 usando aprendizado de máquina				
A carga horária curricular foi cumprida integralmente?	(X) Sim ( ) Não				

Reuniu-se na plataforma Microsoft Teams, da Universidade Federal de Uberlândia, a Banca Examinadora, designada pelo Colegiado do Curso de Graduação em Ciência da Computação, assim composta: Professores: Dr. Marcelo Zanchetta do Nascimento - FACOM/UFU; Dr. Rodrigo Sanches Miani - FACOM/UFU; e Dr. Paulo Henrique Ribeiro Gabriel - FACOM/UFU orientador(a) do(a) candidato(a).

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(a) Nota 100

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 **Paulo Henrique Ribeiro Gabriel**, **Professor(a) do Magistério Superior**, em 22/06/2023, às 19:16, 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 **Marcelo Zanchetta do Nascimento**, **Professor(a) do Magistério Superior**, em 22/06/2023, às 19:16, 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 22/06/2023, às 19:16, 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 **4580159** e o código CRC **B6674B5F**.

Resumo

A indústria do cinema é extremamente relevante e vem recebendo cada vez mais atenção com a chegada dos serviços de vídeo sobre demanda na última década. Parte da repercussão e visibilidade das obras lançadas anualmente é impulsionado por diversas premiações e cerimônias que ocorrem durante o ano. O objetivo desse trabalho é analisar a possível relação entre a aceitação do público, se baseando nas notas providas por críticos e usuários comuns dentro da plataforma Metacritic, e os ganhadores das premiações Oscar de melhor filme, Globo de Ouro de melhor drama e Globo de Ouro de melhor comédia no ano de 2023. Foi coletado todo o conjunto de dados dos filmes indicados às três cerimônias no período de 2007 a 2023, de onde foi extraído as características necessárias para realizar uma análise preliminar e treinar modelos de aprendizado de máquina com os algoritmos *naive bayes*, floresta aleatória e KNN com o intuito de prever os vencedores do último ano. Inicialmente foi separado uma base de treino e de teste,a partir dos dados anteriores a 2023, onde as métricas de acurácia, precisão, revocação e pontuação F1 foram medidas. Por fim os modelos foram treinados com os dados de 2007 a 2022 e o resultado da classificação dos indicados de 2023 foi comparando com o resultado real das premiações. Ao concluir o trabalho não foi possível perceber uma ligação clara entre as notas de cada obra e seu resultado nas cerimônias. O desempenho dos algoritmos, apenas considerando a base de teste, também não foi satisfatório apresentando métricas abaixo do esperado. A análise final não demonstrou nenhuma correlação clara entre os resultados das premiações do Oscar de melhor filme e Globo de Ouro de melhor drama com as avaliações apresentadas no Metacritic, apesar de apontar possíveis candidatos mostrando que os vencedores foram sim bem avaliados. O Globo de Ouro de melhor comédia apresentou um caso a parte onde de fato o ganhador era o indicado com maior média das notas de críticos porém com uma posição mais baixa em relação ao público geral, o que possivelmente fez com que não fosse escolhidos pelos modelos de aprendizagem como favorito.

Palavras-chave: Oscar, Globo de Ouro, Aprendizado de máquina, predição, Metacritic.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O crescimento da indústria do cinema se confunde com parte da história do entretenimento nas últimas décadas. Movimentando bilhões de dólares todos os anos, filmes e séries fazem parte do cotidiano de boa parte das pessoas do globo diariamente (AFONSO, 2021). Ao mesmo tempo, o acesso a esse tipo de conteúdo tornou-se cada vez mais simples. Plataformas de vídeo sobre demanda disponibilizam catálogos de lançamentos para consumo a qualquer momento, mantendo sempre o espectador entretido com novas produções (DUMELA, 2020). É dentro desse cenário tão competitivo, onde centenas de obras são lançados por ano (FOLLOWS, 2017), que vemos a existência de eventos e premiações como o Oscar e Globo de Ouro chamando atenção do público e da mídia.

Criado em 1929 e mantido até hoje pela Academia de Artes e Ciências Cinematográficas, o Oscar é uma tradicional premiação de cinema. Seu principal objetivo é premiar com uma estatueta de ouro, seguindo os critérios da academia, os melhores filmes do ano de acordo com cada categoria, tendo o melhor filme do ano como seu maior prêmio da noite (OSCARS.ORG, 2022). Ultimamente com o advento das redes sociais e uma maior interação entre o público geral, essa premiação ganhou ainda mais peso no meio midiático. Com várias pessoas conectadas comentando e acompanhando a noite de premiações, o Oscar é sempre um assunto muito discutido pelo amantes de cinema.

Seguindo um caminho similar tem-se o Globo de Ouro. Criado pela Associação de Imprensa Estrangeira de Hollywood na década de 1940, essa premiação tem como objetivo premiar os profissionais da indústria de cinema de Hollywood. Com um leque de categorias os prêmios do Globo de Ouro são decididos por 93 membros de uma Associação de Correspondentes Estrangeiros de Hollywood associados à mídia (GOLDENGLOBES.COM, 2022).

Apesar dessas categorias apenas acontecerem no início de cada ano, a indústria do cinema gera engajamento do público a todo momento. A cada nova obra lançada, críticos de cinema prontamente disponibilizam em seus portais notas e críticas para que qualquer um possa saber com antecedência o que esperar de cada filme. Nesse contexto então surgem diversas ferramentas especializadas em agregar notas e análises, facilitando a vida dos usuários ao buscar informações e fomentando discussões sobre essas obras. Através desses *sites* geralmente é possível não só consumir informações de especialistas sobre qualquer filme que o usuário queira como também externar suas opiniões e avaliações sobre qualquer conteúdo cadastrado, para que outras pessoas tenham acesso.

Com um grande volume de dados de avaliações existentes nesses agregadores é possível então realizar análises afim de relacionar o sucesso das obras vencedoras com sua

recepção pelo público e pela crítica, aplicando ferramentas de aprendizado de máquina com o objetivo de rotular novos indicados das cerimônias, verificando a existência de possíveis padrões com a tentativa de prever os próximos ganhadores dos prêmios Oscar de melhor filme, Globo de Ouro de melhor drama e Globo de Ouro de melhor comédia no ano de 2023.

1.1 Justificativa

No mundo digital e conectado que atual, boa parte da população influencia e é influenciado a todo momento por qualquer informação presente na Internet que seja interessante para os mesmos. As pessoas frequentemente tomam decisões de qual conteúdo consumir através de avaliações online e premiações, incluindo os próprios jurados presentes nestes eventos. Vemos que muita gente busca assistir todos os premiados assim que são anunciados para ter suas próprias opiniões e compartilhar seus sentimentos nas redes sociais em cada etapa da premiação.

Através desse trabalho, alcançando o objetivo de relacionar o resultado dos premiados através de notas disponibilizadas abertamente pela Internet, será possível estipular qual é o impacto das notas na opinião dos jurados da academia e da associação de imprensa, tentando entender de que forma ambas são influenciadas por informações externas e qual a diferença entre elas nesse quesito. Também podemos descobrir se há alguma relação prática entre o que essas associações identificam como valioso nas obras premiadas e o que de fato o público geral gosta de consumir e avaliar de forma positiva.

Ao estimar o ganhador com antecedência também seria possível prever possíveis próximos assuntos em redes sociais e permitir que marcas preparem planos de marketing e engajamento na data desses grandes eventos. Serviços de *streamming* podem disponibilizar filmes com alta probabilidade de vitória em seus catálogos previamente com garantia de sucesso imediato pós premiação.

Por fim será possível comparar este trabalho com outros que possuem objetivos semelhantes mas utilizam outras métricas,verificando se apenas utilizar notas para analisar o desempenho de um filme já seria suficiente para tomar conclusões sobre seu sucesso nesse tipo de evento ou se dados mais complexos como análises escritas e comentários são mais efetivos no caso dessa análise.

1.2 Objetivos

O objetivo geral desse trabalho visa a utilização de métodos de aprendizado de máquina para verificar se há alguma relação entre as notas de críticos e usuários comuns

dos filmes indicados ao Oscar e Globo de Ouro e o resultado final das premiações. Para isso foram seguidos os seguintes objetivos específicos:

- Definir a origem dos dados, analisando os pontos positivos e negativos de cada agregador de notas, que será usado para a realização do trabalho.
- Desenvolver uma ferramenta para extração de análises.
- Criar uma base com todos os dados necessários para realizar o treinamento dos algoritmos além de definir quais características serão extraídas.
- Treinar os algoritmos escolhidos a partir dos atributos gerados para apresentar a qualidade de cada modelo no problema proposto além de expor os indicados de 2023 aos modelos treinados e verificar a qualidade da classificação.
- Analisar o resultado e expor as conclusões encontradas ao fim do trabalho.

2 Referencial Teórico

Neste capítulo será realizada uma revisão pelo referencial teórico necessário para o desenvolvimento e o entendimento deste trabalho. Inicialmente, na seção 2.1 sera apresentado com mais detalhes o que são os sites agregadores de críticas. A seção 2.2 irá descrever como funciona a extração dos dados necessários para esse tipo de trabalho, definindo quais dificuldades foram encontradas, quais técnicas foram abordadas e como a base de dados foi construída. A seção 2.3 desenvolve os conceitos relacionados a aprendizado de máquina, dando um panorama geral em relação ao conceito em si e se aprofundando nos algoritmos utilizados para desenvolvimento deste estudo. Por fim a seção 2.4 irá trazer os trabalhos relacionados a este que buscam objetivos parecidos em relação a predição e classificação utilizando aprendizado de máquina.

2.1 Agregadores de criticas

Os agregadores de críticas, ou agregadores de análises, são sítios na Internet que agrupam análises e notas sobre filmes, livros, jogos, séries e outros conteúdos relacionados a entretenimento. Alguns agregadores e seu funcionamento serão exemplificados.

O IMDB (IMDB, 2022b) é uma grande referência entre esses agregadores. Neste *site* é possível ver detalhes sobre filmes, séries e programas de TV, descobrir curiosidades, acompanhar rankings e ver notícias sobre celebridades do cinema internacional. Os usuários da plataforma conseguem criar análises sobre os conteúdos que assistem com comentários e uma nota final medido em estrelas, indo de 0 a 10,0, onde a média ponderada das avaliações de todos os usuários se torna a nota final da obra na plataforma.

O Rotten Tomatoes (ROTTENTOMATOES, 2022b) possui uma abordagem parecida com o IMDB. No geral, seu objetivo é oferecer um apanhado de notas e críticas de veículos e críticos especializados sobre filmes e séries, porém trazendo algumas peculiaridades. Primeiramente, as notas do Rotten são indicadas por porcentagem, indo de 0 a 100. Há também um indicador visual, onde um tomate indica a qualidade do filme, sendo o tomate maduro um bom filme e um tomate podre um filme ruim. Além disso, tem-se a adição também de uma métrica separada da crítica de usuários comuns, onde há uma nota que agrega as opiniões da audiência de modo geral.

O Metacritic (METACRITIC, 2022b) possui características semelhantes aos outros dois exemplos, porém é mais abrangente. Com um catálogo não só de séries, filmes e televisão, também há um grande foco na área de jogos eletrônicos e de música, além de uma área focada em notícias. Este site por sua vez utiliza uma combinação de cores e

números de 0 a 100 para indicar as pontuações das obras. Notas com a cor verde são

Figura 1 – Tabela da relação entre cores e notas.

Significado Geral da Pontuação	Filmes, TV e Música	Jogos
Elogios Universais	81 - 100	90 - 100
Avaliações Geralmente Favoráveis	61 - 80	75 - 89
Avaliações Mistas ou Médias	40 - 60	50 - 74
Avaliações Geralmente Desfavoráveis	20 - 39	20 - 49
Desagrado Universal	0 - 19	0 - 19

Fonte: Traduzido de Metacritic (2022b).

Figura 2 – Tabela de conversão entre notas de 0 a 4 e notas de A a F para a escala do metascore.

Escala de 4 Estrelas	
Sua Nota	Converte para
4	100
3.5	88
3	75
2.5	63
2	50
1.5	38
1	25
0.5	12
0	0

Notas em Letras	
Sua Nota	Converte para
A ou A+	100
A-	91
B+	83
B	75
B-	67
C+	58
C	50
C-	42
D+	33
D	25
D-	16
F+	8
F ou F-	0

Fonte: Traduzido de Metacritic (2022b).

consideradas boas, amarela são as intermediárias e vermelho indica baixa qualidade. Outra diferença é que, apesar das notas dos veículos especializados e da nota geral, chamada de *metascore*, seguir a mesma escala do Rotten Tomatoes, as notas dos usuários da plataforma são tratados de 0 a 10,0 e não entram na casa das centenas. Outro ponto a mencionar é que o *metsacore* é decidido através de uma média ponderada entre as notas dos especialistas, sendo o peso da nota definida de acordo com a credibilidade do avaliador. É importante

notar que nem todos os portais de cinema utilizam as mesmas métricas para avaliação, para isso é utilizada uma tabela de conversão, como demonstrado na figura 2.

