UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Guilherme Rodrigues Rodovalho

Análise da influência das avaliações de críticos e usuários nos resultados do The Game Awards

Uberlândia, Brasil

2025

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Guilherme Rodrigues Rodovalho

Análise da influência das avaliações de críticos e usuários nos resultados do The Game Awards

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

Orientador: Prof. Dr. Paulo Henrique Ribeiro Gabriel

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2025

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Faculdade de Computação
Av. João Naves de Ávila, 2121, Bloco 1B - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4393 - www.facom.ufu.br - facom@ufu.br

ATA DE DEFESA - GRADUAÇÃO

Curso de Graduação em:	Bacharelado em Sistemas de Informação				
Defesa de:	FACOM31802 – Trabalho de Conclusão de Curso 2				
Data:	03/08/2026	Hora de início:	17:05	Hora de encerramento:	18:10
Matrícula do Discente:	11911BSI226				
Nome do Discente:	Guilherme Rodrigues Rodovalho				
Título do Trabalho:	Análise da influência das avaliações de críticos e usuários nos resultados do The Game Awards				

A carga horária curricular foi cumprida integralmente?	(X) Sim ( ) Não

Reuniu-se no Anfiteatro/Sala virtual via plataforma MS Teams, da Universidade Federal de Uberlândia, a Banca Examinadora, designada pelo Colegiado do Curso de Graduação em Sistemas de Informação, assim composta: Professores: Dr. Bruno Augusto Nassif Travençolo - FACOM/UFU; Dr. Marcelo Zanchetta do Nascimento - 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 Nota 95 (Somente números inteiros)

OU

( ) Aprovado(a) sem nota.

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

Documento assinado eletronicamente por **Bruno Augusto Nassif Travençolo**, **Professor(a) do Magistério Superior**, em 03/08/2026, às 18:13, 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 03/08/2026, às 18:13, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Paulo Henrique Ribeiro Gabriel**, **Professor(a) do Magistério Superior**, em 03/08/2026, às 18:13, 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 **7555052** e o código CRC **A6A57911**.

Referência: Processo nº 23117.052868/2026-21 SEI nº 7555052

Agradecimentos

Agradeço primeiramente a Deus, por ter me dado a graça da vida e o desejo profundo de entender como as coisas funcionam para além do que é óbvio.

Agradeço à minha família pelas oportunidades; que me foram dadas com tanto sacrifício e que eu serei eternamente grato e levarei comigo o fardo de nunca ser capaz recompensá-los na mesma medida.

Agradeço à minha namorada, que nos momentos de dúvidas sempre me apoiou e voltou meu olhar ao que realmente importa, e nos momentos de alegria regozijou-se comigo pelas conquistas.

Agradeço ao meu professor orientador, que em aulas, para mim, foi um verdadeiro instigador da curiosidade, e na orientação deste trabalho foi paciente e compassivo com o aluno atípico que às vezes sou.

Agradeço aos meus professores, que me mostraram a beleza do trabalho intelectual, me ensinaram que o aprendizado é construído com esforço, e lições que guardarei para sempre no coração, mesmo quando tiver esquecido onde as aprendi.

Resumo

Este trabalho investiga a correlação entre as avaliações de críticos e usuários registradas no agregador Metacritic e a probabilidade de um jogo eletrônico ser premiado na categoria principal do *The Game Awards* (TGA). Para isso, adaptou-se ao contexto dos jogos eletrônicos uma metodologia previamente aplicada a premiações cinematográficas, na qual não se identificou relação clara entre as notas agregadas e os vencedores do Oscar e do Globo de Ouro. As avaliações individuais dos jogos indicados entre 2014 e 2025 foram coletadas por meio de *web scraping*, com um protocolo de filtragem temporal que descarta notas publicadas após cada cerimônia, prevenindo o vazamento de dados. A partir das notas consolidadas de todas as plataformas, foram extraídas doze *features* estatísticas e treinados três classificadores — Naive Bayes, K-Nearest Neighbors e Random Forest — com as edições de 2014 a 2023, testados em seguida nas edições de 2024 e 2025, em quatro cenários de combinação de fontes. O Random Forest atingiu 91,67% de acurácia no cenário de combinação de fontes e 100% no cenário ponderado 90/10, que replica os pesos atribuídos pelo TGA ao júri de críticos e à votação popular, com precisão e revocação perfeitas. Os resultados indicam uma correlação forte e mensurável entre as avaliações agregadas e o resultado da premiação, diferentemente do observado no cenário cinematográfico, e sugerem que a previsibilidade de uma premiação está positivamente correlacionada com a transparência de seus critérios de votação.

Palavras-chave: Aprendizado de máquina. Predição de premiações. The Game Awards. Metacritic. Web scraping.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AJAX	*Asynchronous JavaScript and XML*
API	*Application Programming Interface*
AutoML	*Automated Machine Learning*
AWS	*Amazon Web Services*
CSS	*Cascading Style Sheets*
CSV	*Comma-Separated Values*
DOM	*Document Object Model*
HTML	*HyperText Markup Language*
IMDb	*Internet Movie Database*
JSON	*JavaScript Object Notation*
KNN	*K-Nearest Neighbors*
NB	*Naive Bayes*
PC	*Personal Computer*
RF	*Random Forest*
ROS	*Random Over-Sampling*
SR	Sistemas de Recomendação
TGA	*The Game Awards*
URL	*Uniform Resource Locator*
WEKA	*Waikato Environment for Knowledge Analysis*
XML	*eXtensible Markup Language*

Sumário

1 Introdução

Desde a década de 1970, os jogos eletrônicos têm evoluído rapidamente, cativando milhões de pessoas ao redor do mundo devido à sua maneira interativa de contar histórias, à jogabilidade envolvente e à inovação artística proporcionada pela mídia (Lambert, 2008). Esse impacto cultural e a ampla acessibilidade dos videogames refletem-se também no seu crescimento econômico: em 2020, a indústria foi avaliada em 159,9 bilhões de dólares, superando a soma dos mercados de filmes e música (Morton, 2023).

Diante dessa relevância, surgiram premiações dedicadas a reconhecer a excelência no setor, sendo o *The Game Awards* (TGA) uma das mais notáveis. Criada em 2014 para eleger os melhores títulos do ano com base no voto do público e de críticos especializados (Awards, 2024), a cerimônia influencia diretamente a visibilidade de novos jogos e as tendências de consumo. Assim, o estudo de premiações culturais como essa mostra-se relevante não apenas para a análise da indústria, mas também para a compreensão de fenômenos midiáticos e do comportamento social em torno do entretenimento interativo.

Neste contexto, este trabalho tem como objetivo analisar a correlação entre as avaliações de usuários e críticos e a probabilidade de um jogo ser premiado no *The Game Awards*. Este trabalho visa empregar a mesma metodologia aplicada por Oliveira (2023) em um trabalho relacionado a premiações cinematográficas. O autor utilizou algoritmos de aprendizado de máquina, como Naive Bayes, Random Forest e K-Nearest Neighbors (KNN), empregando-os para analisar como a avaliação de críticos e usuários influencia as premiações do Oscar e do Globo de Ouro. Embora não tenha obtido resultados expressivos, Oliveira (2023) apontou que tal limitação pode ter ocorrido porque o Oscar tende a considerar fatores internos e corporativos mais do que as avaliações externas ao escolher o vencedor.

Devido aos critérios específicos e documentados do prêmio, espera-se que essa metodologia proporcione resultados mais expressivos neste contexto do que no cenário cinematográfico. Para tanto, foram coletados dados de um agregador de críticas e aplicados algoritmos de classificação para prever as chances de vitória com base nas pontuações. A escolha das técnicas de aprendizado de máquina justifica-se por sua capacidade de identificar padrões complexos em grandes volumes de dados. Além disso, essas técnicas podem gerar previsões usando características tanto quantitativas quanto qualitativas das críticas (Bishop; Nasrabadi, 2006).

A etapa de coleta de dados foi viabilizada por meio da técnica de *web scraping*, voltada à extração automatizada de informações de páginas da internet. A linguagem Python foi empregada para implementar esse processo. Para isso, utilizou-se o *Selenium*

WebDriver para navegar em páginas com carregamento dinâmico. Além disso, a biblioteca *BeautifulSoup* foi usada para analisar e extrair o conteúdo HTML resultante (Lowe, 2022). Por fim, na etapa de aprendizado de máquina, utilizou-se a biblioteca *scikit-learn*, que disponibiliza os algoritmos de classificação aplicados, permitindo a posterior avaliação e comparação dos modelos preditivos (Pedregosa *et al.*, 2011).

O restante desta monografia está organizado da seguinte maneira. No Capítulo 2, apresenta-se o referencial teórico, abordando sistemas de recomendação e de predição, agregadores de críticas, a técnica de *web scraping*, os fundamentos de aprendizado de máquina e os trabalhos relacionados. O Capítulo 3 descreve o desenvolvimento do trabalho, incluindo a seleção do agregador de críticas, a coleta e a filtragem temporal dos dados, a engenharia de *features* e a estratégia de treinamento e validação dos classificadores. O Capítulo 4 analisa e discute os resultados obtidos nos quatro cenários de dados avaliados, comparando-os com o trabalho de referência. Por fim, o último capítulo apresenta as conclusões e as perspectivas de trabalhos futuros.

2 Referencial Teórico

2.1 Sistemas de Recomendação e Sistemas de Predição

Sistemas de Recomendação (SR) são ferramentas e técnicas computacionais que têm o objetivo de sugerir itens de interesse a usuários. Assim, auxiliam na tomada de decisão em cenários com excesso de informações (Ricci; Rokach; Shapira, 2015). Tais sistemas tornaram-se ubíquos no cotidiano digital e são pilares estratégicos de plataformas líderes de mercado. Por exemplo, a Netflix usa esses sistemas para sugestão de conteúdo audiovisual, enquanto a Amazon os emprega para recomendações personalizadas de consumo. Já o Spotify utiliza esses sistemas para curadoria musical.

De modo geral, essas ferramentas são divididas em três abordagens principais. A primeira é a filtragem colaborativa, que se baseia na similaridade de comportamento entre usuários. A segunda é a filtragem baseada em conteúdo, que analisa as características intrínsecas dos itens. A terceira são os sistemas híbridos, que tentam combinar as duas estratégias para superar as limitações de cada uma delas (Aggarwal, 2016). No entanto, a implementação desses sistemas enfrenta desafios clássicos da área. Um deles é o problema de *cold start*, que se refere ao início a frio. Além disso, há a esparsidade das matrizes de dados e a necessidade de escalabilidade para lidar com volumes massivos de informações (Adomavicius; Tuzhilin, 2005).

Embora compartilhem bases tecnológicas, é importante distinguir os conceitos de recomendação e predição. Enquanto os SRs buscam sugerir itens de forma personalizada com base em perfis individuais, a predição foca na estimativa de valores específicos ou na classificação de instâncias em categorias predefinidas (Shani; Gunawardana, 2011).

Este trabalho enquadra-se no escopo da predição binária, cujo objetivo é classificar se um jogo nomeado ao *The Game Awards* (TGA) pertencerá à classe de vencedor (*winner*) ou não-vencedor (*loser*). Apesar da distinção funcional, observa-se uma convergência metodológica: ambos os campos utilizam avaliações (*ratings*) e análises (*reviews*) como as principais variáveis preditivas (*features*) para construir modelos estatísticos. Nesse sentido, técnicas consolidadas em sistemas de recomendação podem ser adaptadas para o refinamento de modelos preditivos em premiações culturais.

Nesse contexto, as avaliações atuam como um *proxy* quantitativo da qualidade percebida de um produto cultural, permitindo mensurar a recepção de uma obra perante diferentes públicos. Essas podem ser classificadas em *ratings* explícitos, quando o usuário atribui deliberadamente uma nota numérica ao item, ou implícitos, quando o interesse é inferido a partir de dados comportamentais (Jannach *et al.*, 2010). Contudo, a análise

de avaliações é suscetível a diversos vieses. Um deles é o viés de seleção, no qual usuários com opiniões extremas tendem a avaliar com maior frequência. Outro é o chamada "efeito manada" (*herding effect*), no qual notas prévias influenciam as percepções subsequentes (Muchnik; Aral; Taylor, 2013). Argumenta-se que agregar dados de múltiplas fontes independentes é uma estratégia robusta. Essa abordagem combina a *expertise* de críticos especializados com o sentimento da massa de usuários. Assim, ela ajuda a atenuar distorções e a obter uma representação mais fiel da qualidade artística e técnica de um título.

2.2 Agregadores de Críticas

Os agregadores de críticas se consolidam como sistemas fundamentais na era da informação digital. Eles atuam como repositórios centrais que reúnem análises de especialistas e opiniões de consumidores. Essas opiniões abrangem diversos produtos culturais, como filmes, música e, especialmente, jogos eletrônicos. Estes sistemas permitem que o público consulte de forma ágil e estruturada o consenso sobre uma obra, mitigando o problema da sobrecarga de informação e auxiliando no processo de decisão de consumo (Ricci; Rokach; Shapira, 2015).

