UNIVERSIDADE FEDERAL DO MARANHÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

Predição do Impacto de Curto Prazo de Artigos: O Título como Indicador Chave

João Pedro Cavalcanti Azevedo

São Luís, 2025

João Pedro Cavalcanti Azevedo

Predição do Impacto de Curto Prazo de Artigos: O Título como Indicador Chave

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Antônio de Abreu Batista Junior
UFMA

São Luís
2025

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Cavalcanti Azevedo, João Pedro.
Predição do Impacto de Curto Prazo de Artigos: O Título como Indicador Chave / João Pedro Cavalcanti Azevedo. - 2025.
50 f.

Orientador(a): Antônio de Abreu Batista Junior.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís - Maranhão, 2025.

1. Predição de Citações. 2. Aprendizado de Máquina. 3. Título de Artigo. 4. Processamento de Linguagem Natural (pln). 5. Análise Bibliométrica. I. de Abreu Batista Junior, Antônio. II. Título.

João Pedro Cavalcanti Azevedo

Predição do Impacto de Curto Prazo de Artigos: O Título como Indicador Chave

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Trabalho Aprovado. São Luís, 01 de Agosto de 2025:

Orientador: Prof. Dr. Antônio de Abreu Batista Junior
Orientador
Universidade Federal do Maranhão

Prof. Dr. Alexandre Cesar Muniz de Oliveira
Examinador Interno
Universidade Federal do Maranhão

Prof. Me. Yonara Costa Magalhães
Examinador Externo
Universidade Estadual do Maranhão

São Luís
2025

Dedicatória: Dedico este trabalho aos meus pais, por todo o sacrifício e apoio incondicional que me proporcionaram, tornando possível cada conquista que alcancei. E aos meus cachorros, Scooby e Snoopy, por todo o amor e companhia fiel nos momentos mais difíceis e felizes da jornada.

Agradecimentos

Foram anos de dedicação intensa desde o término do ensino médio em 2018, até a entrada na UFMA em 2019, quando comecei a perseguir o curso dos meus sonhos. Entreguei-me de corpo e alma a essa conquista e hoje estou prestes a concluí-la. Foram momentos de muita luta, com algumas reprovações, mas também excelentes aprovações, incluindo a vitória em uma maratona de programação durante um congresso.

Durante esse percurso, sempre mantive atividades em paralelo à graduação. Iniciei uma segunda graduação em 2019 junto com está, que concluí em 2021, além de cursar três pós-graduações: duas já concluídas, uma em 2022 e outra em 2023, e uma terceira que será finalizada ainda neste ano de 2025. Em 2023, comecei o mestrado em Ciência da Computação, que já foi concluído. Hoje, sou mestre em Ciência da Computação. Em verdade, conquistei o título de mestre antes mesmo de concluir a graduação, pois precisei sacrificar muito dessa etapa para alcançar tantas outras realizações.

Porém, chegou o momento de concluir essa jornada acadêmica e dar à graduação o desfecho digno que ela merece. Tornar-me mestre foi uma imensa alegria, mas ser bacharel em Ciência da Computação representa a concretização dos sonhos daquele garoto que abriu mão de aproveitar o último ano do ensino médio para ser aprovado em uma universidade federal, no curso que sempre desejou.

Cada madrugada sacrificada, cada esforço, tudo valeu a pena. No final, todo o caminho trilhado se justifica pela realização desse sonho.

Ao longo desta trajetória, tive o privilégio de vivenciar experiências que moldaram não apenas minha formação acadêmica, mas também meu desenvolvimento pessoal. Apoios, aprendizados e desafios foram fundamentais para cada conquista, e é com gratidão que registro aqui o reconhecimento àqueles que, direta ou indiretamente, contribuíram para a realização deste trabalho.

Ao Prof. Dr. Antonio de Abreu Batista Junior, deixo minha mais profunda gratidão pela orientação ao longo da graduação, pelo conhecimento compartilhado nas disciplinas cursadas, pelas valiosas sugestões e pela constante disponibilidade. Agradeço por sua paciência, confiança e dedicação, especialmente durante a condução desta pesquisa.