2.2 Extração de dados

A extração de dados de uma página web depende essencialmente de um desses dois fatores: A existência de uma interface externa que permite acessar os dados de forma oficial através de qualquer aplicação ou através do código HTML da página, de onde também é possível extrair as informações, técnica conhecida como *web scraping*.

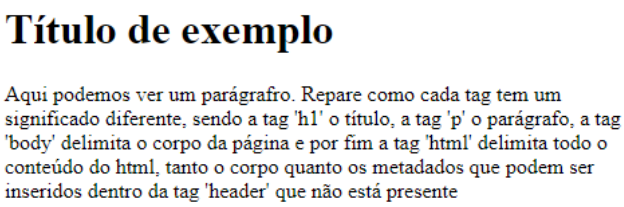
2.2.1 HTML e DOM

O HTML, do inglês *HyperText Markup Language*, ou do português Linguagem de Marcação de HiperTexto, se posiciona como uma linguagem utilizada para estruturar páginas na Internet. Através de blocos demarcados por anotações, ou *tags*, é possível montar todo o esqueleto do site, de forma semântica ou não, demarcando quais áreas de conteúdo ele terá e o que terá dentro delas, podendo adicionar blocos de texto, botões, tabelas, listas, caixas de entrada e diversas outras estruturas que irão compor a página (HTML, 2022). A figura 3 mostra a página gerada pelo html a seguir:

```
<html>
    <body>
        <h1>Título de exemplo</h1>
        <p>Aqui podemos ver um parágrafro. Repare como cada tag tem um
        significado diferente, sendo a tag 'h1' o título,
        a tag 'p' o parágrafo,a tag 'body' delimita o corpo
        da página e por fim a tag 'html' delimita todo o
        conteúdo do html, tanto o corpo quanto os metadados que podem
        ser inseridos dentro da tag 'header' que não está presente</p>
    </body>
</html>
```

Para manipular o conteúdo da página de forma dinâmica, ou então reconhecermos estruturas e navegarmos pela página de forma programática, precisamos utilizar o DOM, do inglês *Document Object Model*, ou em português Modelo de Objeto de Documento. Através de uma representação estruturada em formato de árvore, como mostrado no exemplo da figura 4, onde as *tags* mais internas são nós filho das *tags* mais externas, é possível facilmente localizar, alterar, excluir ou até adicionar novo elementos na composição da página. Isso será utilizado para, através de uma aplicação em python, encontrar

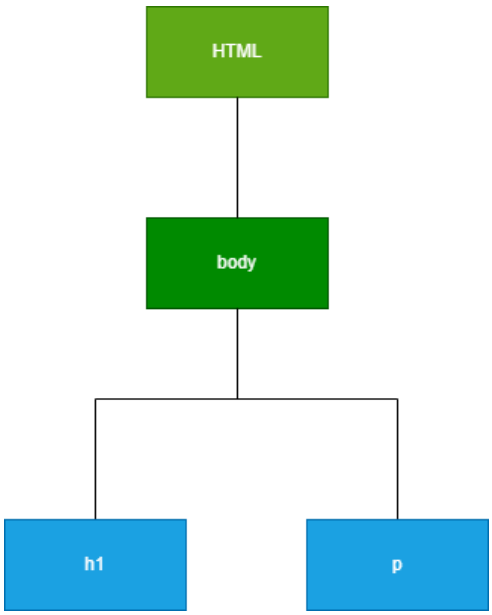
Figura 3 – Resultado gerado pelo html demonstrado.



Fonte: Autor do trabalho.

os blocos HTML que correspondem a análises de usuários da página e extrair os dados que forem úteis (DOM, 2022).

Figura 4 – Exemplo da árvore gerada pelo DOM do HTML citado na figura 3.



Fonte: Autor do trabalho.

2.2.2 Web Scraping

O *Web scraping* consiste em técnicas de extração de dados de páginas web, de forma manual ou automatizada, para que os dados de uma página sejam armazenados em um sistema de arquivos ou em um banco de dados e futuramente analisados (ZHAO, 2017).

Para realizar esse tipo de tarefa com *python* pode-se usar a biblioteca *Beautiful Soup* que permite tirar dados de arquivos HTML ou XML (BEAUTIFULSSOUP, 2022). O código fonte do site será obtido através de uma chamada HTTP do tipo GET diretamente no domínio alvo. Tendo o conteúdo em mãos, a biblioteca irá fazer o *parse* do DOM

da página e gerar uma árvore de *tags* que poderão ser acessadas facilmente de forma programática. A figura 5 demonstra um HTML que será processado pelo *beatiful suop* na figura 6

Figura 5 – Html usado como exemplo para demonstrar o beautiful soap.

```
<html>
    <head>
        <title>The Dormouse's story</title>
    </head>
    <body>
        <p class="title"><b>The Dormouse's story</b></p>

        <p class="story">Once upon a time there were three little sisters; and their names were
        <a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
        <a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
        <a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
        and they lived at the bottom of a well.</p>

        <p class="story">...</p>
    </body>
</html>
```

Fonte: BEAUTIFULSSOUP (2022).

Figura 6 – Beautiful soup sendo aplicado para extrair informações do html da figura 5. É possível ver o resultado de cada atributo acessado nos comentários do código.

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')

soup.title
# <title>The Dormouse's story</title>

soup.title.name
# u'title'

soup.title.string
# u'The Dormouse's story'

soup.title.parent.name
# u'head'

soup.p
# <p class="title"><b>The Dormouse's story</b></p>

soup.p['class']
# u'title'

soup.a
# <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>

soup.find_all('a')
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]