Antes da Internet, a crítica cultural encontrava-se fragmentada em veículos de comunicação tradicionais, como jornais impressos, revistas especializadas e programas televisivos. Nesse cenário, o consumidor enfrentava dificuldades inerentes para acessar uma pluralidade de opiniões, dependendo muitas vezes de uma ou duas fontes de confiança para pautar suas escolhas. O poder dos críticos individuais era desproporcionalmente elevado, visto que não havia um mecanismo eficiente para contrastar diferentes perspectivas em larga escala.

O surgimento dos agregadores de críticas no final da década de 1990 e início dos anos 2000 alterou radicalmente esse paradigma. O Rotten Tomatoes, fundado em 1998, foi pioneiro no setor cinematográfico ao introduzir uma métrica binária de recepção crítica (Rotten Tomatoes, 2026). Posteriormente, o Metacritic, lançado em 2001, expandiu o conceito ao abranger múltiplas mídias e implementar um sistema de pontuação numérica normalizada (Metacritc, 2026). Essa democratização da crítica permitiu ao público visualizar o "sentimento da massa" de forma instantânea. No entanto, ela também trouxe um novo desafio: o excesso de informação e a potencial simplificação de análises complexas em números absolutos.

O impacto desses agregadores na indústria de jogos é profundo e multifacetado. Um dos casos mais emblemáticos é o do jogo *Fallout: New Vegas*, lançado em 2010. A desenvolvedora Obsidian Entertainment perdeu um bônus contratual porque o jogo atingiu 84 pontos no Metacritic. A cláusula de bônus exigia uma pontuação mínima de 85

pontos (Purchese, 2012). Tais práticas evidenciam que os escores agregados transcendem o mero guia de consumo, influenciando decisões de *marketing*, bônus corporativos e até a valorização de estúdios perante publicadoras e investidores. Contudo, a metodologia opaca de certas plataformas e a concentração excessiva de poder nesses índices permanecem como pontos de crítica recorrente no setor.

2.2.1 Metacritic

O Metacritic fundamenta sua operação em três componentes principais. O primeiro é o *Metascore*, que representa o consenso da crítica especializada. O segundo é o *User Score*, derivado de avaliações do público geral. O terceiro é a distribuição qualitativa das notas, que pode ser positiva, mista ou negativa. Para garantir a comparabilidade entre diferentes fontes, o sistema realiza uma normalização de escalas. Ele converte formatos diversos, como notas de 5 estrelas, conceitos de A a F ou pontuações de 0 a 10, para uma escala única de 0 a 100 (Oliveira, 2019). Adicionalmente, as avaliações são segmentadas por plataforma, reconhecendo que a experiência técnica de um jogo pode variar significativamente entre consoles e computadores.

Diferente de uma média aritmética simples, o cálculo do *Metascore* utiliza uma média ponderada. A plataforma atribui pesos distintos aos veículos de imprensa com base em sua relevância, reputação e consistência histórica na indústria. Embora a fórmula exata e os pesos individuais não sejam públicos, sabe-se que publicações de renome internacional possuem maior influência no índice final do que *blogs* independentes. Essa abordagem visa mitigar o impacto de análises destoantes ou de veículos com menor rigor editorial, buscando uma representação mais fiel da *expertise* crítica *mainstream*.

Em contrapartida, o *User Score* permite que qualquer usuário cadastrado atribua uma nota de 0 a 10 ao título. Embora esse indicador capte o sentimento popular, ele é frequentemente suscetível a fenômenos de *review bombing*. Nesse fenômeno, massas de usuários avaliam negativamente um jogo de forma coordenada, motivadas por controvérsias ideológicas ou problemas técnicos específicos. Para combater tais distorções, o Metacritic implementou medidas, como o bloqueio de avaliações de usuários nas primeiras 48 horas após o lançamento. Essa ação busca assegurar que o público tenha tido tempo hábil para consumir o produto antes de opinar.

A escolha do Metacritic como base de dados para este trabalho justifica-se por sua consolidação como o agregador mais influente e longevo da indústria. Sua ampla cobertura abrange desde títulos independentes até grandes produções AAA[1], mantendo um histórico estruturado que cobre todo o período de existência do *The Game Awards* (2014–2025).

[1] Jogos AAA, ou "triplo A", são aqueles lançados por empresas de renome e já consolidada no mercado. Geralmente, são jogos com alto custo de produção e contam com grande potencial de venda, o suficiente para trazer o retorno desse investimento e os lucros.

Além disso, o formato de dados do Metacritic, ao separar claramente críticos e usuários, fornece o insumo necessário para simular a metodologia de votação do TGA, que pondera essas duas fontes em proporções distintas.

2.2.2 Outros agregadores

Apesar da dominância do Metacritic, surgiram alternativas mais modernas que buscam sanar suas limitações. O OpenCritic, fundado em 2015, destaca-se por sua transparência, não aplicando ponderações ocultas às análises e permitindo que os usuários escolham quais críticos desejam seguir (OpenCritic, 2026). Embora o OpenCritic tenha ganhado popularidade por sua interface moderna e foco exclusivo em jogos, ele não foi utilizado como fonte primária neste estudo. Isso ocorreu devido ao seu histórico temporal mais limitado e à menor penetração em cláusulas contratuais da indústria, em comparação ao Metacritic.

Outro agregador relevante foi o GameRankings, que operou entre 1999 e 2019, antes de ser descontinuado e absorvido pela rede CBS Interactive. Durante duas décadas, o GameRankings foi a principal referência para o registro histórico de pontuações de jogos. Com o tempo, sua obsolescência tecnológica e a migração de tráfego para plataformas mais dinâmicas levaram ao seu encerramento. Hoje, ele é apenas uma referência histórica.

Por fim, é válido notar que a agregação de críticas manifesta-se de forma distinta em outras mídias. Enquanto o Rotten Tomatoes foca no percentual de aprovação (índice "Fresh/Rotten") em vez de uma média numérica, plataformas como o Goodreads (para livros) operam de forma quase inteiramente com base em usuários. Modelos que considerem diferentes aspectos são importantes devido à complexidade dos jogos eletrônicos.

2.3 Web Scraping

A maioria das páginas na internet é construída usando três tecnologias principais: HTML, CSS e JavaScript (Boyan; Freitag; Joachims, 1996). O HTML utiliza *tags* organizadas hierarquicamente, formando uma estrutura conhecida como *HTML Document Object Model* (DOM). A DOM representa a estrutura da página em uma árvore de dados, possibilitando a extração de informações por meio de técnicas de *web scraping*.

Web scraping, ou extração de dados de páginas web, é uma técnica amplamente utilizada para coletar e organizar informações publicamente disponíveis em sites. Consiste na análise da estrutura HTML da página, convertendo seu conteúdo em formatos mais legíveis e apropriados para análise ou uso em outros sistemas (Gunawan *et al.*, 2019).

Com o advento da World Wide Web em 1989, surgiram os primeiros programas de navegação e coleta de dados na internet, conhecidos como robôs ou *crawlers*. Em junho

de 1993, foi criado o primeiro robô com a finalidade de mensurar o tamanho da web. Em dezembro do mesmo ano, desenvolveu-se o primeiro motor de busca, utilizando robôs para navegar e indexar páginas web. Esse é um dos primeiros exemplos de *Web Crawler*, uma técnica fundamental para a evolução do *web scraping* (Wall, 2017).

Diferentes abordagens são empregadas dependendo da natureza técnica do sítio eletrônico. O *BeautifulSoup* é uma biblioteca amplamente utilizada para extração de dados de arquivos HTML e XML, funcionando de forma eficiente em páginas estáticas, cujo o conteúdo é fornecido diretamente pelo servidor (Lowe, 2022). Contudo, sítios modernos frequentemente utilizam JavaScript e técnicas de AJAX para carregar conteúdo de forma assíncrona (*lazy loading*), o que torna o conteúdo invisível para extratores estáticos. Nesses cenários, ferramentas de automação de navegador, como o *Selenium WebDriver*, tornam-se essenciais. Isso ocorre porque permitem a emulação completa de uma sessão de usuário, executando *scripts* no lado do cliente e interagindo com elementos dinâmicos antes da extração de dados (Selenium, 2024).

Além dos desafios técnicos, a prática do *web scraping* deve observar diretrizes éticas e legais. O arquivo `robots.txt`, localizado na raiz da maioria dos domínios, serve como um protocolo de exclusão que orienta robôs sobre quais áreas do sítio podem ou não ser acessadas. Respeitar as taxas de requisição para evitar a sobrecarga dos servidores (*throttling*) e priorizar a coleta de dados de domínio público são aspectos importantes para a execução responsável de pesquisas que dependem dessa técnica.

2.4 Aprendizado de Máquina

Aprendizado de máquina é um campo da inteligência artificial focado no estudo e desenvolvimento de algoritmos estatísticos que aprendem com os dados de entrada e generalizam para dados não vistos (Simon, 2013). Dentro do aprendizado de máquina, existem três tipos principais de tarefas: aprendizado supervisionado, não supervisionado e por reforço (Stuart J. Russell, 2009). O aprendizado de máquina tornou-se essencial na análise e interpretação de grandes volumes de dados, possibilitando avanços significativos em diversas áreas. Por exemplo, algoritmos de processamento de linguagem natural, como modelos de linguagem baseados em transformadores, aprimoram a compreensão e a geração de texto, viabilizando aplicações como tradução automática e assistentes virtuais (Chowdhary; Chowdhary, 2020). No setor financeiro, modelos de aprendizagem supervisionada auxiliam na detecção de fraudes em transações, identificando comportamentos anômalos em tempo real (Souza; Bordin Jr, 2023). Além disso, sistemas de recomendação utilizam técnicas de filtragem colaborativa, personalizando a experiência do usuário em plataformas de *streaming* e *e-commerce* ao sugerirem conteúdos e produtos alinhados às preferências individuais (Portugal; Alencar; Cowan, 2018).

O termo "aprendizado de máquina" foi cunhado por Samuel (1959), pioneiro na área de inteligência artificial e jogos computacionais. A história do aprendizado de máquina remonta aos anos 1940. Nessa época, Donald Hebb propôs uma teoria sobre interações neuronais, que fundamentou estruturas artificiais de redes neurais (Milner, 1993). Além disso, McCulloch e Pitts (1990) desenvolveram modelos matemáticos para simular processos cognitivos humanos. Na década de 1960, a Raytheon Company desenvolveu o *Cybertron*, uma máquina experimental de aprendizado. Essa máquina utilizava reforço para reconhecer padrões em sinais de sonar e eletrocardiogramas, amplificando o uso inicial de algoritmos de aprendizado com supervisão humana (Time, 1961). Mitchell (1997) formalizou a definição de aprendizado de máquina como algoritmos que aprimoram seu desempenho em tarefas específicas com a experiência. Essa definição está alinhada à pergunta de Alan Turing sobre a capacidade das máquinas de replicar processos cognitivos humanos (Mitchell, 1997).

As pesquisas mais recentes propõem diversas abordagens no campo do aprendizado de máquina. Por exemplo, existe o AutoML, ou aprendizado de máquina automático. Essa abordagem propõe um treinamento mais automatizado de modelos de aprendizado de máquina. Assim, quem está treinando não precisa ter um grande domínio do contexto que está sendo trabalhado (Elshawi; Maher; Sakr, 2019). Por outro lado, há o *Human-in-the-loop Machine Learning*, ou aprendizado com o humano no ciclo, que busca inserir o *feedback* humano diretamente no processo de aprendizado de máquina (Mosqueira-Rey *et al.*, 2023). Entre outras técnicas, destacam-se o aprendizado ativo, o aprendizado de máquina interativo, o ensino de máquina (Mosqueira-Rey *et al.*, 2023) e o meta-aprendizado (Elshawi; Maher; Sakr, 2019).

2.4.1 Algoritmos Empregados neste Trabalho

A escolha dos algoritmos para este trabalho baseia-se na diversidade de abordagens estatísticas, incluindo um modelo probabilístico (Naive Bayes), um modelo baseado em instâncias (KNN) e um modelo de conjunto (*ensemble*), baseado em árvores (Random Forest). Assim, é possível comparar desde *baselines* simples até modelos robustos de alto desempenho (Bishop; Nasrabadi, 2006).

Naive Bayes

O classificador Naive Bayes fundamenta-se no Teorema de Bayes, que descreve a probabilidade de um evento com base em conhecimentos prévios relacionados a ele. A característica "ingênua" (naive) do algoritmo reside na suposição de que todas as características (*features*) são condicionalmente independentes entre si, dado o rótulo da classe (Rish, 2001).

Matematicamente, para um vetor de características $X = (x_1, x_2, \ldots, x_n)$, a probabilidade de pertencer à classe C é dada por:

$$\Pr(C \mid X) = \frac{\Pr(C) \prod_{i=1}^{n} \Pr(x_i \mid C)}{\Pr(X)} \tag{2.1}$$

Embora a independência seja raramente observada em dados reais, o Naive Bayes apresenta resultados eficazes em problemas de classificação binária. Além disso, é computacionalmente eficiente e requer um volume relativamente pequeno de dados para treinamento (Mitchell, 1997).

K-Nearest Neighbors

O algoritmo KNN é um método de aprendizado não-paramétrico e baseado em instâncias. Diferente de outros algoritmos, o KNN não constrói um modelo explícito durante o treinamento. Em vez disso, ele armazena todos os exemplos de treino. Quando realiza uma classificação, ele baseia-se na similaridade de distância entre os exemplos (Cover; Hart, 1967).