Aos ilustres membros da banca examinadora, Prof. Dr. Antonio de Abreu Batista Junior, Prof. Dr. Alexandre Cesar Muniz de Oliveira e Prof. Me. Yonara Costa Magalhães, registro meus sinceros agradecimentos pelo tempo dedicado à leitura e avaliação deste trabalho, bem como pelas contribuições e sugestões enriquecedoras, que certamente agregaram valor à pesquisa desenvolvida.

Aos meus pais, Vânia e Werbeth, minha eterna gratidão pelo amor incondicional, pelo exemplo de dedicação e pelos inúmeros sacrifícios que permitiram que eu chegasse até aqui. Aos meus tios, tias, primos, primas e avós, agradeço pelo apoio constante, pelos conselhos nos momentos difíceis, pela força transmitida e, sobretudo, pela amizade e carinho que sempre me acompanharam ao longo dessa jornada.

Aos amigos que tive a felicidade de conhecer ao longo desta graduação, e àqueles que sempre estiveram ao meu lado em todos os momentos, expresso minha sincera gratidão. Em especial, agradeço à minha prima Sabrina, e aos queridos amigos Uhilma, Karla, Adonias, Rosy e Thaís, pela amizade genuína, pelo apoio constante e pelas palavras de incentivo que fizeram toda a diferença ao longo desta jornada.

Aos colegas da UFMA e do Laboratório de Sistemas Distribuídos (LSDI), minha profunda gratidão pela parceria, pelos aprendizados compartilhados e pela valiosa contribuição para o desenvolvimento deste trabalho. O companheirismo e a colaboração de vocês foram fundamentais em minha trajetória acadêmica.

Aos professores da graduação, deixo meu sincero agradecimento por cada disciplina ministrada, cada orientação oferecida e por todo o conhecimento transmitido ao longo desses anos. Em especial, estendo minha gratidão também àqueles com os quais enfrentei maiores desafios — inclusive nas disciplinas em que fui reprovado. Essas experiências, ainda que difíceis, foram decisivas para o meu amadurecimento acadêmico e pessoal, fortalecendo minha persistência, resiliência e determinação em seguir em frente.

Por fim, mas não menos importante, minha gratidão aos meus dois companheiros de quatro patas, Scooby e Snoopy. Em meio às longas madrugadas de estudo e aos momentos de cansaço, vocês sempre estiveram por perto, oferecendo companhia, afeto e conforto silencioso. A presença de vocês foi um alívio em dias difíceis e uma alegria constante ao longo dessa jornada.

"Ama-se mais o que se conquista com esforço." Benjamin Disraeli.

Resumo

Este trabalho propõe o desenvolvimento de modelos preditivos capazes de classificar artigos científicos como altamente ou pouco citados em um horizonte de curto prazo, utilizando majoritariamente informações extraídas dos títulos. A hipótese investigada é a de que o título de um artigo, mesmo de forma isolada, pode conter indícios suficientes para prever seu potencial impacto na comunidade científica. Para isso, foram construídos quatro conjuntos de dados (com e sem metadados, balanceados e desbalanceados) e testados nove algoritmos clássicos de aprendizado supervisionado. Os experimentos revelaram que, mesmo com entradas textuais mínimas (somente o título), é possível obter desempenhos competitivos na predição de impacto. A adição de metadados forneceu ganhos marginais em algumas configurações, mas não superou significativamente os resultados baseados apenas no título. Essa constatação reforça a relevância dos títulos como indicadores-chave na avaliação precoce do impacto científico.

Palavras-chave: Predição de Citações; Aprendizado de Máquina; Título de Artigo; Processamento de Linguagem Natural (PLN); Análise Bibliométrica; Citações.