soup.find(id="link3")
# <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>
```

Fonte: BEAUTIFULSSOUP (2022).

2.3 Aprendizado de máquina

O termo aprendizado de máquina, do inglês *machine learning*, se refere a capacidade de alguns algoritmos de aprenderem padrões através de uma base de dados e utilizarem o conhecimento adquirido para executar tarefas simulando um comportamento humano (NAQA; MURPHY, 2015). No caso deste trabalho, serão utilizados técnicas de aprendizagem com um objetivo de classificação, para classificar filmes como possíveis ganhadores do Oscar e Globo de Ouro.

A área de *machine learning* se ramifica em diversos tipos de algoritmos com objetivos diferentes, sendo que a efetividade de cada um dependerá de cada problema (MAHESH, 2020).

2.3.1 Não supervisionado

Algoritmos não supervisionados são algoritmos onde não há uma resposta certa para a classificação ou agrupamento de dados. O programa é alimentado com uma base de informações e de acordo com esses dados tenta entender e criar grupos de objetos com relações entre si (IBM, 2023b).

Esse tipo de aprendizado é utilizado quando não se sabe ao certo quais classes podem ser encontradas, porém é necessário que os dados sejam agrupados para que padrões sejam encontrados ou então apenas deseja-se encontrar relações não conhecidas previamente entre diferentes objetos do banco de dados.

Em suma, esse tipo de aprendizado tem três diferentes casos de uso:

- **Agrupamento**: Técnicas de agrupamento são usadas para segregar dados não classificados em uma quantidade pré-determinada de grupos, baseado em suas similaridades ou diferenças, permitindo que esses agrupamentos sejam posteriormente analisados.
- **Regras de associação**: São algoritmos com foco em detectar padrões entre diferentes variáveis de um grupo de dados, permitindo relacionar objetos através de, por exemplo, comportamentos de usuário e similaridade entre conteúdos consumidos anteriormente.
- **Redução de dimensionalidade**: Tem como objetivo transformar dados em alta-dimensão para dados em menores dimensões, tentando ao máximo manter suas características, permitindo simplificar objetos que possuem grandes quantidades de variáveis que nem sempre são úteis para possíveis análises.

2.3.2 Supervisionado

Aprendizado supervisionado se refere ao tipo de aprendizado onde o algoritmo é treinado para classificar novas entradas de dados baseado em um conhecimento pré-adquirido através de treinamento (IBM, 2023a).

Em processo de treinamento de um algoritmo supervisionado são inseridos diversos objetos já classificados, onde se sabe não só todos os parâmetros de entrada como o resultado desejado para análise. O método escolhido então tem o papel de, através de um modelo matemático, ajustar os pesos de cada atributo, relacionando-os com cada classe, de forma a conseguir prever a probabilidade de uma nova entrada ainda não classificada pertencer a um determinado rótulo.

Existe uma variedade de algoritmos supervisionados disponíveis para diversos casos de uso, onde a efetividade de cada um varia de acordo com o caso de uso. Esse tipo de

aprendizado por ser usado para uma infinidade de objetivos como predição, classificação, reconhecimento de imagem, geração de imagem, *upscalling e downscalling*, detecção de vírus e muitas outras possibilidades.

Alguns exemplos de algoritmos utilizados em aprendizado de máquina supervisionado:

2.3.2.1 Naive Bayes

O algoritmo *naive bayes* é um algoritmo simples, eficiente e que frequentemente produz bons resultados em classificação. Sua base consiste em considerar que os atributos de um determinado dado sendo analisado são condicionalmente independentes, ou seja, o valor de um atributo não interfere no valor de outro atributo, utilizando o teorema de bayes como base (WEBB, 2010).

O teorema de bayes é utilizado para determinar a probabilidade P da existência de um atributo A dado uma classe B:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}. \tag{2.1}$$

Para treinar o algoritmo *naive bayes*, trabalha-se com múltiplos atributos A_i para determinar a probabilidade P de um objeto dado uma classe B, onde n é o número total de atributos:

$$P(A|B) = \prod_{i=1}^{n} P(A_i|B). \tag{2.2}$$

Por ser um método supervisionado de aprendizado, será necessário separar a base de dados entre uma base de treinamento e uma base de teste para utilizar o *naive bayes*, analisando sua eficiência em dados já classificados.

2.3.2.2 KNN (*K Nearest Neighbor*)

KNN, ou K vizinhos mais próximos, é um algoritmo simples de classificação em aprendizado de máquina. Ele consiste em usar medidas de distância para comparar uma nova entrada com entradas já existentes classificadas (PETERSON, 2009). Ao inserir um novo objeto no plano ele é comparado com os K vizinhos mais próximos, sendo K um valor pré definido, e sua classe se torna a classe predominante entre os K vizinhos.

Dado um novo objeto x_1 a ser classificado e um objeto x_2 já existente, onde cada um possui j características (x_{i1},x_{i2},...,x_{ij}) sendo i o índice do objeto, a distância entre dois objetos é calculada através de alguma medida de similaridade ou dissimilaridade, por exemplo:

- **Similaridade por cosseno (similaridade):**

$$sim_{cos}(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{||\overrightarrow{x}||||\overrightarrow{y}||} \tag{2.3}$$

- **Distância euclidiana (dissimilaridade):**

$$d(x_1,x_2) = \sqrt{\sum_{z=1}^{j}(x_{2z} - x_{1z})^2} \tag{2.4}$$

2.3.3 Árvore de decisão

Árvore de decisão, ou *decision tree*, é um tipo de classificador que se utiliza de uma estrutura de árvore para tomar decisões em relação a novas classificações. Ao receber um novo dado, a árvore é percorrida onde cada nó corresponde a um atributo do objeto e direciona, baseado no seu valor, seu fluxo para um próximo nó. Ao chegar em um nó folha a árvore obtêm a classe prevista para aquele objeto (KINGSFORD; SALZBERG, 2008).

Para realizar o treinamento desse modelo é utilizado um sistema de ganho de informação. Como cada nó da árvore é um atributo, então se escolhe o atributo que traga o maior ganho de informação possível para cada posição. Métricas possíveis para tomar essa decisão é o ganho de entropia e o índice *geni*.

Onde S e X são características, c os seus valores possíveis e pi a probabilidade p de uma característica i aparecer nas linhas sendo analisadas:

$$Entropia(S) = \sum_{i=1}^{c} -p_i log_2 p_i \tag{2.5}$$

$$Ganho(S,X) = Entropia(S) - Entropia(S|X) \tag{2.6}$$

2.3.4 Floresta Aleatória

A Floresta Aleatória, ou *Random Forest*, é um classificador que combina diversas árvores de decisão para a classificação de novos objetos. Cada árvore é construída utilizando vetores aleatórios de objetos provenientes da base de dados de treinamento, gerando uma série de subgrupos. Embora esses subgrupos possuam o mesmo tamanho da base de treinamento original, os elementos podem variar, pois os objetos nesses vetores são selecionados aleatoriamente e podem se repetir. Além disso, para cada vetor, também são escolhidos aleatoriamente alguns atributos que serão utilizados na análise. A partir dessas listas de vetores e atributos selecionados, cada conjunto de dados forma uma árvore de decisão (RIGATTI, 2017).

Ao inserir um novo objeto a ser classificado, cada árvore retorna uma classe. A classe mais frequente entre as árvores é escolhida como a classificação final do objeto.

2.3.5 Interpretando o resultado

Os resultados de cada classificação serão analisados utilizando métricas providas por uma matriz de confusão. Considerando que este trabalho leva em consideração duas

classes possíveis: ganhador(G) ou perdedor(P), a matriz irá expressar quatro dados diferentes:

- Filmes que ganharam a premiação e foram classificados corretamente (G).
- Filmes que não ganharam a premiação e foram classificados como ganhadores (FG).
- Filmes que perderam a premiação e foram classificados corretamente (P).
- Filmes que ganharam a premiação mas foram classificados como perdedores (FP).

Figura 7 – Um exemplo do funcionamento da matriz de confusão.

Fonte: Autor do trabalho.

Como mostrado na figura 7, é possível observar que cada quadrante irá representar a quantidade de objetos classificados em cada categoria, permitindo então avaliar a acurácia do algoritmo para o caso de uso.

Utilizando os valores providos pela tabela serão utilizados como parâmetro de avaliação:

- **Precisão:** Precisão determina a exatidão dos objetos classificados como ganhadores, com a precisão Pr igual a $\frac{G}{G+FG}$.
- **Acurácia:** Mede o número de acertos do modelo baseado no total de registros testados, ou seja, a acurácia A é igual a $\frac{G+P}{G+P+FG+FP}$.
- **Revocação(ou *Recall*):** A Revocação tem como objetivo classificar com qual exatidão o modelo classificou objetos ganhadores, sendo a Revocação R igual a $\frac{G}{G+FP}$.

- **Pontuação F1(ou *F1-score*):** A Pontuação F1 é uma média harmônia entre a precisão e a revocação. Sua intenção é balancear a proporção entre essas duas métricas e analisar o resultado do modelo de forma mais sensível. A Pontuação F1 F é igual a $2 * \frac{Pr*R}{Pr+R}$

2.4 Trabalhos relacionados

Há diversos trabalhos na área de predição através de aprendizado de máquina que foram utilizados como inspiração e referência para este trabalho.