A classificação de uma nova instância é definida pela maioria dos votos de seus k vizinhos mais próximos no espaço de características. A métrica de distância mais comum é a euclidiana, definida como:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.2}$$

A escolha do hiperparâmetro k é fundamental. Valores muito baixos podem causar *overfitting* por causa do ruído nos dados. Por outro lado, valores muito altos podem suavizar demais as fronteiras de decisão, ignorando padrões locais (Hastie; Tibshirani; Friedman, 2009).

Random Forest

O Random Forest (ou Floresta Aleatória) é um método de aprendizado em conjunto que opera construindo múltiplas árvores de decisão durante o treinamento. A predição final é obtida por meio da média das predições (regressão) ou da votação majoritária (classificação) de todas as árvores individuais (Breiman, 2001). Este algoritmo introduz duas formas de aleatoriedade para aumentar a robustez:a

1. **Bagging (Bootstrap Aggregating):** cada árvore é treinada em uma amostra aleatória do conjunto de dados, com reposição.
2. **Seleção aleatória de atributos:** em cada divisão (*split*) de nó, apenas um subconjunto aleatório de características é considerado, o que reduz a correlação entre as árvores e melhora a generalização (Fawagreh; Gaber; Elyan, 2014).

O Random Forest é especialmente eficaz para este estudo, pois consegue lidar com dados tabulares complexos e fornece uma métrica de importância de atributos, conhecida como *feature importance*. Essa métrica revela quais estatísticas do Metacritic têm maior influência na vitória no TGA

2.4.2 Tratamento de Dados e Validação

Um desafio inerente à predição de premiações é o severo desbalanceamento das classes: para cada vencedor (*winner*), existem múltiplos indicados que não recebem o prêmio (*losers*). Modelos treinados em conjuntos desbalanceados tendem a apresentar um viés em direção à classe majoritária, resultando em uma alta acurácia nominal, mas falhando em detectar a classe de interesse (He; Garcia, 2009). Para mitigar esse problema, utiliza-se a técnica de *Random Over-Sampling* (ROS), que consiste em replicar aleatoriamente instâncias da classe minoritária até que haja um equilíbrio entre as categorias no conjunto de treinamento. Essa abordagem garante que o algoritmo aprenda os padrões distintivos dos vencedores sem a perda de informação que ocorreria em técnicas de subamostragem (Lemaître; Nogueira; Aridas, 2017).

Além disso, em problemas em que os dados possuem uma ordem cronológica, a técnica tradicional de validação cruzada aleatória (k-fold) é inadequada como estratégia de avaliação final, pois pode causar o vazamento de dados (*data leakage*): o modelo "prevê o passado" usando informações do futuro (Hyndman; Athanasopoulos, 2018). O uso do k-fold permanece, contudo, válido de forma restrita ao conjunto de treinamento para a seleção de hiperparâmetros, uma vez que, nesse contexto, não há acesso a dados futuros.

Neste trabalho, adota-se a validação temporal, em que o modelo é treinado com dados de edições anteriores do TGA (ex.: 2014–2023) e testado exclusivamente em edições subsequentes (2024–2025). Esse método simula a aplicação real do sistema, cujo objetivo é prever o resultado da próxima cerimônia sem conhecimento prévio de seus vencedores.

2.4.3 Métricas de Avaliação

Para quantificar a eficácia dos modelos, além da acurácia, são utilizadas métricas derivadas da matriz de confusão, com foco no desempenho da classe positiva (vencedores) (Fawcett, 2006):

- Acurácia: proporção total de predições corretas.
- Precisão (*Precision*): proporção de vencedores reais entre todos os jogos que o modelo classificou como vencedores. É necessária para evitar "falsos positivos".
- Revocação (*Recall*): proporção de vencedores reais que o modelo foi capaz de identificar. Indica a sensibilidade do sistema.

- F1-Score: média harmônica entre precisão e revocação, fornecendo uma métrica única que balanceia ambos os aspectos, sendo adequada para classes desbalanceadas.

2.5 Trabalhos Relacionados

Esta seção apresenta estudos que aplicaram técnicas de aprendizado de máquina e mineração de opiniões à predição de premiações e do sucesso de produtos culturais em diferentes domínios, como cinema, jogos eletrônicos e literatura. Inicialmente, discute-se o trabalho que serviu de base para esta monografia; na sequência, os demais estudos são apresentados em ordem cronológica, destacando-se, para cada um, o objetivo, a metodologia empregada, os principais resultados e a relação com o presente trabalho.

O trabalho de Oliveira (2023) teve como objetivo investigar a relação entre a aceitação do público, com base nas notas fornecidas por críticos e usuários na plataforma Metacritic, e os vencedores das principais premiações cinematográficas, incluindo o Oscar de Melhor Filme, Globo de Ouro de Melhor Drama e Globo de Ouro de Melhor Comédia, no ano de 2023. Para tal, foram coletados dados de filmes indicados às premiações entre 2007 e 2023, os quais serviram como base para treinar modelos de aprendizado de máquina voltados à previsão dos vencedores do último ano. Os resultados da pesquisa indicaram que não foi possível estabelecer uma correlação clara entre as avaliações no Metacritic e os resultados das premiações. A abordagem proposta pelo autor serve como inspiração para o presente trabalho, que adapta essa metodologia ao contexto dos jogos eletrônicos, buscando analisar a correlação entre avaliações de críticos e usuários e a probabilidade de um jogo ser vencedor no *The Game Awards*.

O estudo pioneiro de Krauss *et al.* (2008) propôs uma abordagem de mineração da Web que combina análise de redes sociais e análise de sentimentos para prever eventos do mercado cinematográfico. Os autores analisaram as discussões do fórum *Oscar Buzz*, da plataforma *Internet Movie Database* (IMDb), entre novembro e dezembro de 2006, e derivaram índices de intensidade, positividade e ruído temporal para os 25 filmes mais comentados pela comunidade. O modelo resultante, baseado exclusivamente no índice de positividade das discussões, identificou corretamente nove dos dez filmes mais bem posicionados como vencedores ou indicados ao Oscar de 2007, dois meses antes da cerimônia. Em um segundo experimento, um modelo análogo, acrescido de um índice de formadores de tendência (*trendsetters*), apresentou correlação de 0,75 com a bilheteria dos vinte filmes de maior arrecadação de 2006. O trabalho é relevante para o presente estudo por evidenciar, de forma pioneira, que a opinião expressa por comunidades *on-line* antecipa resultados de premiações — premissa que esta monografia adota ao utilizar as avaliações de críticos e usuários do Metacritic como variáveis preditivas.

Na mesma direção, Correa *et al.* (2018) investigaram se o sentimento expresso

por usuários do Twitter seria capaz de antecipar os vencedores do Oscar de 2017. Os autores construíram uma base com 889.840 *tweets* em inglês sobre os nove filmes indicados à categoria de Melhor Filme, coletados entre o anúncio dos indicados e a véspera da cerimônia, e compararam três abordagens de classificação de sentimentos: Naive Bayes, aprendizado por supervisão distante (Sentiment140) e função de polaridade (TextBlob). O classificador Naive Bayes obteve o melhor desempenho, com acurácia de 74,1%, e foi utilizado para rotular a base completa. A comparação entre os *rankings* derivados dos *tweets* e um *ranking* ponderado de indicações e vitórias revelou apenas correlações fracas a moderadas (coeficiente de Spearman máximo de 0,43); ainda assim, a metodologia permitiu identificar corretamente o filme mais prestigiado e os menos prestigiados da edição analisada. O estudo aproxima-se do presente trabalho por buscar correlacionar a opinião do público com o resultado de premiações e por empregar o Naive Bayes, um dos classificadores aqui avaliados; difere, contudo, quanto à fonte de dados, uma vez que esta monografia substitui o texto espontâneo de redes sociais por notas estruturadas de um agregador de críticas.

No domínio dos jogos eletrônicos, Zhou (2022) empregou plataformas de aprendizado de máquina automatizado (AutoML) — EasyDL, Vertex AI e Azure — para analisar avaliações e vendas de jogos a partir de dados do Metacritic e do VGChartz. Foram treinados modelos de classificação, para categorizar os jogos como recomendados ou não pelos usuários, e modelos de regressão, para prever a nota dos usuários com base nas notas da crítica, em informações básicas dos títulos e nas vendas regionais. Os modelos de classificação alcançaram F1-Score superior a 70% nas plataformas EasyDL e Azure, enquanto os de regressão obtiveram coeficiente de determinação (R^2) em torno de 0,5. Em todos os cenários avaliados, a nota da crítica especializada figurou entre as variáveis mais importantes, o que levou o autor a concluir que há forte correlação entre a avaliação da mídia e a avaliação dos usuários, mas não entre essas avaliações e o desempenho comercial dos jogos. Esses achados são relevantes para o presente trabalho porque as notas de críticos e usuários do Metacritic — cuja inter-relação foi evidenciada pelo autor — constituem exatamente o conjunto de *features* aqui empregado para a predição dos vencedores do *The Game Awards*.

Por fim, Wijekoon e Rupasingha (2023) abordaram a predição de premiações no domínio literário, com o objetivo de identificar livros de sucesso a partir da conquista de prêmios. Utilizando uma base com 800 livros publicados entre 2000 e 2021, extraída da plataforma Goodreads por meio do repositório Kaggle, os autores modelaram o problema como uma classificação binária — livro premiado ou não premiado — a partir de dez atributos, entre eles nota média, contagem de avaliações, número de resenhas, autor, gênero e editora. Seis algoritmos foram comparados na ferramenta WEKA (Naive Bayes, *Random Forest*, árvore de decisão J48, *Multilayer Perceptron*, *Support Vector Machine* e Regressão Logística), tendo o Naive Bayes alcançado o melhor desempenho, com acurácia

de 86,32% e os menores índices de erro. O estudo guarda estreita relação metodológica com o presente trabalho: em ambos, a conquista de um prêmio é tratada como variável-alvo binária e predita a partir de atributos derivados de avaliações agregadas em plataformas *on-line*, com o emprego de classificadores em comum, como o Naive Bayes e o *Random Forest*.

Em síntese, os trabalhos revisados indicam que sinais de opinião disponíveis na Web — sejam discussões em fóruns, publicações em redes sociais ou notas agregadas por plataformas especializadas — carregam informação preditiva sobre premiações e sobre o sucesso de produtos culturais em domínios distintos. O presente trabalho diferencia-se dos demais ao aplicar essa premissa ao *The Game Awards*, utilizando exclusivamente variáveis estatísticas derivadas das notas de críticos e usuários do Metacritic, com filtragem temporal para prevenir o vazamento de dados, e ao comparar três classificadores em múltiplos cenários de validação.

3 Desenvolvimento

Neste capítulo, são abordadas todas as etapas do desenvolvimento desta monografia. A seção 3.1 descreve o processo de seleção da fonte de dados primária, comparando as alternativas disponíveis e justificando a escolha do Metacritic. A seção 3.2 detalha a implementação do módulo de extração automatizada de dados, incluindo a abordagem híbrida de coleta e o protocolo de filtragem temporal adotado para prevenir o vazamento de dados. A seção 3.3 apresenta a etapa de pré-processamento e a engenharia de *features*, descrevendo as doze variáveis estatísticas extraídas e a fórmula de ponderação que emula o critério oficial do prêmio. A seção 3.4 expõe a estratégia de treinamento e validação temporal, incluindo a configuração de cada classificador e o protocolo de balanceamento de classes. Por fim, a Capítulo 4 analisa e discute os resultados obtidos, comparando os modelos nos quatro cenários de dados avaliados. O código-fonte e as bases de dados utilizados estão disponíveis em `https://github.com/GuilhermRodovalho/award-prediction`.

3.1 Agregador de Críticas

A definição da fonte primária de dados é uma decisão metodológica central. Ela deve atender a dois requisitos simultâneos: primeiro, separar explicitamente as avaliações da crítica especializada das opiniões do público geral, para permitir a emulação da regra de votação 90/10 do *The Game Awards* (TGA). Segundo, essa fonte deve possuir cobertura histórica que abranja integralmente o período de existência da cerimônia, de 2014 a 2025. Para fundamentar a escolha, foram consideradas três plataformas de forma ampla utilizadas pela indústria: o IMDb, o OpenCritic e o Metacritic.

O *Internet Movie Database* (IMDb) é um dos maiores repositórios de dados culturais digitais e oferece uma API[1] pública por meio dos serviços da *Amazon Web Services* (AWS). No entanto, não foi adotado como fonte primária devido a uma limitação estrutural: a plataforma não diferencia formalmente as avaliações de críticos credenciados das avaliações do público geral para jogos eletrônicos. Sua métrica principal é uma nota média calculada sobre votos de usuários cadastrados, sem a segmentação especializada necessária para emular a ponderação assimétrica do TGA. Utilizar o IMDb implicaria tratar toda a base de avaliações como homogênea, o que comprometeria a premissa metodológica do presente estudo.