Abstract

This study proposes the development of predictive models capable of classifying scientific articles as highly or lowly cited in the short term, based primarily on information extracted from their titles. The central hypothesis investigated is that an article's title alone may contain sufficient signals to forecast its future scientific impact. To test this, four datasets were built (with and without metadata, balanced and unbalanced), and nine classical supervised learning algorithms were evaluated. The experiments demonstrated that even with minimal textual input (titles only), competitive performance can be achieved in citation prediction. Adding metadata provided marginal improvements in some settings but did not significantly outperform the models based solely on titles. These findings highlight the value of titles as key indicators for early impact assessment in scientific research.

Keywords: Citation Prediction; Machine Learning; Article Title; Natural Language Processing (NLP); Bibliometric Analysis; Citations.

Lista de ilustrações

Lista de tabelas

Lista de Siglas

APS *American Physical Society.*

AUC *Area Under the Curve.*

BGW *Bag-of-Words.*

GBDT *Gradient Boosting Decision Trees.*

GPT *Generative Pre-Trained Transformer.*

LightGBM *Light Gradient Boosting Machine.*

Linear SVC *Linear Support Vector Classifier.*

LLMs *Large Language Models.*

LSTM *Long Short-Term Memory.*

MLP *MLPClassifier.*

MNB *Multinomial Naive Bayes.*

PLN Processamento de Linguagem Natural.

ROC *Receiver Operating Characteristic.*

SGDC *Stochastic Gradient Descent.*

SVM *Support Vector Machine.*

TF-IDF *Term Frequency-Inverse Document Frequency.*

XGBoost *eXtreme Gradient Boosting.*

Sumário

1 Introdução

O número de citações que um artigo científico recebe é amplamente utilizado como uma métrica para avaliar o impacto de uma pesquisa e a relevância de seus autores (GOODWIN, 1980; BORNMANN, 2017). No entanto, esse indicador só se torna disponível após um período significativo de tempo, geralmente anos após a publicação, o que limita sua utilidade para avaliações imediatas e decisões estratégicas em ambientes acadêmicos e institucionais (WANG; SONG; BARABÁSI, 2013).

Diante desse cenário, torna-se relevante o desenvolvimento de modelos capazes de prever, com base em informações disponíveis no momento da publicação, o potencial de um artigo vir a ser altamente citado (ZHANG; WU, 2020; STEGEHUIS; LITVAK; WALTMAN, 2015a). Essa tarefa, embora desafiadora, pode contribuir para acelerar a identificação de trabalhos promissores e otimizar processos de seleção, divulgação e financiamento de pesquisas (SADEQI-ARANI; KADKHODAIE, 2023).

Estudos anteriores demonstraram que é possível prever com razoável precisão a contagem futura de citações por meio do uso combinado de técnicas de aprendizado de máquina e variáveis como texto do título, resumo, termos de indexação e características bibliométricas dos autores e instituições. Em especial, Fu e Aliferis (2008) mostraram que modelos supervisionados podem antecipar o impacto de artigos biomédicos até uma década após a publicação, utilizando exclusivamente dados disponíveis no momento da submissão. Seus resultados indicam que o uso integrado de variáveis textuais e bibliométricas permite alcançar altos níveis de acurácia preditiva, especialmente quando combinados em modelos robustos como *Support Vector Machine* (SVM).

Inspirado por essa linha de pesquisa, neste trabalho, avaliamos a capacidade dos modelos de aprendizagem máquina para prever o impacto de artigos científicos a partir de informações limitadas provenientes dos metadados associados, principalmente o título. O foco está em identificar se tais dados, geralmente disponíveis logo após a submissão de um artigo, podem ser suficientes para estimar sua probabilidade de atingir altos níveis de citação no futuro. Avaliamos diferentes arquiteturas de modelos com informações limitadas, desde técnicas tradicionais até modelos baseados em representações semânticas mais sofisticadas, buscando não apenas confirmar nossa suspeita, mas também compreender quais características mais contribuem para o sucesso de um artigo na literatura científica.