Em Corrêa (2017) foi desenvolvida uma metodologia para a análise de possíveis ganhadores do Oscar 2017 através de comentários, sobre os filmes indicados, no *twitter*. Para isso, foi utilizado um sistema de análise de sentimentos que buscava classificar as mensagens de usuários do *twitter* através das palavras utilizadas e alimentar algoritmos de aprendizado supervisionado com esses dados com o objetivo de realizar a predição do Oscar 2017.

O trabalho de Masih e Ihsan (2019) utiliza *naive bayes*, redes bayesianas, floresta aleatória e arvore de decisão para estimar o sucesso de filmes indianos de acordo com seus gêneros, atores principais e diretores, identificando se os mesmos já foram premiados anteriormente ou não. O objetivo desse estudo era encaixar o filme em uma possível faixa de retorno em dinheiro, indicando seu sucesso.

Com uma abordagem parecida do primeiro trabalho citado Kim, Hwang e Park (2021) extraíram postagens de usuários de uma rede social chamada *Reddit* para realização de uma análise de sentimentos sobre possíveis ganhadores do Oscar. Foi criado uma metodologia para extrair apenas comentários relacionados aos filmes nomeados e que possuíam alguma relação com a premiação de melhor filme. Desses comentários certas palavras foram agrupadas e suas frequências utilizadas como entrada do algoritmo de aprendizado supervisionado, onde a probabilidade de um filme ganhar ou não a premiação dependia das palavras usadas pelo usuários do *reddit* para se referenciar a ele.

Em sua monografia Bortolini (2022) utilizou algoritmos de aprendizado de máquina para diagnosticar casos de COVID-19 através de seus exames de sangue e urina com o objetivo de auxiliar a detecção da doença na população. Esse trabalho obteve acesso a uma base de dados de exames de pacientes já classificados onde pode extrair, limpar e transformar esses dados para que fossem utilizados na pesquisa. Foram realizados testes com diversas formas de algoritmos de aprendizado e por fim foi determinado o *naive bayes* para a análise final e conclusão do trabalho, comparando os resultados obtidos com as classificações reais dos pacientes. Apesar de não estar diretamente atrelado com o assunto abordado nesse trabalho, a metodologia apresentada serviu de grande referência para a fase de classificação durante o desenvolvimento.

É possível perceber que pesquisas relacionadas a previsão de premiações e classificação de comportamentos de consumidores de cinema estão ocorrendo com frequência atualmente. Através das pesquisas citadas pode-se concluir que os resultados ainda são totalmente conclusivos, porém pode-se perceber que o tema está avançando e recebendo mais atenção a cada dia e já contendo resultados interessantes.

3 Desenvolvimento

Nesta seção serão abordadas todas as etapas do desenvolvimento dessa monografia, além de uma análise final dos resultados obtidos. Nas seções 3.1 e 3.2 será mostrado todo o processo de extração e descoberta dos dados, falando sobre a escolha do Metacritic como fonte, o método utilizado para obtenção das avaliações e como ele foi implementado. Na seção 3.3 será mostrado como os dados foram organizados, assim como cada atributo foi escolhido e calculado, mostrando sua distribuição por cada classe. Na seção 3.4 será mostrado como os dados foram separados para treino e teste, quais algoritmos foram utilizados, e porque foram utilizados, além das métricas obtidas em cada modelo. Na seção 3.5 será analisado o resultado final do desenvolvimento, aplicando os modelos nos indicados do ano de 2023 e verificando se a hipótese inicial foi concretizada.

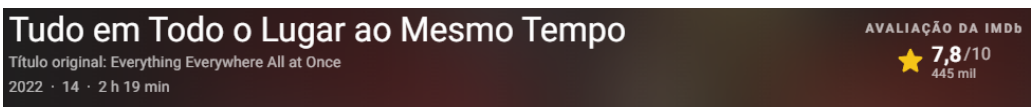
O código fonte, assim como as bases de dados, utilizado nesse trabalho pode ser encontrado no repositório <https://github.com/kauelso/oscars-golden-globe-prediction>

3.1 Escolha do agregador

Antes de realizar qualquer extração dos dados necessários, primeiro será necessário escolher qual fonte de dados será utilizada. Para isso, foram consideradas três opções dentre os maiores agregadores de críticas da Internet: O IMDB, o Rotten Tomatoes e o Metacritic.

Iniciando com o IMDb (2022b), em inglês *internet Movie Database*, sua principal vantagem vem do grande volume de filmes e séries disponíveis com uma quantidade enorme de avaliações, tendo sua existência precedendo sequer a data do primeiro navegador de Internet (ROTHMAN, 2015). Um ponto positivo para esse domínio seria a existência de uma *api* pública (IMDB, 2021) para a coleta dos dados , providenciada pela AWS (Amazon Web Service). Por outro lado o sistema de notas do IMDB não possui diferenciação entre tipos diferentes de usuários, mesclando notas de críticos e usuários comuns em uma mesma categoria, impossibilitando distinguir os dois tipos de avaliação. Devido a essa limitação, esse *website* foi descartado inicialmente.

Figura 8 – Exemplo de nota geral do IMDb para o filme *Tudo em Todo Lugar ao Mesmo Tempo*



Fonte: IMDb (2022a).

O segundo candidato como fonte de informação seria o RottenTomatoes (2022b). Com o objetivo de classificar filmes de uma forma inusitada e um pouco lúdica, esse agregador possui dois tipos de avaliação: Para a crítica especializada é feito um sistema onde notas são extraídas de diversos jornais e portais especializados e convertidos em um sistema binário de pontuação onde a obra é classificada com um tomate fresco(positivo) ou um tomate apodrecido(negativo) e com isso a qualidade do filme é avaliada pela relação entre votos positivos e o total de votos. Para os usuários comuns é permitido realizar uma avaliação de 1,0 a 5,0. Não foi encontrado nenhuma interface pública oficial para a extração dos dados dessa base de dados, apesar disso, ele se mantém como alternativa possível já que possui os dois tipos de notas necessários para o trabalho.

Figura 9 – Exemplo de nota geral do Rotten Tomatoes para o filme *Tudo em Todo Lugar ao Mesmo Tempo*

Fonte: RottenTomatoes (2022a).

O terceiro e último candidato é o Metacritic (2022b). Também de muita relevância e com uma vasta base de dados, apesar de não ter um foco primário em filmes e ser menos reconhecido que as outras duas opções nessa categoria, o Metacritic é uma fonte sólida para a extração das notas. Suas avaliações consistem em um sistema de notas que varia de 0 a 100 para a crítica especializada e de 0,0 a 10,0 para os usuários da plataforma, gerando ao final um *metascore* que avalia a qualidade da obra. Assim como o Rotten Tomatoes esse *website* não possui uma *api* pública oficial, tornando a extração dos dados possível primariamente via *web scraping*. Com uma base sólida de avaliações e duas medidas não binárias de notas, o Metacritic parece ser uma opção boa e flexível para o problema proposto nesta monografia, portanto será utilizado como fonte de dados para este trabalho.

Figura 10 – Exemplo de nota geral do Metacritic para o filme *Tudo em Todo Lugar ao Mesmo Tempo*

Fonte: Metacritic (2022a).

3.2 Extração dos dados

A extração de dados é uma etapa sensível e vital para o sucesso do projeto. Garantir que os dados foram coletados de forma correta e segura é parte fundamental para garantir a qualidade das conclusões tomadas. Para obter os dados necessários foram executadas quatro etapas: a escolha do período que seria extraído e quais avaliações seriam consideradas, criação da base de dados com o nome dos filmes indicados e suas respectivas classes (ganhador ou perdedor), criar a mesma relação com as obras indicadas do ano de 2023 e por fim extrair os conjuntos de notas de cada filme. Foram consideradas três bases de dados diferentes para análise do trabalho. Para o Oscar foi considerado a categoria de melhor filme como citado anteriormente, já no Globo de Ouro não há uma categoria única equivalente, portanto foram consideradas tanto a categoria de melhor drama quanto a categoria de melhor comédia, extraídas separadamente.

Para definir o período de onde os dados seriam extraídos foi preciso entender a partir de que ano o Metacritic já se tornaria uma fonte sólida de dados. Como o *site* foi lançado no ano de 2001 apenas premiações após esse ano poderiam ser consideradas. Além da data de criação do portal também é importante entender quando ele começa a ter um volume considerável de informação, pelo menos para os filmes indicados. Analisando as

opções disponíveis, foi decidido manter a base de treinamento contemplando as cerimônias de 2007 até 2022. Após escolher o período da extração outra questão se torna importante, quais avaliações seriam consideradas válidas? Para isso surgiram três alternativas: considerar todas as notas disponíveis, considerar apenas as notas anteriores à premiação e por fim considerar apenas as notas entre o período da indicação até data do evento. A primeira opção fornece o maior número de dados, porém com um alto risco dessas notas serem influenciadas pelo próprio Oscar ou Globo de Ouro em *reviews* realizados após as cerimônias. A segunda opção parece sólida e considera toda a jornada do filme anterior à qualquer premiação, com um volume razoável de análises. A terceira opção é interessante, apesar de possuir um menor volume de dados ela permitiria entender qual o impacto de um filme ser sequer nomeado para um dos dois prêmios. A opção escolhida para esse caso foi a segunda opção por garantir um número razoável de amostras e já ser suficiente para o objetivo em questão.