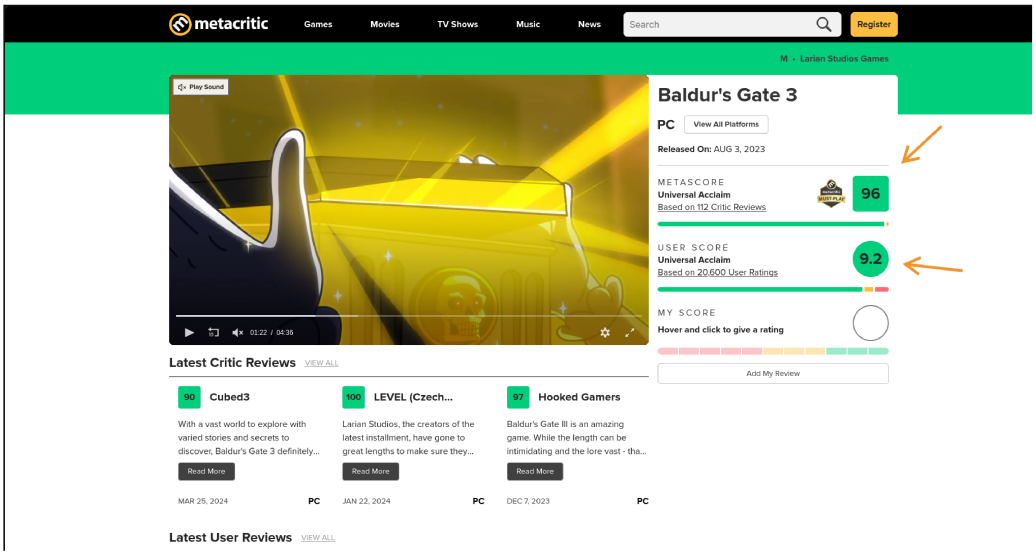
O OpenCritic foi lançado em 2015 e é uma alternativa moderna e transparente. Diferentemente do Metacritic, ele não aplica ponderações ocultas às análises dos veículos,

[1] *Application Programming Interface*

o que torna seu sistema de agregação mais rastreável academicamente (OpenCritic, 2026). Seu foco exclusivo em jogos eletrônicos e sua interface estruturada são pontos positivos inegáveis. Contudo, o início de sua operação em 2015 representa um obstáculo intransponível para este trabalho. Isso ocorre porque a primeira edição do The Game Awards foi realizada em dezembro de 2014. Excluir essa cerimônia do conjunto de treinamento reduziria desnecessariamente o já escasso histórico disponível. Além disso, o OpenCritic possui menor penetração em cláusulas contratuais da indústria, o que indica menor influência institucional comparada ao Metacritic.

O Metacritic, lançado em 2001, consolidou-se como o agregador de referência para a indústria de jogos eletrônicos (Metacritc, 2026). A adoção desta fonte como principal neste trabalho é justificada por três critérios objetivos. Primeiro, ela oferece cobertura histórica integral do período de 2014 a 2025, sem lacunas. Segundo, há uma separação explícita entre dois índices independentes: o *Metascore*, que representa o consenso da crítica especializada em uma escala de 0 a 100, e o *User Score*, que reflete a opinião do público geral em uma escala de 0 a 10, como ilustrado na Figura 1. Essa estrutura permite a emulação direta da regra 90/10 do TGA. Terceiro, ela possui relevância institucional consolidada, evidenciada pelo uso do Metascore como métrica de desempenho em contratos comerciais entre publicadoras e desenvolvedoras, conforme discutido na subseção subseção 2.2.1 (Purchese, 2012).

Figura 1 – Exemplo de página de jogo no *Metacritic* exibindo simultaneamente o *Metascore* (crítica especializada, escala 0–100) e o *User Score* (público geral, escala 0–10).



Fonte: Extraído de Metacritc (2026).

Um aspecto relevante para a modelagem é a diferença de escala entre os dois índices do Metacritic: o *Metascore* varia de 0 a 100, enquanto o *User Score* varia de 0 a 10.

Para garantir a compatibilidade entre as fontes, especialmente nos algoritmos baseados em distância, como o KNN, e na fórmula de ponderação 90/10, as notas individuais da crítica especializada foram normalizadas. Essa normalização foi feita dividindo-se as notas por 10, ajustando-as para a escala decimal $[0, 10]$. Após essa transformação, ambas as fontes operam na mesma grandeza, permitindo que a variável ponderada V_{final} seja calculada de forma consistente, conforme detalhado na seção 3.3.

3.2 Web Scraping e Extração de Dados

Para a coleta sistemática das avaliações, desenvolveu-se um módulo de *web scraping* em Python, implementado pela classe `MetacriticScraper`. Esta seção descreve as cinco etapas principais desse processo. A primeira é a geração dos endereços de acesso às páginas dos jogos, detalhada na subseção 3.2.1. A segunda é a extração híbrida do conteúdo dinâmico, abordada na subseção 3.2.2. A terceira envolve o protocolo de filtragem temporal aplicado para prevenir o vazamento de dados, na subseção 3.2.3. A quarta é a consolidação das avaliações entre plataformas, na subseção subseção 3.2.4. Por fim, a quinta etapa trata da estrutura dos dados resultantes, apresentada na subseção 3.2.5.

3.2.1 Geração de *slugs* e acesso às páginas

Cada jogo no Metacritic possui uma página cujo endereço segue o padrão `https://www.metacritic.com/game/[slug]/`, em que o *slug* é uma versão normalizada do título: letras minúsculas, espaços e caracteres especiais substituídos por hífens e acentos removidos. Por exemplo, o jogo *The Legend of Zelda: Breath of the Wild* é acessado pelo endereço `/game/the-legend-of-zelda-breath-of-the-wild/`. Para automatizar essa derivação, implementou-se a função `slugify_title()`. Essa função tem um comportamento ilustrado na Figura 2.

Como títulos homônimos em anos diferentes podem compartilhar o mesmo *slug* base, a função `process_game_reviews()` implementa um mecanismo de *fallback*: caso a URL primária retorne erro, o sistema tenta automaticamente uma variação com o ano de lançamento concatenado ao *slug* (ex: `doom-2016`). As variações são testadas em sequência até que avaliações sejam encontradas ou todas as possibilidades se esgotem.

3.2.2 Extração Híbrida: Selenium e BeautifulSoup

O Metacritic emprega carregamento dinâmico de conteúdo (*lazy loading*): as listas de avaliações são renderizadas via JavaScript no lado do cliente, tornando-as invisíveis a ferramentas de análise HTML estática. Para contornar essa limitação, adotou-se uma abordagem híbrida em duas fases.

Figura 2 – Implementação da função `slugify_title()`, responsável por converter o título de um jogo no identificador de URL utilizado pelo *Metacritic*.

```python
def slugify_title(title: str) -> str:
    slug = title.lower()
    slug = slug.replace("’", "")
    slug = slug.replace(":", "")
    slug = slug.replace("&", "and")
    slug = slug.replace("\u0113", "e")  # e com macron
    slug = slug.replace("\u00e9", "e")  # e com acento
    slug = slug.replace("\u00f6", "o")  # o com trema
    slug = re.sub(r"[^a-z0-9\s-]", "", slug)  # remove especiais
    slug = re.sub(r"\s+", "-", slug)          # espacos -> hifens
    slug = re.sub(r"-+", "-", slug)           # hifens duplos
    slug = slug.strip("-")
    return slug
```

Fonte: Elaborado pelo autor.

Na primeira fase, o *Selenium WebDriver* opera em modo *headless* (sem interface gráfica), emulando um navegador real. O *driver* navega até a página do jogo. Em seguida, executa os *scripts* JavaScript que expõem as listas de avaliações. Depois, realiza uma rolagem vertical (*scroll*) de forma iterativa na página. Durante esse processo, aciona o elemento `div.load-trigger`, que dispara o *Intersection Observer* responsável pelo carregamento incremental de conteúdo. O processo encerra-se automaticamente quando dez tentativas consecutivas de rolagem não produzem novas avaliações no DOM, ou quando o limite máximo de iterações é atingido. Na segunda fase, o HTML estabilizado é transferido para a biblioteca *BeautifulSoup*, que percorre a árvore DOM em busca dos elementos de avaliação identificados pelo atributo `data-testid="review-card"`. Para cada elemento, o *parser* extrai a nota numérica do componente `c-siteReviewScore` e a data de publicação do componente `review-card__date`, conforme ilustrado na Figura 3.

3.2.3 Protocolo de Filtragem Temporal

Um risco inerente a esse tipo de problema é o vazamento de dados, conhecido como *data leakage*, causado pela contaminação temporal. Premiações de grande visibilidade podem gerar picos de avaliações retrospectivas nas semanas seguintes à cerimônia. Essas avaliações são publicadas por usuários que consumiram o jogo motivados pelo resultado da premiação. Incorporar essas avaliações ao conjunto de treinamento equivale a treinar o modelo com informações que não estariam disponíveis no momento real da predição.

Para eliminar esse viés, implementou-se um protocolo de filtragem temporal no nível de cada avaliação individual. Durante a extração, a data de publicação de cada *review* é comparada à data oficial da cerimônia correspondente, ambas armazenadas no formato `DD/MM/AAAA` no arquivo CSV de entrada. Avaliações com data posterior à cerimônia são

descartadas antes mesmo de serem armazenadas, como demonstrado na Figura 3.

Figura 3 – Fragmento da função `_extract_reviews_from_soup()` responsável pela filtragem temporal. A data de cada avaliação é comparada à data da cerimônia; registros posteriores ao evento são descartados.

```python
def _extract_reviews_from_soup(self, soup, cerimony_date, critic):
    reviews = []
    cerimony_date_obj = datetime.strptime(cerimony_date, "%d/%m/%Y").
       date()

    review_elements = soup.find_all(
        "div", attrs={"data-testid": "review-card"})

    for review_div in review_elements:
        score_element = review_div.find(
            "div", class_=re.compile(r"\bc-siteReviewScore\b"))
        if not score_element:
            continue
        score_span = score_element.find("span")
        if not score_span:
            continue
        score = score_span.text.strip()

        date_el = review_div.find("div", class_="review-card__date")
        if date_el:
            try:
                review_date = datetime.strptime(
                    date_el.text.strip(), "%b %d, %Y").date()
                if review_date > cerimony_date_obj:
                    continue  # descarta review posterior a cerimonia
            except ValueError:
                pass

        try:
            int(score)
            reviews.append(score)
        except ValueError:
            pass
    return reviews
```

Fonte: Elaborado pelo autor.

Um exemplo da relevância desse protocolo é o caso do jogo *Ghost of Tsushima*, indicado ao TGA 2020 (cerimônia realizada em 10/12/2020). O título, lançado em julho de 2020, contava com 8.459 avaliações de usuários no Metacritic publicadas entre seu lançamento e a data da cerimônia, o maior volume do conjunto de dados de treinamento, reflexo de cinco meses de exposição intensa antes do evento. Sem o protocolo de filtragem, esse volume seria inflado por avaliações retrospectivas publicadas após a cerimônia, motivadas pela indicação e pela cobertura midiática do prêmio. O mesmo padrão foi observado em outros títulos de longa exposição, como *The Last of Us Part II* (3.141 avaliações filtradas), também indicado ao TGA 2020, e *Elden Ring* (7.806 avaliações), indicado ao TGA

2022.

3.2.4 Consolidação Multiplataforma

O Metacritic segmenta as avaliações por plataforma de execução, reconhecendo que a experiência técnica de um mesmo jogo pode variar entre consoles e computadores. Contudo, o *The Game Awards* avalia a obra como um todo, independentemente da plataforma em que o júri a experienciou. Para alinhar a fonte de dados ao critério da premiação, o módulo de coleta detecta automaticamente todas as plataformas disponíveis para um título, interagindo com o componente *dropdown* personalizado de filtro de plataformas, que é identificado pelo botão com atributo `button[aria-haspopup]` no DOM. Em seguida, realiza a extração de avaliações de cada plataforma e consolida os resultados em uma única lista.

Um exemplo concreto é o jogo *Elden Ring*, vencedor do TGA 2022, que estava disponível em cinco plataformas: PlayStation 4, PlayStation 5, Xbox One, Xbox Series X/S e PC (Windows). O *scraper* coletou avaliações de críticos de todas as cinco versões, totalizando 166 notas individuais de críticos e 7.806 avaliações de usuários após a aplicação do filtro temporal. Calculadas sobre esse conjunto unificado, as estatísticas refletem a recepção global do jogo, indicando que a abordagem é coerente com o critério multiplataforma do TGA.

3.2.5 Estrutura dos Dados Coletados

O resultado da etapa de extração é armazenado em arquivos no formato JSON[2]. Cada arquivo tem uma chave que corresponde ao nome do jogo, cujo valor é um objeto com vários campos. O campo `critic-reviews` é uma lista de notas individuais da crítica especializada, já filtradas temporalmente, na escala de 0 a 100. O campo `user-reviews` é uma lista de notas de usuários, também filtradas, na escala de 0 a 10. O campo `year` indica o ano de lançamento do jogo. O campo `cerimony-date` registra a data da cerimônia no formato DD/MM/AAAA. Por fim, o campo `winner` informa se o jogo venceu aquela edição. A Tabela 1 apresenta uma amostra representativa do conjunto de dados resultante.