1.1 Caracterização do Problema

A principal limitação do uso de citações como métrica de impacto é sua natureza retrospectiva: somente após meses ou anos da publicação é possível observar sua influência na literatura. Isso cria um vácuo temporal entre a produção do conhecimento e sua efetiva validação pela comunidade científica (AGARWAL *et al.*, 2016; BORNMANN; DANIEL, 2008).

Diante disso, surge a seguinte questão de pesquisa: **é possível prever, com base apenas no título ou em um conjunto limitado de informações iniciais de um artigo, se ele será altamente citado no futuro?** Essa tarefa apresenta múltiplos desafios, tanto técnicos quanto conceituais.

Do ponto de vista técnico, destacam-se: (i) a seleção adequada de variáveis preditoras, geralmente limitadas ao título, autores e outros metadados disponíveis no momento da submissão; e (ii) o desbalanceamento das classes, já que a maioria dos artigos científicos tende a receber um número reduzido de citações ao longo do tempo.

Conceitualmente, o impacto de um artigo depende de múltiplos fatores, nem todos acessíveis ou quantificáveis no momento da publicação, como a relevância do tema, a visibilidade do periódico, o prestígio dos autores e até mesmo aspectos contextuais e temporais que influenciam sua difusão (TAHAMTAN; AFSHAR; AHAMDZADEH, 2016; ZHANG *et al.*, 2020).

1.2 Relevância do Trabalho

A relevância deste estudo reside na possibilidade de transformar o processo de avaliação científica, tradicionalmente reativo, em uma abordagem proativa e orientada por dados. A capacidade de prever o impacto futuro de artigos pode beneficiar diferentes atores no ecossistema da pesquisa: editores podem destacar trabalhos promissores, instituições podem alocar recursos com maior eficiência, e pesquisadores podem compreender melhor os fatores que influenciam a visibilidade e o reconhecimento de seus trabalhos.

Além disso, ao explorar o potencial de modelos baseados em representações textuais, com foco no título dos artigos e técnicas modernas de aprendizado de máquina, esta pesquisa contribui para o avanço do estado da arte na interseção entre bibliometria, Processamento de Linguagem Natural (PLN) e ciência de dados.

1.3 Hipótese de Pesquisa

Nossa hipótese é que, para publicações científicas, modelos de previsão de alta citação baseados exclusivamente no título demonstram uma capacidade preditiva superior

ou comparável em relação a modelos que incorporam o título e outros metadados do artigo.

1.4 Objetivos

1.4.1 Geral

Este trabalho propõe o desenvolvimento de um modelo preditivo para classificar artigos científicos com base em seu potencial de citação a curto prazo, utilizando exclusivamente o título ou o título com informações adicionais simples (ano de publicação, número de autores e presença de pontuação).

1.4.2 Específicos

Os objetivos específicos desta monografia são:

1. Construir um conjunto de dados contendo metadados de artigos científicos, com ênfase nas informações extraídas dos títulos e na quantidade de citações recebidas;
2. Desenvolver um classificador binário capaz de prever se um artigo pertence ao grupo dos mais citados, utilizando o título como principal fonte de informação;
3. Investigar se a confiança do modelo preditivo é superior na classificação de artigos com base apenas no título, em comparação àqueles que utilizam também metadados adicionais.

1.5 Organização do Trabalho

Este trabalho está estruturado em seis capítulos, organizados da seguinte forma:

- O **Capítulo 2** apresenta os fundamentos teóricos necessários para o desenvolvimento do estudo, incluindo conceitos sobre aprendizado de máquina, processamento de linguagem natural e sobre a citação em si.
- O **Capítulo 3** discute os principais trabalhos relacionados à predição de citações, destacando abordagens, técnicas e lacunas presentes na literatura atual.
- O **Capítulo 4** descreve a metodologia adotada, detalhando a construção do conjunto de dados, os critérios de seleção, os modelos utilizados e as etapas do processo experimental.
- O **Capítulo 5** apresenta os resultados obtidos, incluindo a análise do desempenho dos modelos, discussão dos achados e comparação entre diferentes abordagens.