O segundo passo para montar a base de dados consiste em colher o nome dos filmes indicados em cada evento, o ano de lançamento do filme(para que seja realizado uma validação ao colher as notas em casos de existirem dois filmes com o mesmo nome), a data da premiação(para que avaliações após a cerimônia não fossem consideradas) e a classificação de cada filme. Como ponto de partida foram utilizados duas planilhas disponíveis no *website Kaggle* com licença de domínio público: Fontes (2022) e Fontes (2020). Esses *datasets* possuem todos os filmes participantes, de todas as categorias, de 1944 a 2020 no caso do Globo de Ouro e de 1927 a 2022 no caso do Oscar, e sua classificação final. A partir desses arquivos foram separados apenas os indicados,dos anos de 2007 a 2022, da categoria de melhor filme, no caso do Oscar,e da categoria de melhor drama e melhor comedia no caso do Globo de Ouro. Ao fim foram removida as colunas desnecessárias, sobrando apenas as colunas com nome, ano e classe. Os filmes de 2021 e 2022 que não estavam presentes foram adicionados manualmente juntamente de uma coluna indicando o ano de lançamento do filme, ao mesmo tempo que a coluna com o ano da premiação foi substituído pela data exata.

Para a terceira etapa, onde as obras de 2023 foram colhidas, as informações foram tiradas diretamente dos portais de cada cerimônia, montando um arquivo *csv*(valores separados por virgulas) no mesmo formato da etapa anterior. O resultado pode ser conferido nas tabelas 1, 2 e 3 onde é mostrado cada um dos filmes indicados, sua classificação final e também a média das notas extraídas, de usuários e críticos.

Por fim é feito a coleta de dados diretamente do site Metacritic. Nesse processo foi utilizado a linguagem *python3*, também utilizado nas fases de pré-processamento e treinamento, junto das bibliotecas *requests*, responsável por obter o código HTML das páginas do *site*, e *Beautiful Soup*, responsável por toda a raspagem dos dados. Para obter o endereço de cada página respectiva de cada filme foi observado que as páginas do Metacrtitic

Tabela 1 – Lista de filmes indicados ao Globo de Ouro na categoria drama.

Nome	**Média Críticos**	**Média Usuários**	**Classe**
Avatar: The way of water	6,84	6,97	Perdedor
Elvis	6,45	7,10	Perdedor
The Fablemans	8,14	6,25	Perdedor
TÁR	8,76	5,47	Perdedor
Top Gun: Maverick	7,75	9,15	Perdedor

Tabela 2 – Lista de filmes indicados ao Globo de Ouro na categoria comédia.

Nome	**Média Críticos**	**Média Usuários**	**Classe**
The Banshees of Inisherin	8,52	6,26	Perdedor
Babylon	6,38	6,39	Perdedor
Everything Everywhere All at Once	8,09	8,13	Perdedor
Glass Onion: A Knives Out Mistery	7,90	4,98	Perdedor
Triangle of Sadness	6,70	6,54	Perdedor

Tabela 3 – Lista de filmes indicados ao Oscar.

Nome	**Média Críticos**	**Média Usuários**	**Classe**
Everything Everywhere All at Once	8,09	8,11	Perdedor
All Quiet on the Western Front	7,59	7,20	Perdedor
Avatar: The way of water	6,84	6,93	Perdedor
The Banshees of Inisherin	8,52	6,38	Perdedor
Elvis	6,45	7,30	Perdedor
The Fablemans	8,14	6,94	Perdedor
TÁR	8,76	5,73	Perdedor
Top Gun: Maverick	7,75	9,08	Perdedor
Triangle of Sadness	6,70	6,25	Perdedor
Women Talking	7,68	7,00	Perdedor

possuem o seguinte padrão: *https://www.metacritic.com/movie/nome-do-filme*, sendo que o nome do filme é escrito em minúsculo, com os espaços e caracteres especiais substituídos por hífen, com exceção do caractere de exclamação, e em casos onde há mais de uma obra com o mesmo nome, a última acompanha o seu ano ao fim do endereço, como por exemplo

o filme *CODA* que possui o endereço <https://www.metacritic.com/movie/coda-2021>. Feito essa observação foi implementado uma função que usa uma expressão regular para adaptar o nome do indicado para o padrão necessário e, caso a página não for encontrada, adiciona o ano de lançamento ao fim do endereço para tentar novamente. Para acessar a lista de notas de usuários, basta adicionar *user-reviews* ao fim da URL, já para acessar análises de críticos basta fazer o mesmo com o caminho *critic-reviews*. Antes de extrair os dados, primeiramente é verificado se não se trata de um falso positivo, comparando o ano de lançamento presente na página com o ano informado anteriormente na base de dados, se for esse o caso é adicionado o ano correto ao final do nome do filme e o processo reiniciado. Com o HTML correto em mãos as notas são extraídas da *div* que possui a classe *user-reviews*, ou *critic-reviews*, onde há uma lista de *divs* que possuem um filho com a classe *metascore_w*, tendo a nota da avaliação. Em cada análise de cada usuário também é verificado sua data através de um *span* com a classe *date* e caso seja posterior à data da premiação, é desconsiderado. Caso o elemento de seleção de página seja encontrado ao fim do documento e o botão *next* esteja disponível, ele é acionado levando o programa a colher as informações da próxima página, até que não existam mais páginas para avançar, indicando o fim da coleta para a obra em questão. Ao final todos os dados colhidos são colocados em um JSON onde cada filme é uma chave e cada chave possui duas listas: uma lista de notas de críticos e outra lista de notas do público. O processo todo é executado seis vezes ao total gerando seis novos arquivos com dados, três para dados de treinamento e três para os dados de 2023, englobando as três categorias.

3.3 Pré-processamento e seleção das características

Pela natureza da base de dados utilizada não foi necessário um grande trabalho de pré-processamento. O principal ponto abordado nessa fase foi a normalização das notas de críticos e a padronização de alguns campos inconsistentes entre as bases após a extração. Observando a metodologia de notas do Metacritic, as notas de críticos e usuários possuem magnitudes diferentes, sendo as notas de críticos indo de 0 a 100, enquanto usuários avaliam com notas de 0,0 a 10,0. Portanto as notas de 0 a 100 foram normalizadas para o intervalo [0,10]. Isso foi feito para reduzir a influência dessa diferença de magnitude em algoritmos baseados em distância, como o KNN.

Com os valores normalizados, foram extraídos doze características na versão final do trabalho, sendo 6 para cada tipo de avaliação. Como inicialmente tem-se apenas duas listas dinâmicas de notas para cada item é preciso utilizar de algumas medidas para entender as características desses vetores e permitir que padrões sejam encontrados. As medidas extraídas foram: média,moda, desvio padrão, mediana e dois percentis, o de 25% e o de 75%. Com a média das notas é possível perceber o desempenho geral da obra na categoria analisada, já a moda permite saber quais notas mais se repetem e encontrar

possíveis semelhanças entre os ganhadores, principalmente analisando a moda das análises de críticos. Mediana, desvio padrão e os percentis permitem verificar a distribuição dos valores, o que pode ser importante para entender qual foi o intervalo de notas que o filme teve mais avaliações e o quão dispersa elas são. Um exemplo das amostras da base de dados pode ser visto na tabela 4. Ao fim do processo, é possível analisar a distribuição dos valores nas figuras 11, 12, 13, 14, 15 e 16:

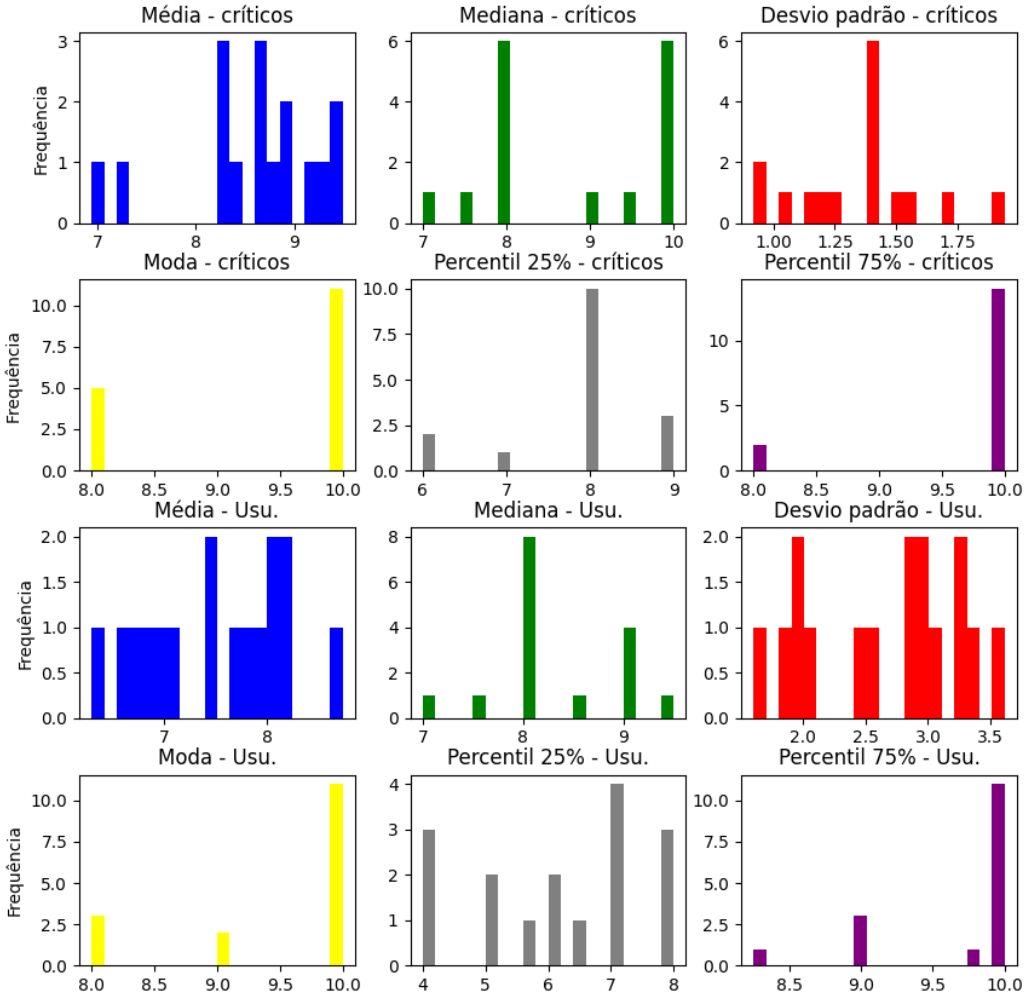
Tabela 4 – Amostra da base de dados dos filmes do Oscar após a geração de atributos.

Ano de lançamento	**Data da cerimônia**	**Nome**	**Ganhador**
2006	25/02/2007	Babel	FALSO
2010	27/02/2011	The King's Speech	VERDADEIRO
2014	22/02/2015	Whiplash	FALSO

A partir dos histogramas gerados é possível notar algumas observações sobre os atributos gerados:

Os candidatos a melhor filme pelo Oscar possuem uma diferença levemente pronunciada entre as classes. É possível observar que nos filmes ganhadores a média se mantém acima dos 8,0, com alguns filmes chegando próximo da média 10,0, no caso de críticos, e acima dos 7,0, no caso de usuários, na maioria dos filmes. Enquanto isso as obras perdedoras possuem médias mais bem distribuídas com várias notas abaixo dos 7,0 pontos, porém com a maioria dos filmes ainda atingindo por volta de 8,0 de média, tanto para críticos quanto usuários. A moda também mostra uma pequena diferença entre as duas classes: apesar de ambos vencedores e perdedores possuírem uma grande quantidade de modas 10,0, enquanto os filmes que obtiveram sucesso estão com moda entre 8 e 10,0, principalmente no caso de usuários, quando se olha para as estatísticas de perdedores há alguns candidatos fora desse intervalo, com casos, apesar de raros, de modas abaixo de 6,0. Por fim olhando para as medidas de distribuição, como desvio padrão, mediana e os percentis de 25% e 75%, o cenário já se torna diferente. Primeiramente observando notas da crítica especializada, os indicados vencedores possuem um desvio padrão mais próximo de 1,0, com uma mediana que varia entre 7,0 e 10,0 e com percentis que demonstram que a grande maioria dos filmes possuem a maior parte das notas acima dos 8,0, mostrado pelo percentil de 25%. Já os perdedores estão com um desvio ligeiramente mais alto, demonstrando notas um pouco mais dispersas, porém ainda dentro ou próximo do intervalo entre 1,0 e 1,5, enquanto a mediana decai comparado com a categoria anterior, apesar de muitos ainda com o valor 8 nesse atributo. Olhando para os percentis, de 25 e 75 por cento, já há uma grande diferença com objetos que possuem avaliações bem mais baixas tendo percentis de 25% de 4,0 e percentis de 75% com nota 6,0. Analisando agora por fim notas do público a diferença se torna menos perceptível. Ambas as classes presentes possuem um desvio padrão em maior parte dentro do intervalo [1,4], com os ganhadores bem esparsos entre 1,5 e 3,5, já os perdedores mais condensados entre o 2,0 e 3,5. Percentis