O conjunto histórico utilizado no treinamento abrange as edições de 2014 a 2023. Após a aplicação do *Random Over-Sampling*, o conjunto de treinamento totaliza 92 instâncias balanceadas (46 vencedores e 46 não vencedores). O conjunto de teste, composto pelas edições de 2024 e 2025, contém 12 jogos, sendo 2 vencedores e 10 não vencedores.

[2] *JavaScript Object Notation*

Tabela 1 – Amostra do *dataset* coletado, com quantidade de avaliações por fonte após filtragem temporal.

Jogo	**Ano TGA**	**Classe**	**Críticos**	**Usuários**
Dragon Age: Inquisition	2014	Vencedor	116	902
Bayonetta 2	2014	Não-vencedor	111	100
The Witcher 3: Wild Hunt	2015	Vencedor	124	1653
Elden Ring	2022	Vencedor	166	7806
Baldur's Gate 3	2023	Vencedor	158	5929
Resident Evil 4	2023	Não-vencedor	175	733
Super Mario Bros. Wonder	2023	Não-vencedor	128	361
Astro Bot	2024	Vencedor	136	1346

Fonte: Elaborado pelo autor.

3.3 Pré-processamento e Engenharia de Features

A etapa de pré-processamento visou converter os dados brutos em vetores de características estruturadas, otimizados para algoritmos de classificação supervisionada.

3.3.1 Extração de variáveis estatísticas

Diferentemente de modelos simplistas baseados unicamente na média aritmética, este trabalho emprega doze estatísticas descritivas para cada título. O objetivo é capturar a distribuição, o consenso e a dispersão das notas, fornecendo ao modelo uma compreensão mais profunda da recepção do jogo. A Tabela 2 apresenta as variáveis calculadas para ambos os grupos (críticos e usuários).

Tabela 2 – Definição das *features* estatísticas extraídas.

Variável	**Definição Estatística**	**Aplicação**
Média	Tendência central aritmética	Críticos e Usuários
Mediana	Ponto central da amostra	Críticos e Usuários
Moda	Valor de maior frequência relativa	Críticos e Usuários
Desvio Padrão	Medida de dispersão absoluta	Críticos e Usuários
Percentil 25	Primeiro quartil da distribuição	Críticos e Usuários
Percentil 75	Terceiro quartil da distribuição	Críticos e Usuários

Fonte: Elaborado pelo autor.

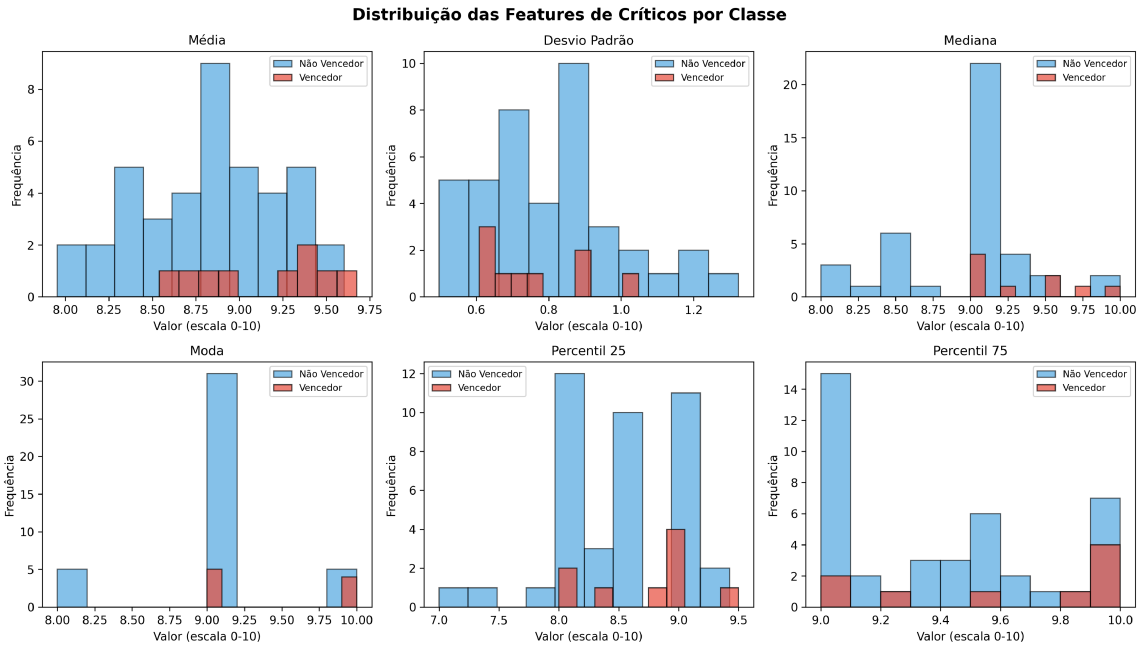
A escolha dessas seis métricas estatísticas, aplicadas a cada grupo de avaliadores, foi motivada pela complementaridade das informações que capturam. A média e a mediana representam a tendência central da recepção; porém, a mediana oferece maior robustez a *outliers* – como avaliações extremas decorrentes de *review bombing*. A moda revela a

nota mais com frequência atribuída, sendo útil para identificar consenso: quando muitos críticos atribuem a mesma nota elevada, há um sinal forte de aclamação uniforme. O desvio padrão quantifica a dispersão das opiniões, diferenciando jogos controversos (desvio alto) de jogos unanimemente aclamados (desvio baixo). Por fim, os percentis 25 e 75 delimitam o intervalo interquartil, revelando o "piso" e o "teto" das avaliações sem distorção pelos extremos da distribuição.

3.3.2 Análise exploratória por classe

A análise exploratória das *features* extraídas revela padrões distintos entre vencedores e não vencedores no conjunto de treinamento (2014–2023). Considerando as avaliações de críticos, os jogos vencedores apresentam uma média de 9,20 na escala normalizada de 0 a 10. Por outro lado, os jogos não vencedores têm uma média de 8,83. Os não vencedores exibem uma distribuição mais ampla de notas, com um desvio padrão de 0,44, enquanto os vencedores têm um desvio padrão de 0,37. O desvio padrão intra-jogo dos vencedores é consistentemente menor (média de 0,72 contra 0,82 dos não vencedores), indicando maior consenso crítico em torno dos títulos premiados. O percentil 25 dos vencedores (média de 8,83) supera o dos não vencedores (8,44), demonstrando que mesmo as avaliações mais baixas dos vencedores permanecem em patamar elevado. A Figura 4 ilustra essas diferenças.

Figura 4 – Distribuição das *features* de críticos para vencedores e não vencedores (2014–2023).

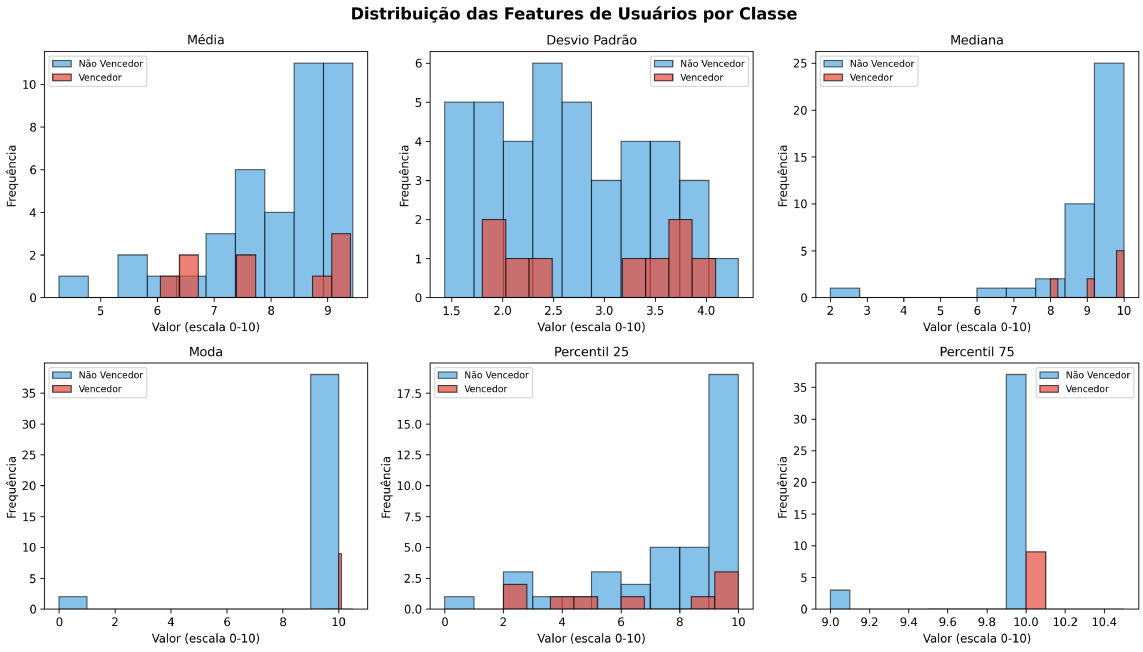


Fonte: Elaborado pelo autor.

Já para as avaliações de usuários, a separação entre as classes é consideravelmente

menos pronunciada. A média dos vencedores (8,21) é praticamente idêntica à dos não vencedores (8,14), e o desvio padrão intra-jogo é semelhante em ambos os grupos (2,67 e 2,70, respectivamente). A amplitude das médias dos vencedores varia de 6,05 a 9,69, enquanto a dos não vencedores vai de 4,27 a 9,44. Essa sobreposição corrobora a hipótese de que o voto do público, que representa apenas 10% do peso decisório no TGA, possui menor poder discriminativo univariado para identificar os vencedores. Contudo, é importante ressaltar que a ausência de separabilidade univariada não elimina o poder preditivo em um contexto multivariado. Essas variáveis, quando combinadas entre si e com as *features* de críticos, podem capturar padrões latentes que não são visíveis na análise isolada. Isso será demonstrado nos resultados da Capítulo 4. A Figura 5 apresenta as distribuições das *features* de usuários.

Figura 5 – Distribuição das *features* de usuários para vencedores e não vencedores (2014–2023).



Fonte: Elaborado pelo autor.

3.3.3 Ponderação 90/10: Emulação do Método TGA

A variável de maior relevância metodológica é a simulação da regra de votação oficial do prêmio, na qual o conselho de críticos detém 90% do peso decisório e a votação popular contribui com os 10% remanescentes. Formalizou-se essa regra por meio da variável ponderada V_{final}, definida pela Equação 3.1:

$$V_{final} = (M_{crit} \times 0{,}9) + (M_{user} \times 0{,}1) \tag{3.1}$$

Nessa formulação, as médias são previamente normalizadas para a mesma escala decimal, permitindo que o modelo incorpore a hierarquia de influência real do evento em seu processo de aprendizado.

3.3.4 Tratamento de Dados e Normalização

Para jogos com amostras insuficientes de avaliações de usuários — como o caso de *PlayerUnknown's Battlegrounds*, único título do *dataset* sem avaliações de usuários coletadas — aplicou-se a imputação estatística por média categórica anual. Nesta abordagem, as *features* ausentes são substituídas pela média das avaliações de usuários de todos os jogos do mesmo ano de cerimônia, preservando a dimensionalidade necessária para o treinamento sem introduzir viés temporal.

Adicionalmente, empregou-se a normalização *Min-Max* nas variáveis contínuas destinadas ao algoritmo KNN, que é sensível à escala dos atributos por se basear na distância euclidiana. Os algoritmos *Naive Bayes* e *Random Forest*, por não dependerem de métricas de distância, receberam os dados sem normalização adicional.

3.4 Treinamento e estratégia de validação

Esta seção detalha as decisões adotadas na fase de treinamento dos classificadores: a estratégia de divisão temporal dos dados, o protocolo de balanceamento das classes e a configuração dos hiperparâmetros de cada modelo.

3.4.1 Divisão por janelas temporais

Optou-se por uma estratégia de validação temporal em detrimento do k-fold aleatório. O conjunto de treinamento é composto por jogos indicados entre 2014 e 2023, balanceado via *Random Over-Sampling* para 92 instâncias (46 vencedores e 46 não vencedores). O conjunto de teste contém 12 jogos das edições de 2024 e 2025, sendo 2 vencedores e 10 não vencedores. A Tabela 3 apresenta a composição completa do conjunto de teste. Essa abordagem simula o desafio real de predição, em que o modelo deve generalizar padrões aprendidos em edições passadas sem contaminação temporal.

3.4.2 Balanceamento via *over-sampling*

O desbalanceamento original do conjunto de treinamento, 10 vencedores contra 46 não vencedores, tende a induzir os classificadores a favorecer sistematicamente a classe majoritária. Para mitigar esse viés, aplicou-se o algoritmo *Random over-sampler* (Lemaître; Nogueira; Aridas, 2017), que replica aleatoriamente instâncias da classe minoritária até equalizar as duas categorias.

Tabela 3 – Jogos utilizados como conjunto de teste (edições de 2024 e 2025).

Jogo	**Edição TGA**	**Classe Real**
Astro Bot	2024	Vencedor
Balatro	2024	Não Vencedor
Black Myth: Wukong	2024	Não Vencedor
Elden Ring: Shadow of the Erdtree	2024	Não Vencedor
Final Fantasy VII Rebirth	2024	Não Vencedor
Metaphor: ReFantazio	2024	Não Vencedor
Clair Obscur: Expedition 33	2025	Vencedor
Death Stranding 2: On the Beach	2025	Não Vencedor
Donkey Kong Bananza	2025	Não Vencedor
Hades 2	2025	Não Vencedor
Hollow Knight: Silksong	2025	Não Vencedor
Kingdom Come: Deliverance 2	2025	Não Vencedor

Fonte: Elaborado pelo autor.