Figura 11 – Histogramas com atributos dos filmes classificados como ganhadores do Oscar.

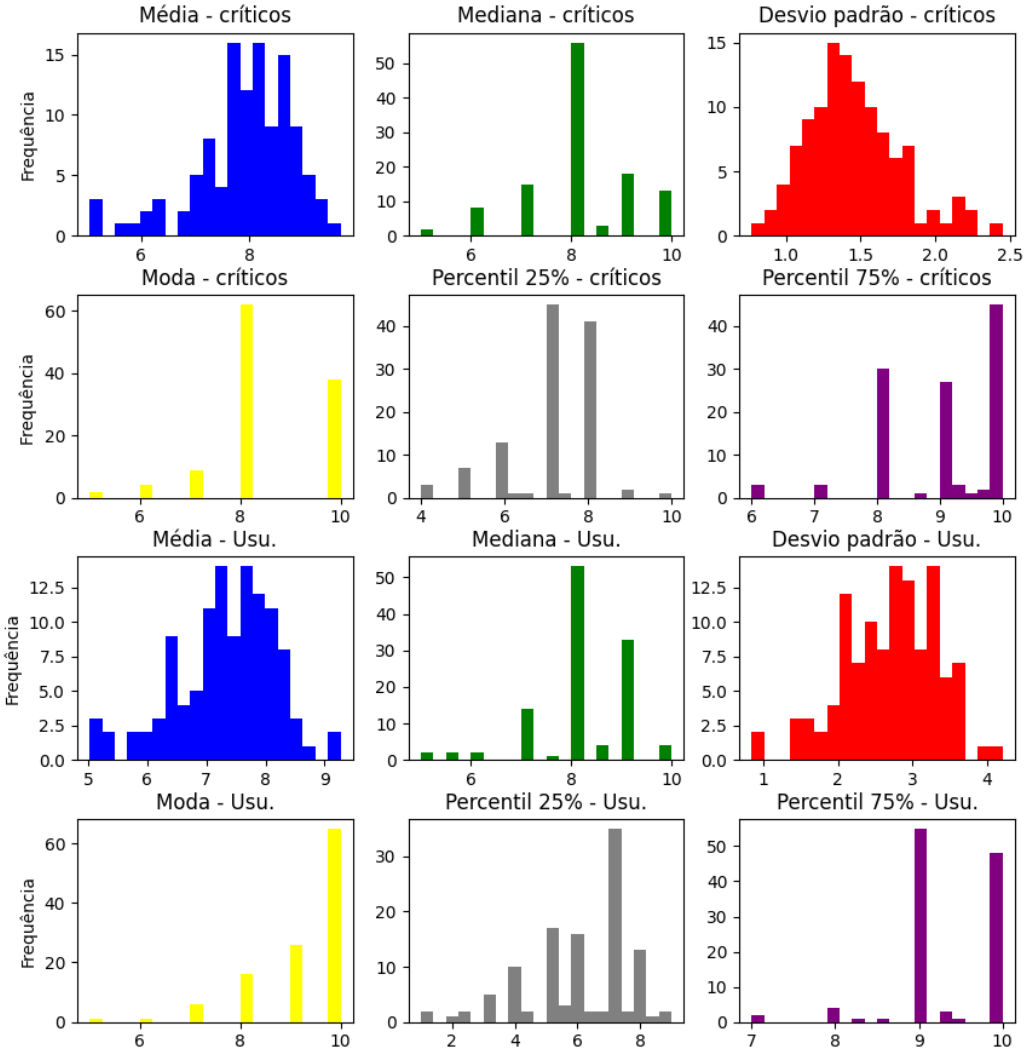


Fonte: Autor do trabalho.

de 75% também não possuem tanta diferença com a maior parte dos filmes dentro de 9,0 a 10,0 de nota nessa métrica. A maior diferença fica por conta do percentil de 25% onde indicados derrotados são distribuídos em valores que vão de 1,0 a 9,0, com a área mais condensada sendo o intervalo de 5 a 8. Os premiados já possuem uma distribuição que varia de 4,0 a 8,0 com uma quantidade relevante de obras em todo o intervalo.

As características dos filmes na categoria drama do Globo de Ouro são um pouco diferentes em relação ao Oscar. Olhando para as médias, de ambos os tipos, perdedores e vencedores possuem números muito parecidos, com a maioria dos objetos se aglomerando no mesmo intervalo. A moda, considerando críticos, também é próxima entre as duas classes, porém com mais filmes com moda 8,0 entre os perdedores e com moda 10,0 entre os ganhadores. Considerando a opinião pública já se vê uma certa divergência em relação

Figura 12 – Histogramas com atributos dos filmes classificados como perdedores do Oscar.



Fonte: Autor do trabalho.

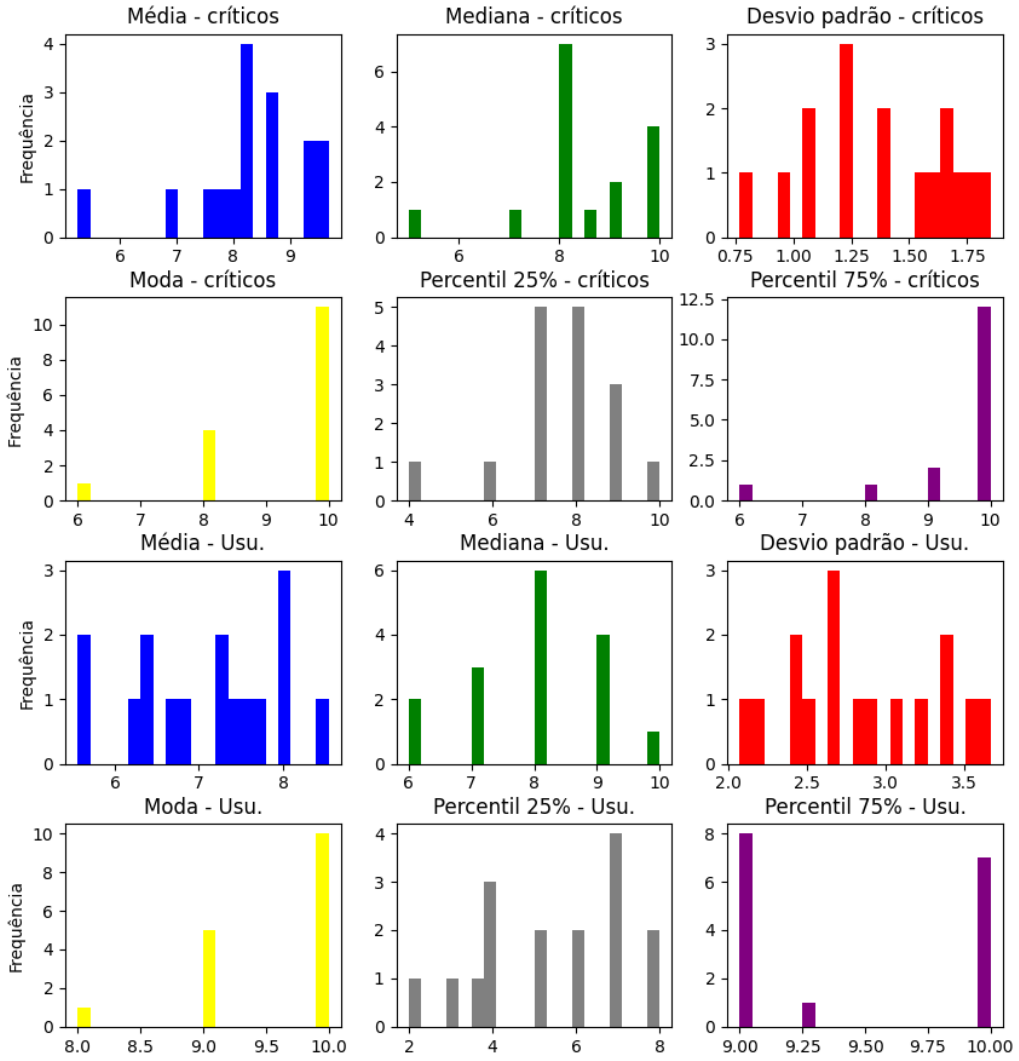
à moda, com obras premiadas tendo números sempre acima dos 8,0 enquanto as outras obras, apesar de ter sua grande maioria nesse intervalo, possui alguns indicados abaixo desse limiar. As outras métricas se mostram bem parecidas dentre as duas categorias: Os percentis estão praticamente no mesmo intervalo, sendo que há produções premiadas tendo percentil de 25% de apenas 4,0 pontos, nas notas de especialistas, enquanto o piso para os perdedores é de 5,0 pontos. O desvio padrão dessas notas também se assemelha nos dois casos: Para críticos a pontuação de vencedores costuma ser menos esparsado. Já para usuários essa métrica é bem parecida em ambos rótulos, com alguns filmes perdedores tendo um desvio menor.

Considerando os indicados como melhor comédia, também do Globo de Ouro, existe um cenário um pouco mais interessante. Diferente da categoria anterior, essa pre-

miação possui algumas diferenças entre as duas classes. Começando pela média e pelos percentis, é possível perceber que ganhadores costumam ser mais bem avaliados do que perdedores, com médias e medianas maiores para críticos e usuários. Os percentis do público não mostram uma grande mudança, porém para a critica especializada a diferença é mais aparente: o percentil de 75% das notas se mantém acima dos 8,0 pontos em ganhadores, enquanto perdedores possuem várias obras no intervalo de 4,0 a 8,0 pontos. O percentil de 25% também mostra um grande variação com premiados tendo um valor, pelo menos, acima de 5,0 e derrotados com várias obras abaixo desse limiar. As medianas de ambas as classes, considerando ambos os tipos de avaliação, também tem uma leve variação ficando com a maioria dos objetos com mediana 8,0, porém com um piso bem mais baixo para perdedores.

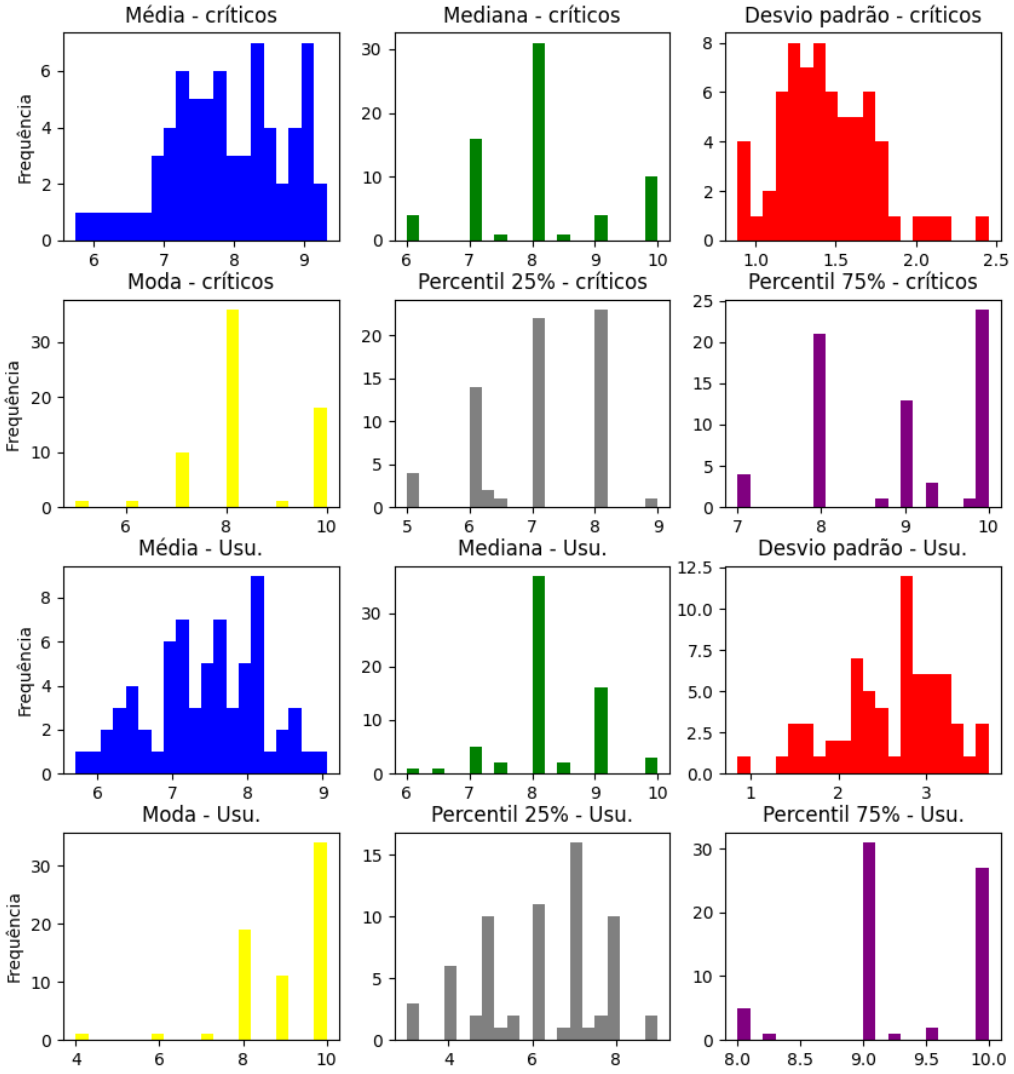
É possível observar que há pequenas diferenças em relação a alguns objetos das diferentes classes nas três categorias, porém a maioria dos filmes indicados se mostra similar aos ganhadores. Apesar de muitas produções que fracassaram possuírem métricas bem abaixo do esperado, também há uma grande quantidade desses filmes que possuem bons números. A utilização de modelos de aprendizado de máquina, presentes na próxima seção, irá ajudar a entender se, apesar disso, existem possíveis padrões que podem ser aprendidos para classificar novos objetos, ajudando a entender se há alguma influência dessas notas nos possíveis premiados.

Figura 13 – Histogramas com atributos dos filmes classificados como ganhadores do Globo de Ouro de melhor drama.



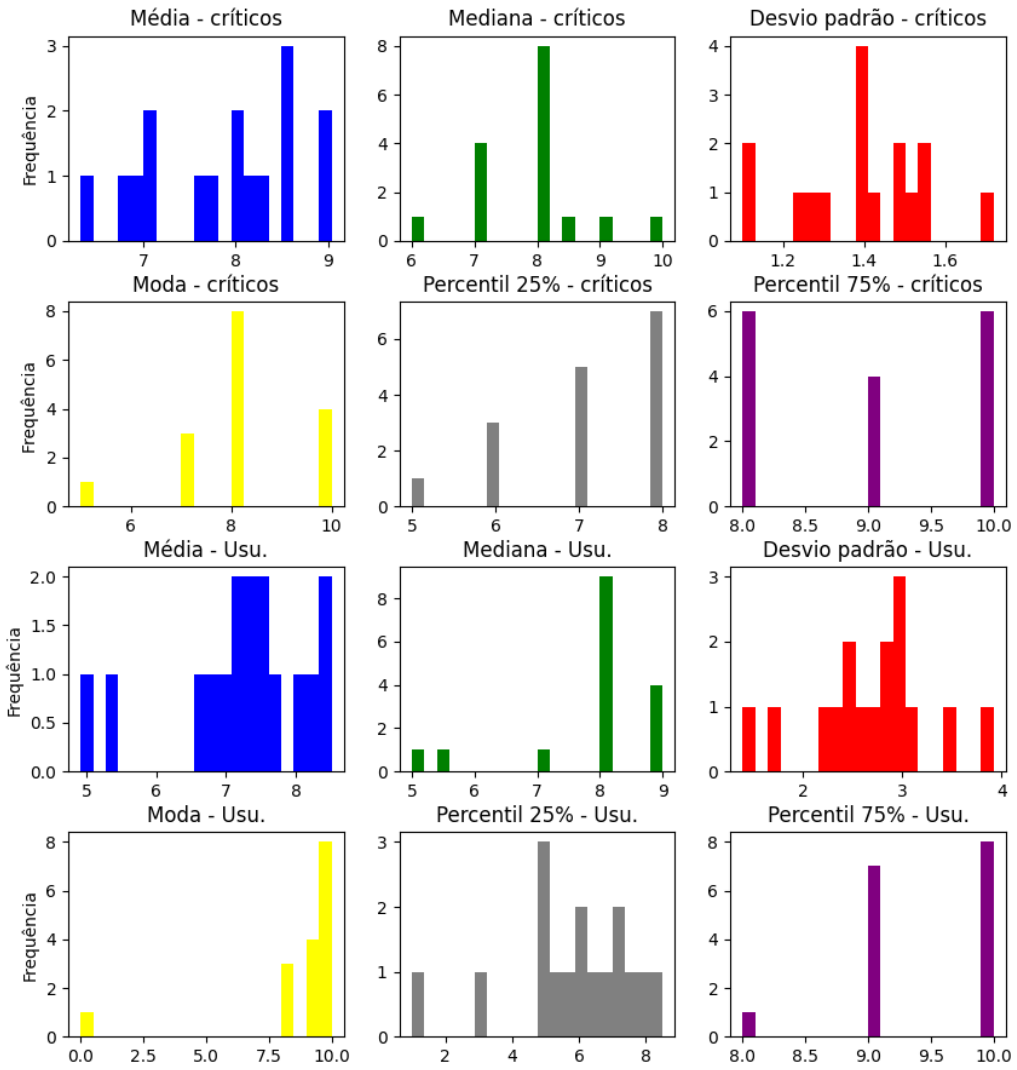
Fonte: Autor do trabalho.

Figura 14 – Histogramas com atributos dos filmes classificados como perdedores do Globo de Ouro de melhor drama.



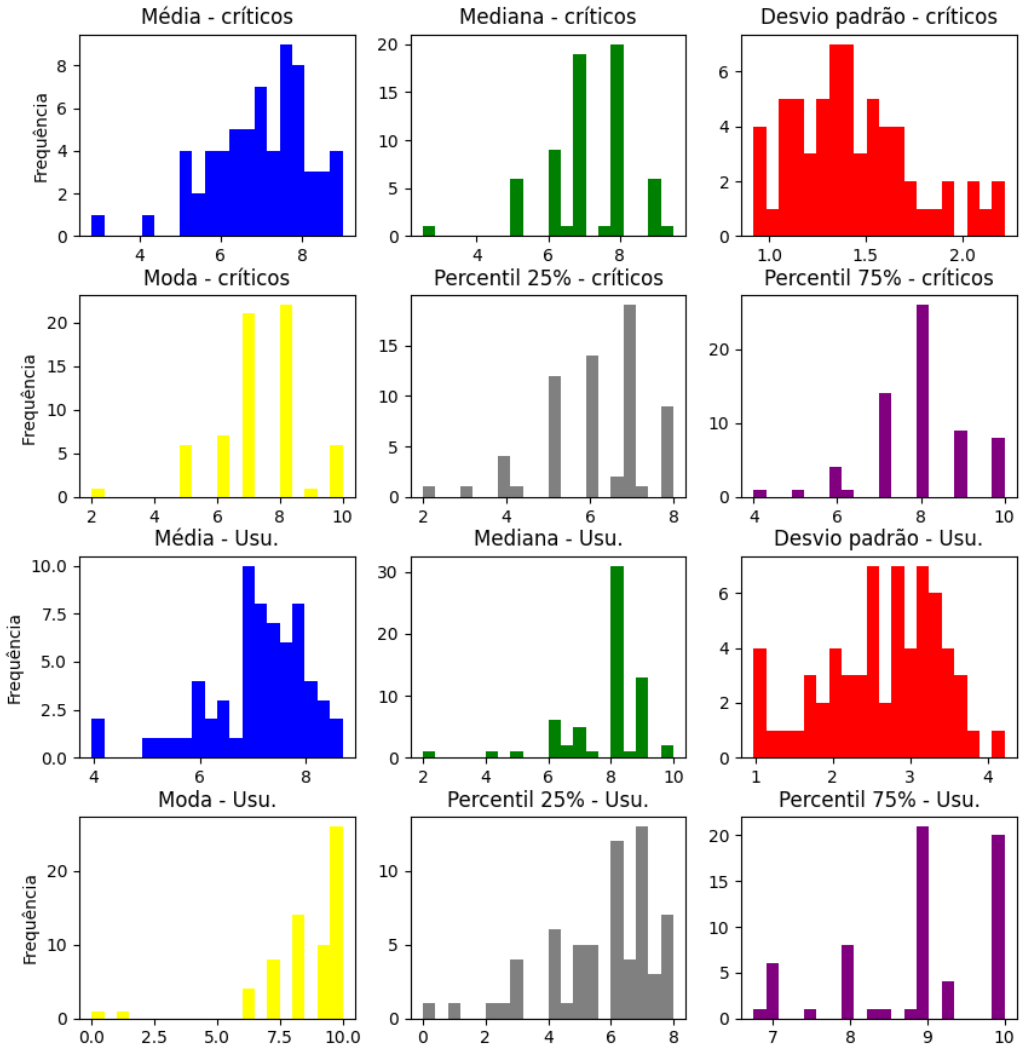
Fonte: Autor do trabalho.

Figura 15 – Histogramas com atributos dos filmes classificados como ganhadores do Globo de Ouro de melhor comédia.



Fonte: Autor do trabalho.

Figura 16 – Histogramas com atributos dos filmes classificados como perdedores do Globo de Ouro de melhor comédia.



Fonte: Autor do trabalho.

3.4 Treinamento e avaliação dos modelos

Com os dados disponíveis, e as características de cada observação extraídas, é possível então treinar modelos de aprendizado para verificar possíveis padrões que permitam a previsão dos premiados. Para essa fase foram escolhidos três algoritmos: *naive bayes*, KNN e floresta aleatória. O objetivo é avaliar o desempenho desses algoritmos em bases de testes, utilizando as métricas de acurácia, precisão, revocação e pontuação F1, em cada base de dados, e posteriormente verificar os possíveis vencedores em 2023 através de suas probabilidades de estarem na classe positiva. Foram utilizadas as bibliotecas Scikit-learn (2023) e imblearn (2023) para todos o processo de treinamento e avaliação dos modelos. Para manter a consistência entre todas as execuções todas as funções que possuem a propriedade *random_state* tiveram esse atributo atribuído com o valor 80, mantendo assim sempre a mesma semente para gerações de números aleatórios.

Para separar as bases de dados em grupos de treino e teste foi utilizado a função *train_test_split* presente no *scikit-learn*. A função foi configurada para gerar o grupo de teste com 20% dos dados originais, além de utilizar a opção *stratify* para separar as categorias de forma mais homogênea. Um grande problema que surge a partir deste ponto é a diferença no número de objetos de cada classe, tornando o treinamento muito desbalanceado. Por isso tanto para a fase de avaliação, quanto para o treinamento com a base completa, foi utilizado o objeto *RandomOverSampler* que permite replicar observações com o rótulo em menor número no conjunto de treino, balanceando as duas classes. Antes do balanceamento os grupos de treino possuíam as seguintes quantidades:

- **Oscar:** 91 filmes perdedores e 13 ganhadores.
- **Globo de Ouro (Drama):** 53 filmes perdedores e 13 filmes ganhadores
- **Globo de Ouro (Comédia):** 51 filmes perdedores e 13 filmes ganhadores

Após o processo de balanceamento o número de filmes em cada classe se igualou, duplicando os vencedores diversas vezes. Os filmes separados para cada grupo de teste estão expostos nas tabelas 5, 6 e 7.

A partir das bases de treino e teste os algoritmos foram configurados e treinados. Para o algoritmo *naive bayes* foi aplicado sua versão gaussiana. Para o KNN foi utilizada a distância euclidiana como medida de dissimilaridade. O valor de k foi definido a partir de executado em cada *dataset*, executando o modelo com k de 1 a 30 e verificando qual número de vizinhos possuía melhor medida F. Para o Oscar foi escolhido $k = 14$ e para as premiações do Globo de Ouro foi escolhido $k = 5$ para drama e $k = 15$ para comédia. Para o treino do algoritmo de floresta aleatória foi mantido os valores padrões da função *RandomForestClassifier* do *scikit-learn* na maioria dos parâmetros. Os únicos parâmetros

Tabela 5 – Lista de filmes separados para testes na base de dados do Ocars.

Filme	**Classe**
Judas and the Black Messiah	Perdedor
Juno	Perdedor
Dune part one	Perdedor
An Education	Perdedor
Three Billboards outside Ebbing, Missouri	Perdedor
Atonement	Perdedor
A Star Is Born	Perdedor
Nebraska	Perdedor
Dunkirk	Perdedor
Life of Pi	Perdedor
No Country for Old Men	Ganhador
Amour	Perdedor
Whiplash	Perdedor
The Queen	Perdedor
Her	Perdedor
The Irishman	Perdedor
Captain Phillips	Perdedor
Mad Max: Fury Road	Perdedor
Phantom Thread	Perdedor
Spotlight	Ganhador
Frost/Nixon	Perdedor
Bohemian Rhapsody	Perdedor
Milk	Perdedor
The Hurt Locker	Ganhador
Inception	Perdedor
The Curious Case of Benjamin Button	Perdedor
Moneyball	Perdedor

modificados foram: ganho de informação alterado de índice *gini* para entropia e o número mínimo de observações em cada nó folha, depois de vários experimentos com valores de 2 a 30, foi definido como 15 para Oscar e 3 para ambas as premiações do globo de Ouro. Outras métricas não foram alteradas pois não surtiram grandes efeitos nos testes. As métricas obtidas, tendo como referência a classe 'Ganhador', podem ser observadas na tabelas 8, 9 e 10. As matrizes de confusão estão presentes nas imagens 17, 18 e 19.