Após o balanceamento, o conjunto de treinamento passou a contar com 46 vencedores e 46 não vencedores, totalizando 92 instâncias. A replicação é aplicada exclusivamente ao conjunto de treinamento; o conjunto de teste permanece inalterado para preservar a validade da avaliação.

3.4.3 Configuração dos classificadores

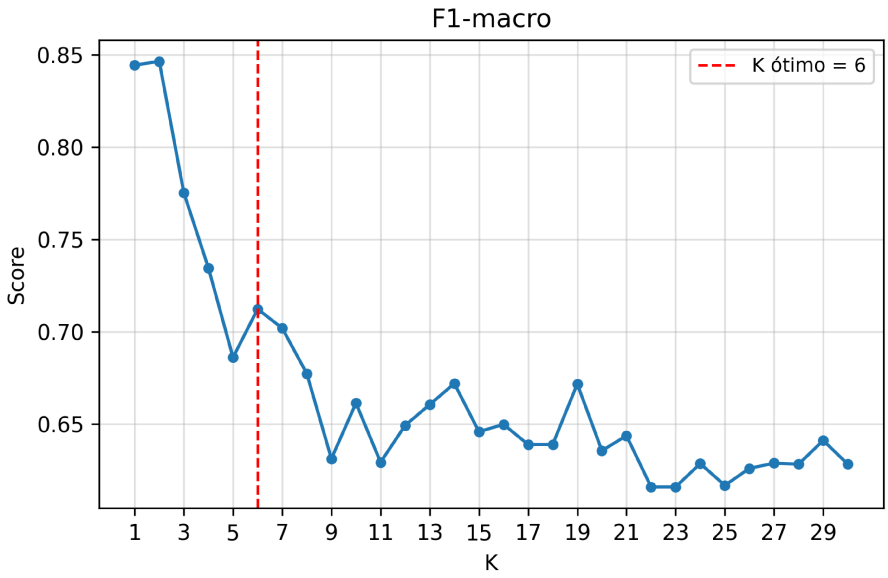
Foram avaliados três algoritmos de classificação, selecionados pela diversidade de abordagens que representam, e disponíveis na biblioteca *scikit-learn* (Pedregosa *et al.*, 2011):

- ***Gaussian Naive Bayes***: modelo probabilístico que assume distribuição gaussiana das features e independência condicional entre elas. Não há hiperparâmetros que exijam ajuste, tornando-o um baseline natural para comparação.
- ***K-Nearest Neighbors* (KNN)**: modelo baseado em instâncias, com classificação por votação dos K vizinhos mais próximos. Para determinar o valor ótimo de K, o algoritmo foi executado para valores de 1 a 30, considerando o cenário de combinação de fontes. Durante esse processo, foram avaliadas três métricas usando validação cruzada de 5 *folds* sobre o conjunto de treinamento balanceado. Essas métricas foram o *F1-macro*, que é a média ponderada entre as classes; o *F1-Winner*, que corresponde ao *F1-Score* da classe positiva; e o *Recall-Winner*, que é a revocação da classe positiva. As Figuras 6, 7 e 8 apresentam os resultados para cada métrica. Embora $K = 1$ e $K = 2$ atinjam os maiores valores em todas as métricas, valores muito baixos de K tendem a ser sensíveis a ruído e propensos a *overfitting*, de forma

particular em conjuntos de dados pequenos, como o deste trabalho. Optou-se por $K = 6$, primeiro pico local acima de $K = 5$ no *F1-macro* (0,712), que representa o melhor equilíbrio entre sensibilidade local e estabilidade de generalização antes da queda contínua observada a partir de $K = 7$. A normalização *Min-Max* foi aplicada previamente às *features* para garantir que todas operem na mesma escala.

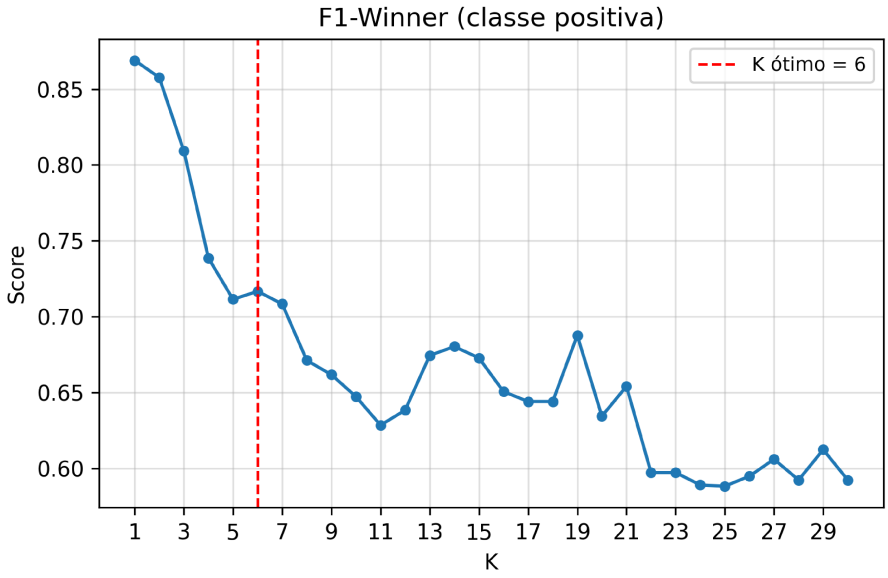
- ***Random Forest***: Modelo de conjunto (*ensemble*) baseado em múltiplas árvores de decisão. A configuração adotada utiliza o critério de entropia para divisão dos nós, restrição de *min_samples_leaf*=3 para limitar a profundidade efetiva das árvores e prevenir o *overfitting*, e *random_state*=80 para reprodutibilidade dos resultados. O número de estimadores foi mantido no valor padrão da biblioteca *scikit-learn* ($n\_estimators = 100$).

Figura 6 – F1-macro em validação cruzada (5 *folds*) para diferentes valores de K no KNN. A linha tracejada indica $K = 6$, valor selecionado.



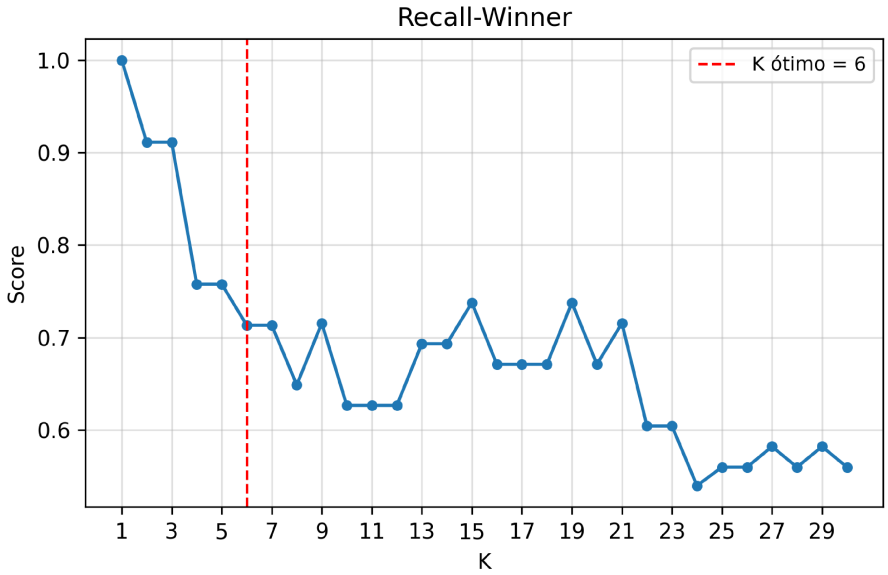
Fonte: Elaborado pelo autor.

Figura 7 – F1-Winner (classe positiva) em validação cruzada (5 *folds*) para diferentes valores de K no KNN.



Fonte: Elaborado pelo autor.

Figura 8 – Recall-Winner em validação cruzada (5 *folds*) para diferentes valores de K no KNN.



Fonte: Elaborado pelo autor.

4 Análise e Discussão dos Resultados

Este capítulo apresenta os resultados obtidos pelos três classificadores nos quatro cenários de dados avaliados. Inicialmente, comparam-se as métricas de acurácia entre os modelos. Em seguida, analisa-se a importância relativa das *features* no Random Forest, detalham-se as probabilidades de vitória atribuídas a cada jogo do conjunto de teste, e discutem-se as matrizes de confusão. Por fim, comparam-se os resultados com os obtidos por Oliveira (2023) no contexto de premiações cinematográficas.

4.1 Comparação de acurácia por cenário

Os modelos foram avaliados em quatro cenários de combinação de *features*: *i*) apenas *features* de críticos, *ii*) apenas *features* de usuários, *iii*) combinação de ambas as fontes, e *iv*) variável ponderada 90/10. A Tabela 4 apresenta a acurácia obtida por cada classificador em cada cenário.

Tabela 4 – Acurácia comparativa por cenário e modelo no conjunto de teste (2024–2025).

Cenário de Dados	Naive Bayes	KNN	Random Forest
Apenas Críticos	41,67%	83,33%	**75,00%**
Apenas Usuários	16,67%	**100,00%**	83,33%
Combinação de Fontes	16,67%	83,33%	**91,67%**
Ponderado (90/10)	16,67%	83,33%	**100,00%**

Fonte: Elaborado pelo autor.

4.2 Análise por modelo

O Naive Bayes apresentou o desempenho mais fraco entre os três classificadores. Nos cenários de combinação de fontes, ponderado e apenas usuários, atingiu apenas 16,67% de acurácia, classificando todos os 12 jogos como vencedores. Esse comportamento é uma consequência direta da violação da suposição de independência condicional. As doze *features* estatísticas são altamente correlacionadas entre si (média, mediana e percentis medem aspectos sobrepostos da mesma distribuição), o que compromete fundamentalmente a premissa do algoritmo. No cenário com apenas críticos, no qual a dimensionalidade é menor (seis *features*), o modelo teve uma degradação menor, de 41,67%. Sua precisão foi de 22,22% e a revocação, de 100%. Isso sugere que a correlação entre fontes distintas, como críticos e usuários, agrava o problema.

O KNN (com $K = 6$) apresentou desempenho consistente em três dos quatro cenários e excelente no restante. Seu melhor resultado foi obtido no cenário de apenas usuários (100,00%), classificando corretamente todos os 12 jogos. Nos cenários de apenas críticos, combinação de fontes e ponderado 90/10, atingiu 83,33% de acurácia (10 acertos em 12), errando apenas na identificação dos dois vencedores reais: *Astro Bot* e *Clair Obscur: Expedition 33*, que foram classificados como não vencedores (indicando um falso negativo). Esse padrão indica que, com $K = 6$, o modelo tende a ser conservador na atribuição da classe positiva. Ele classifica corretamente todos os não vencedores, mas não consegue distinguir os vencedores quando suas *features* estão na fronteira de decisão.

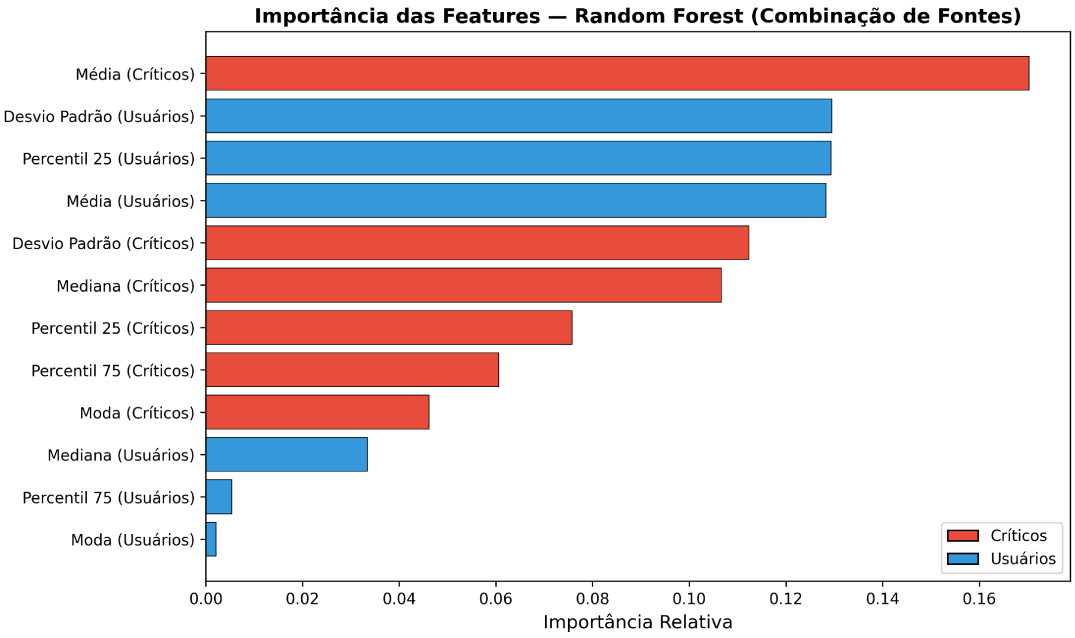
Finalmente, o Random Forest apresentou o desempenho mais consistente e robusto. Seu melhor resultado foi obtido no cenário ponderado 90/10 (100,00%), classificando corretamente todos os 12 jogos do conjunto de teste, com precisão e revocação perfeitas. No cenário de combinação de fontes, atingiu 91,67% de acurácia, com precisão de 100% e revocação de 50%. O único erro foi não detectar *Clair Obscur: Expedition 33* como vencedor (falso negativo). No cenário de apenas usuários, o modelo obteve 83,33%, errando apenas na identificação dos dois vencedores. A superioridade do *ensemble* de árvores de decisão pode ser atribuída a três fatores. Primeiro, a aleatoriedade na seleção de *features* em cada divisão reduz o impacto da correlação entre variáveis. Segundo, a agregação de múltiplas árvores estabiliza a predição e tende a evitar o *overfitting* a padrões espúrios. Por fim, o critério de entropia permite que as árvores identifiquem limiares discriminativos nas *features* mais informativas, como a média dos críticos. No cenário de apenas críticos, o desempenho foi de 75,00%, inferior ao dos demais cenários que combinam ambas as fontes, sugerindo que a integração de avaliações de usuários contribui para a robustez do modelo.

4.3 Importância das *Features*

A Figura 9 apresenta a importância relativa das doze *features* no modelo Random Forest treinado com a combinação de fontes. A *feature* mais influente é a média dos críticos (importância de 0,170), confirmando que o *Metascore* é o preditor dominante para o TGA. Esse resultado é coerente com a regra oficial da premiação, que estabelece um peso de 90% para o júri especializado.

Surpreendentemente, três *features* de usuários ocupam posições de destaque: o desvio padrão (0,130), o percentil 25 (0,129) e a média (0,128). Isso indica que, embora a análise exploratória tenha mostrado pouca separabilidade univariada nas *features* de usuários, o Random Forest consegue explorar interações multivariadas entre essas variáveis e as dos críticos. A consistência e o acordo entre os críticos são o que realmente influenciam a avaliação. Em contrapartida, a moda dos usuários (0,002) e o percentil 75 dos usuá-

Figura 9 – Importância relativa das *features* no modelo *Random Forest* (cenário de combinação de fontes).



Fonte: Elaborado pelo autor.

rios (0,005) apresentaram importância desprezível, sugerindo que o teto das avaliações populares não diferencia vencedores de não vencedores.

4.4 Probabilidades por jogo no conjunto de teste

Para compreender o comportamento dos classificadores em nível individual, as tabelas 5 e 6 apresentam as probabilidades de vitória atribuídas a cada jogo do conjunto de teste por cada modelo, no cenário de combinação de fontes.

Tabela 5 – Probabilidades de vitória por jogo na edição de 2024 (cenário de combinação de fontes).

Jogo	**NB (%)**	**KNN (%)**	**RF (%)**	**Classe Real**
Astro Bot	100,0	50,0	**70,2**	Vencedor
Elden Ring: Shadow of the Erdtree	100,0	50,0	33,6	Não Vencedor
Final Fantasy VII Rebirth	100,0	0,0	19,6	Não Vencedor
Metaphor: ReFantazio	100,0	0,0	16,7	Não Vencedor
Balatro	100,0	33,3	19,9	Não Vencedor
Black Myth: Wukong	100,0	0,0	4,8	Não Vencedor

Fonte: Elaborado pelo autor.

Tabela 6 – Probabilidades de vitória por jogo na edição de 2025 (cenário de combinação de fontes).

Jogo	**NB (%)**	**KNN (%)**	**RF (%)**	**Classe Real**
Clair Obscur: Expedition 33	100,0	50,0	**48,8**	Vencedor
Death Stranding 2: On the Beach	100,0	0,0	3,2	Não Vencedor
Donkey Kong Bananza	100,0	33,3	30,9	Não Vencedor
Hades 2	100,0	0,0	31,7	Não Vencedor
Hollow Knight: Silksong	100,0	33,3	5,8	Não Vencedor
Kingdom Come: Deliverance 2	100,0	0,0	1,6	Não Vencedor

Fonte: Elaborado pelo autor.

A análise caso a caso revela padrões relevantes. Na edição de 2024, o algoritmo Random Forest atribuiu a *Astro Bot* a maior probabilidade de vitória, de 70,2%, e ele foi corretamente classificado como vencedor por esse método. O Naive Bayes também identificou *Astro Bot* como vencedor. Por outro lado, o KNN atribuiu uma probabilidade de 50,0% e classificou-o como não vencedor, resultando em um falso negativo. O segundo colocado em probabilidade pelo Random Forest foi *Elden Ring: Shadow of the Erdtree* (33,6%), que de fato foi amplamente considerado o principal concorrente – um resultado coerente com os dados disponíveis. Finalmente, *Black Myth: Wukong* recebeu a menor probabilidade (4,8%), coerente com sua recepção crítica mais polarizada (desvio padrão elevado), apesar do sucesso comercial expressivo.

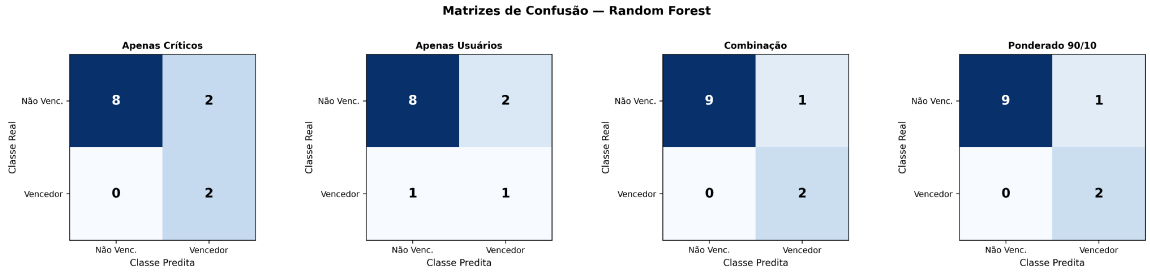
Já na edição de 2025, o Random Forest atribuiu a maior probabilidade a *Clair Obscur: Expedition 33* (48,8%), vencedor da edição, mas classificou-o como não vencedor (falso negativo) no cenário de combinação de fontes. O KNN também classificou *Clair Obscur: Expedition 33* como não vencedor (50,0%), sendo este o único erro do modelo na edição de 2025. Todos os não vencedores de 2025 foram corretamente identificados pelo KNN e pelo Random Forest.

O comportamento do Naive Bayes é muito revelador: ao atribuir probabilidades próximas de 100% a praticamente todos os jogos, o modelo demonstra uma incapacidade completa de discriminar entre as classes, confirmando a inadequação da suposição de independência para este conjunto de dados.

4.5 Matrizes de confusão

A Figura 10 apresenta as matrizes de confusão do Random Forest nos quatro cenários de dados. No cenário de combinação de fontes, o modelo obteve: 10 verdadeiros negativos (não vencedores corretamente identificados), 1 verdadeiro positivo (*Astro Bot*), nenhum falso positivo e 1 falso negativo (*Clair Obscur: Expedition 33*).

Figura 10 – Matrizes de confusão do *Random Forest* nos quatro cenários de dados.



Fonte: Elaborado pelo autor.

No cenário de combinação de fontes, a revocação de 50,00% indica que o modelo identificou um dos dois vencedores reais (*Astro Bot*), não conseguindo detectar *Clair Obscur: Expedition 33*. A precisão de 100% na classe vencedora significa que o único jogo classificado como vencedor era, de fato, vencedor, resultando em *F1-Score* de 66,67%.

No cenário ponderado 90/10, o Random Forest atingiu a configuração ideal: dois verdadeiros positivos, 10 verdadeiros negativos, nenhum falso positivo e nenhum falso negativo – precisão, revocação e F1-Score de 100%. Esse resultado é expressivo por ter sido obtido no cenário que melhor replica a metodologia oficial do TGA. No cenário com apenas usuários, o algoritmo KNN manteve 100% de acurácia. Por outro lado, o Random Forest obteve 83,33% de acertos. Ele classificou incorretamente os dois vencedores como não vencedores. Isso indica que, quando usado isoladamente, o perfil de avaliações de usuários não é suficiente para que o *ensemble* de árvores identifique os vencedores de forma confiável.

4.6 Comparação com o trabalho de referência

Os resultados deste trabalho contrastam significativamente com os obtidos por Oliveira (2023) no contexto cinematográfico. Embora a análise de correlação entre avaliações do Metacritic e os vencedores do Oscar e Globo de Ouro não tenha produzido resultados expressivos, o Random Forest aplicado ao TGA atingiu 100% de acurácia no cenário ponderado 90/10. Este cenário é aquele que mais fielmente replica a metodologia oficial da premiação. Além disso, obteve 91,67% de acurácia no cenário de combinação de fontes. O KNN atingiu 100% no cenário de apenas usuários e 83,33% na combinação.

Essa diferença pode ser atribuída à natureza distinta dos processos de votação: o Oscar depende de uma Academia com critérios internos opacos, fortemente influenciados por campanhas de lobby e dinâmicas políticas internas. Em contraste, o TGA possui uma metodologia documentada e reproduzível, na qual 90% do júri é composto por críticos e 10% pelo público. A transparência dessa regra permite que as notas do Metacritic, que agregam de forma precisa a opinião de críticos e usuários, atuem como um *proxy* mais

direto do processo decisório real. Notavelmente, o cenário ponderado 90/10, ao aplicar os mesmos pesos do TGA às *features* estatísticas, produziu o melhor desempenho isolado de qualquer modelo neste estudo. Isso reforça que a previsibilidade de uma premiação está positivamente correlacionada com a transparência de seus critérios de votação.

5 Conclusão

Este trabalho investigou a correlação entre as avaliações de críticos e usuários registradas no Metacritic e a probabilidade de um jogo ser premiado no The Game Awards. Para isso, adaptou-se ao contexto dos jogos eletrônicos a metodologia proposta por Oliveira (2023) para premiações cinematográficas. As avaliações individuais dos jogos indicados à categoria principal entre 2014 e 2025 foram coletadas por meio de *web scraping*, com um protocolo de filtragem temporal que descarta notas publicadas após cada cerimônia. A partir desses dados, foram extraídas doze *features* estatísticas. Em seguida, treinaram-se três classificadores: Naive Bayes, KNN e Random Forest, usando as edições de 2014 a 2023. Esses classificadores foram testados nas edições de 2024 e 2025, em quatro cenários de combinação de fontes.

Os resultados permitem concluir que existe uma correlação forte e mensurável entre as avaliações agregadas e o resultado da premiação. Diferentemente do cenário cinematográfico, no qual Oliveira (2023) não identificou relação clara entre as notas do Metacritic e os vencedores do Oscar e do Globo de Ouro, o Random Forest aplicado ao TGA atingiu 91,67% de acurácia no cenário de combinação de fontes. Além disso, no cenário ponderado 90/10, a acurácia foi de 100%, com precisão e revocação perfeitas. A hipótese levantada na introdução, portanto, confirma-se: a metodologia produz resultados mais expressivos quando aplicada a uma premiação com critérios específicos e documentados.

O achado central do trabalho reside justamente no cenário ponderado 90/10. Ao aplicar às *features* estatísticas os mesmos pesos que o TGA atribui ao júri de críticos (90%) e à votação popular (10%), obteve-se o melhor desempenho isolado de qualquer modelo neste estudo. Esse resultado sugere que a previsibilidade de uma premiação está positivamente correlacionada com a transparência de seus critérios de votação. Quando a regra decisória é pública e reproduzível, dados abertos que refletem as opiniões dos votantes atuam como um *proxy* eficaz do processo real de escolha.

Quanto ao comportamento dos classificadores, o Random Forest mostrou-se o mais robusto e consistente, beneficiado pela seleção aleatória de atributos, que atenua a correlação entre as *features*, e pela agregação de múltiplas árvores. O KNN (com $K = 6$) obteve acurácias elevadas — 100% no cenário de apenas usuários e 83,33% nos demais —, mas revelou-se conservador na atribuição da classe positiva, produzindo falsos negativos justamente nos dois vencedores reais. O Naive Bayes mostrou-se inadequado para o problema, apresentando acurácias entre 16,67% e 41,67%. Isso ocorre porque a forte correlação entre as estatísticas extraídas de uma mesma distribuição de notas viola a suposição de independência condicional que fundamenta o algoritmo.

A análise da importância das *features* no Random Forest reforça a coerência entre os dados e a regra da premiação. A média das notas dos críticos foi o preditor dominante (0,170), em consonância com o peso de 90% do júri especializado. As *features* de usuários, embora apresentem baixa separabilidade univariada entre as classes, contribuem para o desempenho quando combinadas às *features* dos críticos. Isso indica que o modelo explora interações multivariadas que não são visíveis na análise isolada de cada variável.

Além dos resultados preditivos, o trabalho traz contribuições metodológicas para estudos futuros no domínio. Destacam-se o protocolo de filtragem temporal no nível de cada avaliação individual. Esse protocolo previne o vazamento de dados causado por avaliações retrospectivas posteriores às cerimônias. Além disso, há a consolidação multiplataforma das notas, que alinha a fonte de dados ao critério de julgamento do TGA. O código-fonte e as bases de dados estão publicamente disponíveis, permitindo a reprodução e a extensão dos experimentos.