Após extrair as métricas de cada modelo proposto usando apenas a base de treino, por fim foi realizado o treinamento completo dos algoritmos. Com os modelos treinados, foi então extraída a probabilidade de cada filme indicado (através do método *predict_proba*), em cada premiação, de estar em cada uma das duas classes. As tabelas 11, 12 e 13 apresentam o chance de cada filme estar na classe 'Ganhador', ou seja, ser vencedor na sua categoria de acordo com cada algoritmo.

Tabela 6 – Lista de filmes separados para testes na base de dados do Globo de outro (Drama).

Filme	**Classe**
No Country for Old Men	Perdedor
The Theory of Everything	Perdedor
The Shape of Water	Perdedor
Call Me By Your Name	Perdedor
The Irishman	Perdedor
The Fighter	Perdedor
Avatar	Ganhador
BlacKkKlansman	Perdedor
Nomadland	Ganhador
Up In The Air	Perdedor
The Curious Case of Benjamin Button	Perdedor
Mank	Perdedor
The Queen	Perdedor
War Horse	Perdedor
Moneyball	Perdedor
Argo	Ganhador
Inglourious Basterds	Perdedor

Tabela 7 – Lista de filmes separados para testes na base de dados do Globo de outro (Comédia).

Filme	**Classe**
Burlesque	Perdedor
I Tonya	Perdedor
Midnight In Paris	Perdedor
In Bruges Name	Perdedor
The Hangover	Ganhador
Trainwreck	Perdedor
West Side Story	Ganhador
My Week With Marilyn	Perdedor
Nine	Perdedor
The Tourist	Perdedor
Hamilton	Perdedor
Mary Poppins Returns	Perdedor
Rocketman	Perdedor
Borat Cultural Learnings Of America For Make Benefit Glorious Nation Of Kazakhstan	Perdedor
Les Miserables	Ganhador
Vice	Perdedor

Tabela 8 – Resultado da avaliação dos algoritmos de aprendizado com 20% da base de dados do Oscar usada como base de teste.

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-score**
Naive bayes	48,14%	100%	17,64%	30%
KNN	74,07%	100%	30%	46,15%
Floresta Aleatória	70,37%	100%	27,27%	42,85%

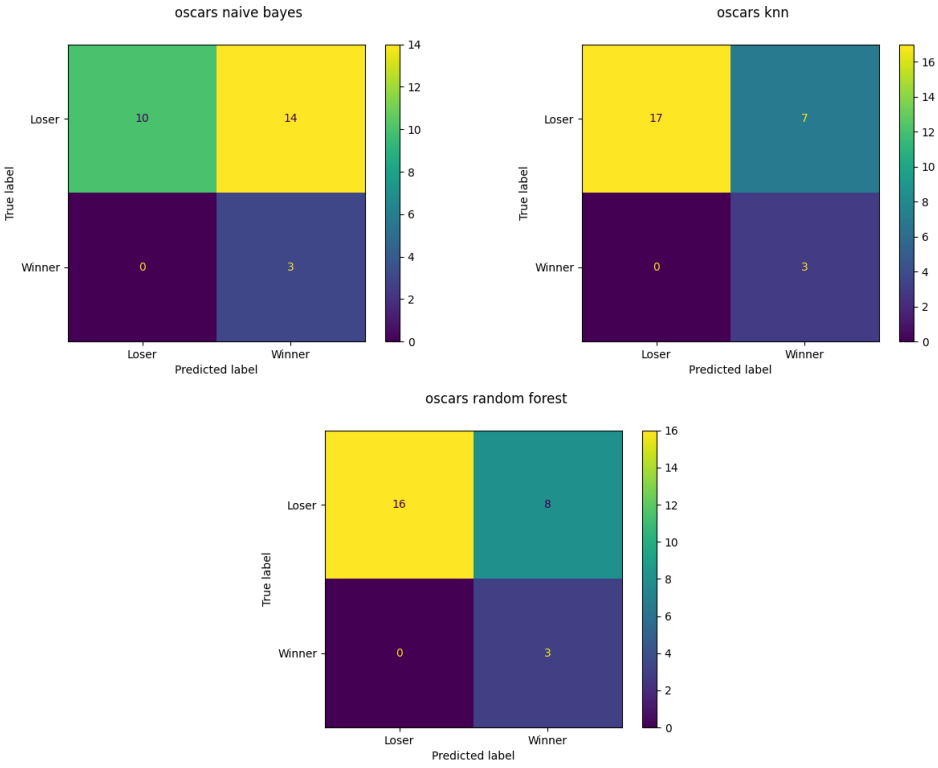
Tabela 9 – Resultado da avaliação dos algoritmos de aprendizado com 20% da base de dados do Globo de Ouro (Drama) usada como base de teste.

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-score**
Naive bayes	58,82%	66,66%	25%	36,36%
KNN	70,58%	100%	37,5%	54,54%
Floresta Aleatória	76,47%	33,33%	33,33%	33,33%

Tabela 10 – Resultado da avaliação dos algoritmos de aprendizado com 20% da base de dados do Globo de Ouro (Comédia) usada como base de teste.

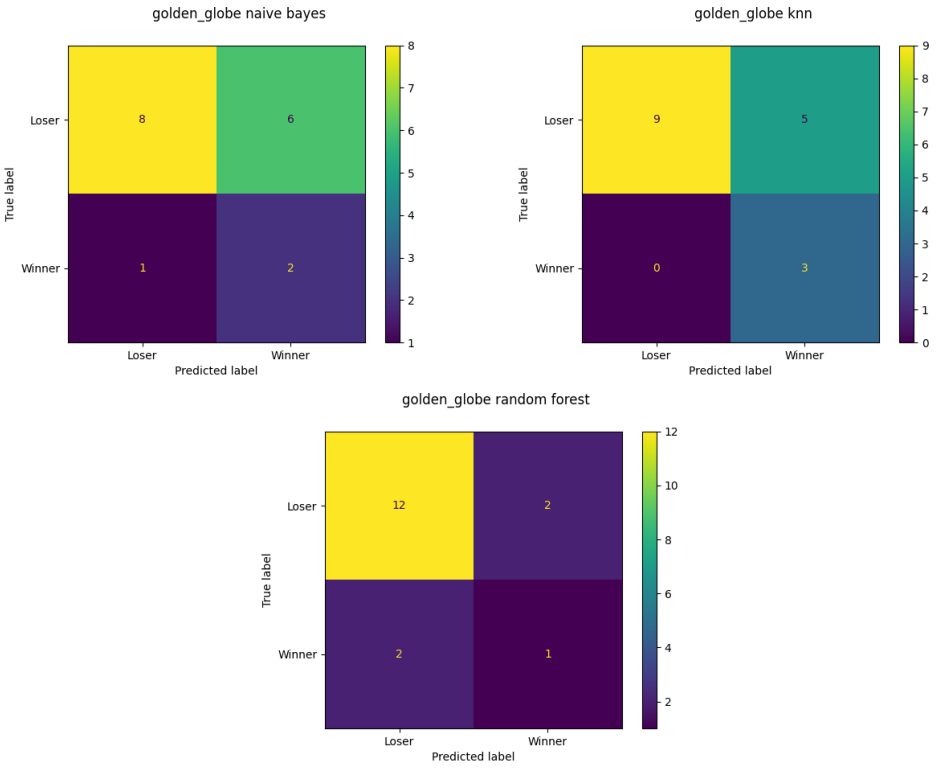
Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-score**
Naive bayes	68,75%	33,33%	25%	28,57%
KNN	81,25%	66,66%	50%	57,14%
Floresta Aleatória	87,5%	33,33%	100%	50%

Figura 17 – Matriz de confusão gerada a partir das previsões dos modelos na base de teste do Oscar.



Fonte: Autor do trabalho.

Figura 18 – Matriz de confusão gerada a partir das previsões dos modelos na base de teste do Globo de Ouro (Drama).



Fonte: Autor do trabalho.

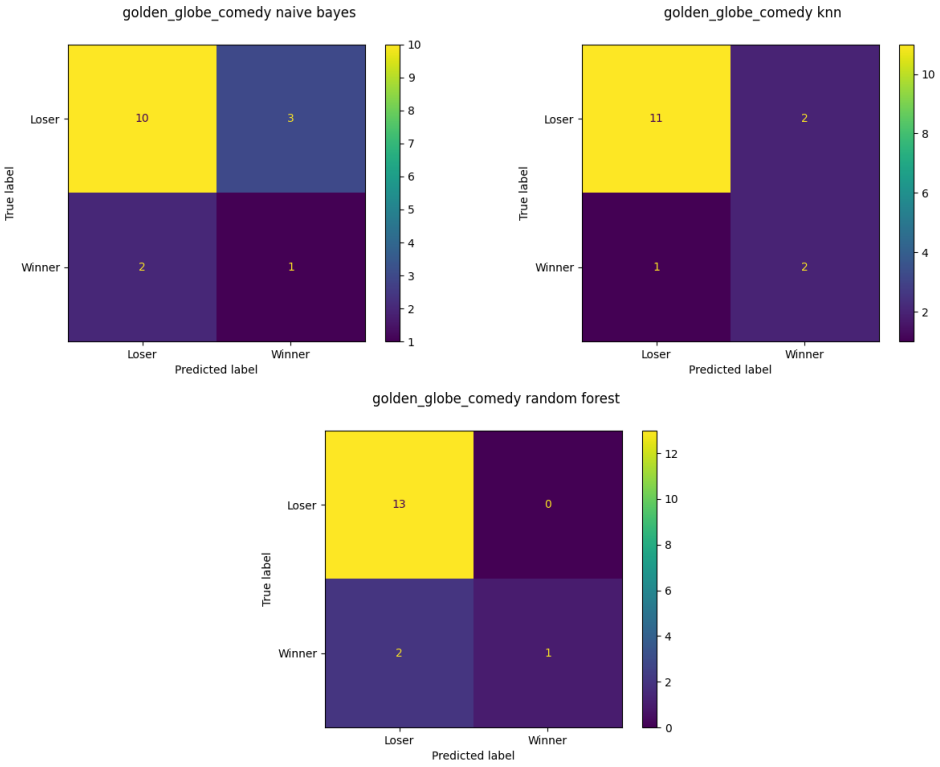
Tabela 11 – Probabilidade de cada filme indicado ao Oscar de melhor filme em 2023 ser vencedor de acordo com cada modelo.

Filme	**Naive bayes**	**KNN**	**Floresta Aleatória**
Everything Everywhere All at once	27,61%	7,14%	21,84%
Avatar: The Way of Water	0,23%	0%	3,12%
All Quiet on The Western Front	0,06%	0%	9,33%
The Banshees of Inisherin	49,15%	21,43%	36,04%
Elvis	0%	14,29%	9,88%
The Fabelmans	33,38%	0%	4,52%
TÁR	0,06%	57,14%	58,02%
Top Gun: Maverick	0,44%	50%	41,37%
Triangle of Sadness	0%	0%	9,64%
Women Talking	0,01%	0%	3,66%

Tabela 12 – Probabilidade de cada filme indicado ao Globo de Ouro de melhor drama em 2023 ser vencedor de acordo com cada modelo.

Filme	**Naive bayes**	**KNN**	**Floresta Aleatória**
Avatar: The Way of Water	17,05%	20%	16,66%
Elvis	14,35%	0%	26,10%
The Fabelmans	0%	0%	25,37%
TÁR	92,54%	80%	72,97%
Top Gun: Maverick	1,12%	0%	3,44%

Figura 19 – Matriz de confusão gerada a partir das previsões dos modelos na base de teste do Globo de Ouro (Comédia).



Fonte: Autor do trabalho.

Tabela 13 – Probabilidade de cada filme indicado ao Globo de Ouro de melhor comédia em 2023 ser vencedor de acordo com cada modelo.

Filme	**Naive bayes**	**KNN**	**Floresta Aleatória**
The Banshees of inisherin	96,44%	66,67%	18,09%
Babylon	0%	26,67%	35,76%
Everything Everywhere All at Once	97,57%	73,33%	48,46%
Glass Onion: A Knives Out Mystery	91,8%	53,33%	36,57%
Triangle of sadness	0%	33,33%	34,41%

3.5 Análise dos resultados

Após colher todos os dados, extrair as informações necessárias, avaliar e treinar os modelos, enfim é possível analisar os resultados obtidos durante o trabalho.

O primeiro ponto a ser analisado são as métricas obtidas em cada um dos algoritmos, através dos dados de teste segregados do *dataset* de treino, demonstradas nas tabelas 8, 9 e 10. O *naive bayes* foi o modelo que obteve os piores resultados de forma geral. Apesar de no Oscar, olhando para a matriz de confusão, acertar todos os filmes premiados, sua acurácia foi baixíssima, acertando apenas 48% das classes alvo. Sua métrica F1 também demonstra que nos três casos ele possivelmente é o pior algoritmo para 2 deles tendo como teto o *F1-score* de 36,36%. O algoritmo Floresta Aleatória por outro lado já teve resultados um pouco mais consistentes, sendo a segunda opção em quase todos os casos. Apesar de uma boa acurácia, sempre acima dos 70%, esse modelo teve uma certa dificuldade com as observações vencedoras do teste. É possível perceber que, com exceção do Oscar que teve uma precisão de 100%, esse algoritmo frequentemente previa objetos da classe negativa como se fossem da classe positiva, tendo uma baixa precisão nas indicações do Globo de Ouro. O algoritmo que se deu melhor durante todo o desenvolvimento foi o KNN. Tendo a maior métrica F entre os três, apesar de possuir uma revocação parecida com a da *random forest*, ele se mostrou muito mais preciso, com duas bases de dados com 100% de precisão e outra com 66%. Sua acurácia também se mantém acima dos 70% e pareceu ser o algoritmo mais interessante de acordo com os testes realizados.

Após o treinamento foi possível calcular as probabilidades de cada filme vencer a premiação, como visto nas tabelas 11, 12 e 13. Olhando para o Oscar, o ganhador da premiação seria *The Banshees of Inisherin*, de acordo com *naive bayes*, e *TÁR*, de acordo com os outros dois modelos. Para o ganhador do Globo de Ouro de melhor drama os resultados são iguais para todos os algoritmos, sendo *TÁR* o vencedor. Já considerando o Globo de Ouro de melhor comédia o escolhido seria *Everything Everywhere All at Once*, também por unanimidade.

Dado que esse trabalho foi desenvolvido após as premiações, é possível comparar os resultados obtidos com os premiados reais. Ao realizar essa comparação percebe-se que, na verdade, nenhum filme escolhido pelos algoritmos foi vencedor em suas respectivas premiações. No caso do Oscar, apesar de ficar entre terceiro e quarto de acordo com os treinamentos, o vencedor foi *Everything Everywhere All at Once*, ao invés de *The Banshees of Inisherin* ou *TÁR*. Para o caso do Globo de Ouro de melhor drama o grande vencedor foi *The Fabelmans*, um dos últimos colocados, com 0% de chance, de acordo com dois dos três algoritmos. Para o ultimo prêmio, Globo de Ouro de melhor comédia, o ganhador foi *The Banshees of Inisherin*. Nesse último caso, apesar dos modelos não terem apontado o filme como o ganhador, ele possuía uma boa colocação e estava em segundo lugar tanto para o *naive bayes* quanto para o KNN.

Observando as tabelas 1, 2 e 3 pode-se ter uma hipótese em relação a esse resultado. No caso do Oscar, o filme com maior média dos críticos de fato foi *TÁR*, com 8,76 pontos, seguido de *The Banshees of Inisherin* com 8,52. *Everything Everywhere All at Once* ficou apenas em quarto lugar de acordo com os especialistas. Considerando as notas dos usuários, o ganhador do Oscar de melhor filme teria sido o segundo lugar, com nota 8,11, enquanto o vencedor seria *Top Gun: Maverick* com 9,08. O filme *TÁR* apresentou uma avaliação de 5,73 na médias do público, enquanto *The Banshees of Inisherin* pontuou 6,38. No caso do Globo de Ouro de melhor drama a maior pontuação, considerando os especialístas, também fica com *TÁR*, tendo *The Fablemans* logo atrás em segundo lugar, com nota 8,14. Nessa categoria todas as médias de usuários estavam com valores baixos, com exceção de *Top Gun*, tendo *The Fablemans* uma nota de 6,25, sendo o quarto lugar nesse quesito. Por fim o prêmio de melhor comédia para o Globo de Ouro tem uma situação levemente diferente. O indicado com maior nota de críticos é de fato o ganhador, *The Banshees of Inisherin*, com a obra indicada pelos modelos ficando em segundo lugar. De acordo com o público a obra premiada teria apenas o quarto lugar.

4 Conclusão

Neste trabalho foram discutidas três grandes premiações do cinema mundial e como a opinião de especialistas e a opinião pública afeta seus resultados. Para coleta dos dados foi escolhido o Metacritic, um dos mais expressivos agregadores de notas disponíveis, de onde foi tirada as avaliações necessárias para o trabalho. A partir dos dados coletados, modelos de aprendizado de máquina foram treinados e avaliados, com a intenção de prever possíveis ganhadores apenas com seu desempenho nas notas anteriores à premiação. Após o treinamento e a obtenção das métricas necessárias, o resultado final foi analisado e comparado com o resultado real das cerimônias.

Ao fim da análise, nota-se que não foi possível criar uma relação direta entre o filme com maior avaliação da crítica e do público com os vencedores das cerimônias em todos os casos. No caso do melhor filme do Oscar e melhor drama do Globo de Ouro, apesar de não possuírem a maior avaliação da lista, suas notas eram próximas do vencedor tanto para críticos quanto para o público, indicando que, apesar da nota não indicar o escolhido, ele pode apontar quais são os indicados mais prováveis em cada cerimônia. O caso do prêmio de melhor comédia é divergente dos outros dois, já que o filme premiado possui de fato a melhor média de notas para os críticos, porém com uma baixa pontuação na média dos usuários, o que pode ser o principal motivo para ter sido o segundo lugar de acordo com os modelos de aprendizado.

Comparando os resultados finais e as premissas impostas do início desta monografia, os resultados surpreendem. Apesar de já ser esperado que o Oscar possui uma menor influência de avaliações externas por ser uma premiação fruto da academia, os resultados em relação ao Globo de Ouro não eram esperados. De forma geral, as premiações do Globo de Ouro foram mais sensíveis às notas do Metacritic, com os ganhadores tendo uma boa pontuação no ranking. Entretanto, o fato dos algoritmos de aprendizado não conseguirem prever esses ganhadores com facilidade revelou que esse prêmio é menos previsível do que se supunha inicialmente. Ao fim, conclui-se que, apesar de não atingir o objetivo de adivinhar o possível ganhador, foi possível entender que, considerando os resultados de 2023, as obras premiadas possuíam uma boa avaliação da crítica, permitindo reduzir a lista de prováveis ganhadores para um subconjunto menor, o que pode servir de grande auxílio para campanhas de marketing em redes sociais e plataformas de video sobre demanda nos dias e meses que antecedem as cerimônias.

Como sugestão para futuros trabalhos, seria interessante realizar uma comparação entre os possíveis diferentes agregadores de notas, entendo a eficácia de cada um, aplicando a metodologia apresentada, e comparando seus desempenhos em prever o resultado. Sendo

o Metacritic um agregador não só de avaliações de filmes mas também de jogos eletrônicos, um caminho muito interessante a se tomar seria testar essa mesma metodologia com o principal prêmio anual da indústria dos *video games*, o prêmio melhor jogo do ano presenteado na cerimônia do *The Game Awards*(AWARDS, 2023), e comparar a influência da opinião pública nessa mídia com o que foi visto em relação ao cinema.

Durante o desenvolvimento deste trabalho diversas disciplinas apresentadas durante curso foram de grande importância. Para desenvolver todo o código fonte foi necessário os conhecimentos de 'Programação procedimental', 'Programação orientada a objetos' e 'Estrutura de dados'. Para realizar as análises, desenvolver os atributos e gerar os modelos o conhecimento aprendido durante as disciplinas 'Estatística', 'Cálculo' e 'Mineração de dados' foram essenciais.

Referências

AFONSO, J. **Luz, câmera, ação e dinheiro**: As cifras bilionárias da indústria do cinema. 2021. Online. Disponível em: <https://blog.nubank.com.br/luz-camera-acao-e-dinheiro-as-cifras-bilionarias-da-industria-do-cinema/>. Acesso em: 13 jun. 2023. Citado na página 8.

AWARDS, T. G. **The game awards**. 2023. Disponível em: <https://thegameawards.com/>. Acesso em: 13 jun. 2023. Citado na página 46.

BEAUTIFULSSOUP. 2022. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Acesso em: 22 dez. 2022. Citado 3 vezes nas páginas 14, 15 e 16.

BORTOLINI, V. **Utilização de aprendizado de máquina para predição do diagnóstico do Covid-19**. 64 p. Monografia (Trabalho de Conclusão de Curso) — Instituto Federal de Santa Catarina, Caçador, maio 2022. Disponível em: <https://repositorio.ifsc.edu.br/handle/123456789/2504>. Acesso em: 11 jan. 2023. Citado na página 21.

CORRÊA, I. T. **Análise dos sentimentos expressos na rede social Twitter em relação aos filmes indicados ao Oscar 2017**. 72 p. Monografia (Trabalho de Conclusão de Curso) — Universidade Federal de Uberlândia, Uberlândia, dez. 2017. Disponível em: <https://repositorio.ufu.br/handle/123456789/20133>. Acesso em: 11 jan. 2023. Citado na página 21.

DOM. 2022. Disponível em: <https://developer.mozilla.org/pt-BR/docs/Web/API/Document_Object_Model>. Acesso em: 21 dez. 2022. Citado na página 14.

DUMELA. **Entenda o que é vídeo on demand e qual é a sua aplicação**. 2020. Online. Disponível em: <https://dumela.tv/blog/video-on-demand/>. Acesso em: 13 jun. 2023. Citado na página 8.

FOLLOWS, S. **How many films are released each year?** 2017. Online. Disponível em: <https://stephenfollows.com/how-many-films-are-released-each-year/>. Acesso em: 30 nov. 2022. Citado na página 8.

FONTES, R. **Golden Globe Awards, 1944 - 2020**. 2020. Disponível em: <https://www.kaggle.com/datasets/unanimad/golden-globe-awards>. Acesso em: 22 mai. 2023. Citado na página 26.

_____. **The Oscar Award, 1927 - 2023**. 2022. Disponível em: <https://www.kaggle.com/datasets/unanimad/the-oscar-award>. Acesso em: 22 mai. 2023. Citado na página 26.

GOLDENGLOBES.COM. **About the HFPA**. 2022. Disponível em: <https://www.goldenglobes.com/>. Acesso em: 29 nov. 2022. Citado na página 8.

HTML. 2022. Disponível em: <https://developer.mozilla.org/pt-BR/docs/Web/HTML>. Acesso em: 21 dez. 2022. Citado na página 13.

IBM. **What is Supervised Learning? | IBM**. 2023. Disponível em: <https://www.ibm.com/topics/supervised-learning>. Acesso em: 09 jan. 2023. Citado na página 17.

_____. **What is Unsupervised Learning? | IBM**. 2023. Disponível em: <https://www.ibm.com/topics/unsupervised-learning>. Acesso em: 09 jan. 2023. Citado na página 17.

IMBLEARN. **imblearn reference page**. 2023. Disponível em: <https://imbalanced-learn.org/stable/#>. Acesso em: 28 mai. 2023. Citado na página 37.

IMDB. **IMDB Developer API**. IMDb.com, 2021. Disponível em: <https://developer.imdb.com/documentation/api-documentation/getting-access/>. Acesso em: 13 jun. 2023. Citado na página 23.

_____. IMDb.com, 2022. Disponível em: <https://www.imdb.com/title/tt6710474/>. Acesso em: 13 jun. 2023. Citado na página 23.

IMDB. 2022. Disponível em: <https://www.imdb.com/>. Acesso em: 20 dez. 2022. Citado 2 vezes nas páginas 11 e 23.

KIM, J.; HWANG, S.; PARK, E. Can we predict the Oscar winner? A machine learning approach with social network services. **Entertainment Computing**, v. 39, p. 100441, ago. 2021. ISSN 1875-9521. DOI: <https://doi.org/10.1016/j.entcom.2021.100441>. Citado na página 21.

KINGSFORD, C.; SALZBERG, S. L. What are decision trees? **Nature Biotechnology**, Springer Nature, v. 26, p. 1011–1013, 2008. DOI: <https://doi.org/10.1038/nbt0908-1011>. Citado na página 19.

MAHESH, B. Machine learning algorithms: A review. **International Journal of Science and Research (IJSR)**, v. 9, p. 381–386, 2020. DOI: <https://doi.org/10.21275/ART20203995>. Citado na página 16.

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**, Science and Information (SAI) Organization Limited, v. 10, n. 2, 2019. DOI: <https://doi.org/10.14569/IJACSA.2019.0100257>. Citado na página 21.

METACRITIC. **Everything everywhere all at once**. 2022. Disponível em: <https://www.metacritic.com/movie/everything-everywhere-all-at-once>. Acesso em: 13 jun. 2023. Citado na página 25.

METACRITIC. 2022. Disponível em: <https://www.metacritic.com/about-metascores>. Acesso em: 20 dez. 2022. Citado 3 vezes nas páginas 11, 12 e 24.

NAQA, I. E.; MURPHY, M. J. What is machine learning? In: NAQA, I. E.; MURPHY, M. J. (Ed.). **Machine learning in radiation oncology**. [S.l.]: Springer, 2015. p. 3–11. DOI: <https://doi.org/10.1007/978-3-319-18305-3_1>. Citado na página 16.

OSCARS.ORG. **Oscar Statuette**. 2022. Online. Disponível em: <https://www.oscars.org/oscars/statuette/>. Acesso em: 29 nov. 2022. Citado na página 8.

PETERSON, L. E. K-nearest neighbor. **Scholarpedia**, v. 4, n. 2, p. 1883, 2009. Revision #137311. DOI: <https://doi.org/10.4249/scholarpedia.1883>. Citado na página 18.

RIGATTI, S. J. Random forest. **Journal of Insurance Medicine**, National Library of Medicine, v. 47, n. 1, p. 31–39, 2017. DOI: <DOI:10.17849/insm-47-01-31-39.1>. Citado na página 19.

ROTHMAN, L. How IMDb can be older than the first web browser. **Time**, Time USA, LLC, out. 2015. Disponível em: <https://time.com/4068036/imdb-history-25th-anniversary/>. Acesso em: 13 jun. 2023. Citado na página 23.

ROTTENTOMATOES. **Everything everywhere all at once**. 2022. Disponível em: <https://www.rottentomatoes.com/m/everything_everywhere_all_at_once>. Acesso em: 13 jun. 2023. Citado na página 24.

ROTTENTOMATOES. 2022. Disponível em: <https://www.rottentomatoes.com/about>. Acesso em: 20 dez. 2022. Citado 2 vezes nas páginas 11 e 24.

SCIKIT-LEARN. **Scikit-learn reference page**. 2023. Disponível em: <https://scikit-learn.org/stable/index.html>. Acesso em: 28 mai. 2023. Citado na página 37.

WEBB, G. I. Naïve bayes. In: SAMMUT, C.; WEBB, G. I. (Ed.). New York, NY: Springer, 2010. p. 713–714. DOI: <https://doi.org/10.1007/978-0-387-30164-8_576>. Citado na página 18.

ZHAO, B. Web scraping. In: SCHINTLER, L. A.; MCNEELY, C. L. (Ed.). **Encyclopedia of big data**. Cham: Springer Living, 2017. DOI: <https://doi.org/10.1007/978-3-319-32001-4_483-1>. Citado na página 14.