5.1 Limitações

Os resultados devem ser interpretados à luz de limitações importantes. A principal delas é o tamanho do conjunto de teste: as edições de 2024 e 2025 somam apenas 12 jogos, dos quais somente dois são vencedores. Nesse contexto, a acurácia de 100% obtida pelo Random Forest no cenário ponderado não possui significância estatística robusta. Além disso, não foram calculados intervalos de confiança, testes binomiais ou reamostragens do tipo *bootstrap*, que poderiam quantificar a incerteza associada ao resultado. Portanto, os resultados devem, portanto, ser lidos como evidência inicial favorável à hipótese investigada, e não como demonstração definitiva de capacidade preditiva.

O escopo do estudo também é restrito: apenas a categoria principal (*Game of the Year*) foi analisada, e o histórico do TGA é curto, com dez edições disponíveis para treinamento. Além disso, há uma limitação metodológica na seleção do hiperparâmetro K do algoritmo KNN. O balanceamento por *Random Over-Sampling* foi aplicado ao conjunto de treinamento antes da validação cruzada. Essa prática permite que réplicas de uma mesma instância apareçam em *folds* diferentes. Como resultado, as métricas obtidas nessa etapa podem estar superestimadas. O procedimento mais rigoroso seria aplicar o balanceamento em cada *fold*, por meio de um pipeline. Por fim, o trabalho depende integralmente do Metacritic, herdando a opacidade da ponderação do Metascore e os vieses inerentes às avaliações de usuários, como o viés de seleção e o *review bombing*, discutidos no Capítulo 2.

5.2 Trabalhos Futuros

A partir dos resultados e das limitações identificadas, quatro direções de pesquisa mostram-se promissoras. A primeira é a extensão do estudo a outras categorias do *The Game Awards* e a outras premiações da indústria com critérios de votação distintos, como o *BAFTA Games Awards* e o *DICE Awards*. Essa ampliação permitiria testar, de forma sistemática, a hipótese de que a previsibilidade de uma premiação cresce com a transparência de seus critérios, além de aumentar o volume de dados disponível para treinamento e avaliação.

A segunda direção é a incorporação do conteúdo textual das avaliações por meio de técnicas de processamento de linguagem natural, como a análise de sentimentos com modelos baseados em transformadores. Trabalhos como os de Krauss *et al.* (2008) e Correa *et al.* (2018) evidenciam que a opinião expressa em texto carrega informação preditiva sobre premiações; combinar esse sinal qualitativo com as estatísticas numéricas aqui empregadas pode aumentar o poder discriminativo dos modelos.

A terceira estratégia consiste em enriquecer o conjunto de *features* com variáveis externas às notas. Essas variáveis incluem dados de vendas, engajamento em redes sociais, o número de indicações de cada título em outras categorias da mesma edição e as notas agregadas pelo OpenCritic (OpenCritic, 2026) como fonte complementar. Essa abordagem visa ampliar as informações disponíveis para análise.

Por fim, a quarta direção é reformular o problema como uma tarefa de ordenação, conhecida como *learning to rank*. Nesse novo formato, o objetivo passa a ser prever o vencedor entre os indicados de cada edição. Isso substitui a abordagem anterior, que classificava cada jogo isoladamente como vencedor ou não vencedor. Essa formulação reflete com mais fidelidade a natureza competitiva da premiação — há exatamente um vencedor por edição — e elimina de forma natural o desbalanceamento de classes, dispensando técnicas de reamostragem.

Referências

ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. **IEEE transactions on knowledge and data engineering**, IEEE, v. 17, n. 6, p. 734–749, 2005.

AGGARWAL, C. C. **Recommender systems**. [*S. l.*]: Springer, 2016.

AWARDS, T. G. **About**. [*S. l.: s. n.*], 2024. https://thegameawards.com/about. [Online; accessed 15-Outubro-2024].

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. [*S. l.*]: Springer, 2006. v. 4.

BOYAN, J.; FREITAG, D.; JOACHIMS, T. A Machine Learning Architecture for Optimizing Web Search Engines. *In:* PROCEEDINGS of the AAAI-96 Workshop on Internet-based Information Systems. [*S. l.: s. n.*], 1996.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

CHOWDHARY, K.; CHOWDHARY, K. Natural language processing. **Fundamentals of artificial intelligence**, Springer, p. 603–649, 2020.

CORREA, I. T.; ABDALA, D. D.; MIANI, R. S.; FARIA, E. R. Sentiment Analysis of Twitter Posts About the 2017 Academy Awards. *In:* SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. ANAIS do XV Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). São Paulo: [*s. n.*], 2018. p. 320–331. DOI: 10.5753/eniac.2018.4427.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE transactions on information theory**, IEEE, v. 13, n. 1, p. 21–27, 1967.

ELSHAWI, R.; MAHER, M.; SAKR, S. **Automated Machine Learning: State-of-The-Art and Open Challenges**. [*S. l.: s. n.*], 2019. arXiv: 1906.02287 [cs.LG]. Disponível em: https://arxiv.org/abs/1906.02287.

FAWAGREH, K.; GABER, M. M.; ELYAN, E. Random forests: from theory to practice. **Theoretical Computer Science**, Elsevier, v. 555, p. 2–20, 2014.

FAWCETT, T. An introduction to ROC analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 9, p. 861–874, 2006.

GUNAWAN, R.; RAHMATULLOH, A.; DARMAWAN, I.; FIRDAUS, F. Comparison of Web Scraping Techniques : Regular Expression, HTML DOM and Xpath. *In:* PROCEEDINGS of the 2018 International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018). [*S. l.*]: Atlantis Press, mar. 2019. p. 283–287. ISBN 978-94-6252-689-1. DOI: 10.2991/icoiese-18.2019.50. Disponível em: https://doi.org/10.2991/icoiese-18.2019.50.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. [*S. l.*]: Springer Science & Business Media, 2009.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on knowledge and data engineering**, IEEE, v. 21, n. 9, p. 1263–1284, 2009.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. [*S. l.*]: OTexts, 2018.

JANNACH, D.; ZANKER, M.; FELFERNIG, A.; FRIEDRICH, G. **Recommender systems: an introduction**. [*S. l.*]: Cambridge University Press, 2010.

KRAUSS, J.; NANN, S.; SIMON, D.; FISCHBACH, K.; GLOOR, P. Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis. *In:* ASSOCIATION FOR INFORMATION SYSTEMS. PROCEEDINGS of the 16th European Conference on Information Systems (ECIS). Galway, Irlanda: [*s. n.*], 2008. p. 2026–2037.

LAMBERT, B. **Brookhaven Honors a Pioneer Video Game**. [*S. l.: s. n.*], 2008. https://www.nytimes.com/2008/11/09/nyregion/long-island/09videoli.html?_r=2. [Online; accessed 15-Outubro-2024].

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. **Journal of Machine Learning Research**, v. 18, n. 1, p. 559–563, 2017.

LOWE, D. **Web Scraping of Machine Learning Mastery Articles Using Python and BeautifulSoup**. [*S. l.: s. n.*], 2022. https://dainesanalytics.blog/2022/08/26/web-scraping-of-machine-learning-mastery-articles-using-python-and-beautifulsoup/. [Online; accessed 17-Outubro-2024].

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biology**, v. 52, p. 99–115, 1990. Disponível em: https://api.semanticscholar.org/CorpusID:15619658.

METACRITC. **Metacritic About**. [*S. l.: s. n.*], 2026. Página de About do Metacritic. Disponível em: https://www.metacritic.com/about-us/%22. Acesso em: 1 mar. 2026.

MILNER, P. M. The Mind and Donald O. Hebb. **Scientific American**, v. 268, n. 1, p. 124–129, jan. 1993. DOI: 10.1038/scientificamerican0193-124.

MITCHELL, T. M. **Machine Learning**. [*S. l.*]: McGraw-Hill Education, 1997.

MORTON, W. A $406bn gaming industry and the melee for advertiser's next frontier. **The Drum**, 2023. Disponível em: https://www.thedrum.com/oupinion/2023/11/17/406bn-gaming-industry-and-the-melee-advertisers-next-frontier. Acesso em: 15 out. 2024.

MOSQUEIRA-REY, E.; HERNÁNDEZ-PEREIRA, E.; ALONSO-RÍOS, D.; BOBES-BASCARÁN, J.; FERNÁNDEZ-LEAL, Á. Human-in-the-loop machine learning: a state of the art. **Artificial Intelligence Review**, Springer, v. 56, n. 4, p. 3005–3054, 2023.

MUCHNIK, L.; ARAL, S.; TAYLOR, S. J. Social influence bias: A randomized experiment. **Science**, American Association for the Advancement of Science, v. 341, n. 6146, p. 647–651, 2013.

OLIVEIRA, K. L. S. **Análise da influência de avaliações de críticos e usuários nas premiações do Oscar e do Globo de Ouro em 2023 usando aprendizado de máquina**. 2023. 49 f. Monografia (Trabalho de Conclusão de Curso) – Universidade Federal de Uberlândia.

OLIVEIRA, R. C. de. Ensaio crítico sobre o site Metacritic. **Revista Cronos**, v. 19, n. 1, p. 07–13, maio 2019. DOI: 10.21680/1982-5560.2018v19n1ID17867.

OPENCRITIC. **OpenCritic - About**. [*S. l.: s. n.*], 2026. Disponível em: https://opencritic.com/about. Acesso em: 19 jan. 2026.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PORTUGAL, I.; ALENCAR, P.; COWAN, D. The use of machine learning algorithms in recommender systems: A systematic review. **Expert Systems with Applications**, Elsevier, v. 97, p. 205–227, 2018.

PURCHESE, R. **Obsidian missed out on Fallout: New Vegas bonus because of Metacritic**. [*S. l.: s. n.*], 2012. [Online; accessed 19-Fevereiro-2026]. Disponível em: https://www.eurogamer.net/obsidian-missed-out-on-fallout-new-vegas-bonus-because-of-metacritic.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Introduction to recommender systems handbook. *In:* RECOMMENDER systems handbook. [*S. l.*]: Springer, 2015. p. 1–34.

RISH, I. An empirical study of the naive Bayes classifier. *In:* IJCAI 2001 workshop on empirical methods in AI. [*S. l.: s. n.*], 2001. v. 3, p. 41–46.

ROTTEN TOMATOES. **Rotten Tomatoes - About**. [*S. l.: s. n.*], 2026. Disponível em: https://www.rottentomatoes.com/about. Acesso em: 19 fev. 2026.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210–229, 1959. DOI: 10.1147/rd.33.0210.

SELENIUM. **Selenium Documentation**. [*S. l.: s. n.*], 2024. Disponível em: https://www.selenium.dev/documentation/. Acesso em: 1 mar. 2026.

SHANI, G.; GUNAWARDANA, A. Evaluating recommendation systems. *In:* RECOMMENDER systems handbook. [*S. l.*]: Springer, 2011. p. 257–297.

SIMON, P. **Too Big to Ignore: The Business Case for Big Data**. [*S. l.*]: Wiley, 2013. (Wiley and SAS Business Series). ISBN 9781118642108. Disponível em: https://books.google.com.br/books?id=Dn-Gdoh66sgC.

SOUZA, D. H. M. de; BORDIN JR, C. J. Detecção de fraude de cartão de crédito por meio de algoritmos de aprendizado de máquina. **Revista Brasileira de Computação Aplicada**, v. 15, n. 1, p. 1–11, 2023.

STUART J. RUSSELL, P. N. **Artificial Intelligence: A Modern Approach**. [*S. l.*]: Prentice Hall, 2009.

TIME. **Science: The Goof Button**. [*S. l.: s. n.*], 1961. https://time.com/archive/6810331/science-the-goof-button/. [Online; accessed 8-Novembro-2024].

WALL, A. **History of Search Engines: From 1945 to Google Today**. [*S. l.: s. n.*], 2017. http://www.searchenginehistory.com/. [Online; accessed 30-Outubro-2024].

WIJEKOON, S. M. S. T.; RUPASINGHA, R. A. H. M. Machine Learning Approach for Discovering Successful Books Based on the Achievement of Award-Winning. *In:* IEEE. 2023 3rd International Conference on Advanced Research in Computing (ICARC). [*S. l.: s. n.*], 2023. p. 24–29. DOI: 10.1109/ICARC57651.2023.10145695.

ZHOU, Z. Automatic Machine Learning-Based Data Analysis for Video Game Industry. *In:* IEEE. 2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI). [*S. l.: s. n.*], 2022. p. 732–737. DOI: 10.1109/CEI57409.2022.9950194.

Apêndices

APÊNDICE A – Declaração de uso de IA generativa

Para a produção deste trabalho, foi utilizado Claude Opus (Anthropic, Versão 4.8) para revisão gramatical, indexação de fontes, produção de código e revisão integral. O uso ocorreu nas seguintes etapas: referencial teórico, desenvolvimento e análise e discussão dos resultados. Todo conteúdo gerado pela ferramenta foi revisado criticamente pelo autor, que assume integral responsabilidade pelo trabalho final.