- O **Capítulo 6** traz as considerações finais, destacando as contribuições do trabalho, suas limitações e sugestões para pesquisas futuras.

2 Fundamentação Teórica

Este capítulo apresenta os conceitos e fundamentos teóricos que embasam o presente trabalho. Inicialmente, discute-se a importância das citações científicas como indicador de impacto acadêmico e as principais variáveis que influenciam o número de citações de um artigo, contextualizando o problema da previsão de artigos altamente citados.

Em seguida, introduz-se o PLN, área fundamental para a análise dos títulos dos artigos, que são a principal fonte de informação utilizada para realizar a predição. São descritas as técnicas de representação textual, em especial a vetorização por *Term Frequency-Inverse Document Frequency* (TF-IDF), que permitem transformar dados textuais em formatos compatíveis com algoritmos de aprendizado de máquina.

O capítulo também aborda os conceitos essenciais do aprendizado supervisionado, metodologia adotada para a construção dos classificadores. Apresentam-se os principais algoritmos utilizados no estudo, incluindo modelos lineares, métodos baseados em árvores e redes neurais, detalhando seus princípios de funcionamento e as razões para sua escolha.

2.1 Citação

A citação científica é um dos principais indicadores de impacto acadêmico, sendo amplamente utilizada para avaliar a relevância e a influência de artigos, autores, periódicos e instituições na comunidade científica (GARFIELD, 1955). Quando um artigo é citado por outro, pressupõe-se que ele contribuiu de alguma forma para o desenvolvimento do trabalho subsequente, seja por fornecer fundamentos teóricos, metodológicos ou resultados relevantes (TAHAMTAN; BORNMANN, 2019).

Apesar de a citação ser um fenômeno complexo e multifacetado, ela é frequentemente utilizada como métrica quantitativa para análises bibliométricas e decisões em políticas científicas, como concessão de financiamentos, promoções acadêmicas e ranqueamento de periódicos (TAHAMTAN; AFSHAR; AHAMDZADEH, 2016). Por essa razão, prever se um artigo será ou não altamente citado tornou-se um problema de interesse tanto para pesquisadores da área de ciência da informação quanto para especialistas em aprendizado de máquina (BORNMANN, 2017).

Entretanto, prever o impacto futuro de uma publicação não é uma tarefa trivial, pois as citações são influenciadas por uma combinação de fatores, incluindo:

- Qualidade e inovação do conteúdo;
- Reputação dos autores e da instituição de afiliação;

- Visibilidade do periódico (fator de impacto, indexação, acesso aberto);
- Título do artigo e presença de palavras-chave relevantes;
- Área de pesquisa e atualidade do tema.

Neste trabalho, adota-se como definição de artigo *altamente citado* aquele cujo número de citações ultrapassa a média geral da base de dados utilizada. Essa definição binária permite transformar o problema em uma tarefa de classificação supervisionada, conforme discutido nas seções anteriores. A abordagem proposta neste estudo busca investigar se, com base apenas no título do artigo (ou no título em conjunto com metadados simples), é possível prever seu potencial de citação futura, contribuindo assim para o avanço de métodos automatizados de avaliação e recomendação científica.

2.2 Processamento de Linguagem Natural

O PLN é uma área da inteligência artificial que estuda a interação entre humanos e computadores por meio da linguagem natural, permitindo que sistemas computacionais analisem, interpretem e extraiam significado de textos (CAMBRIA *et al.*, 2013). No presente trabalho, o PLN é aplicado exclusivamente aos títulos de artigos científicos, visando prever se um artigo será altamente citado. As principais técnicas utilizadas:

- Tokenização: Os textos foram segmentados em palavras (tokens), permitindo sua posterior vetorização.
- Vetorização com TF-IDF: A principal técnica de representação textual empregada neste trabalho foi o TF-IDF, que transforma os títulos em vetores numéricos, atribuindo maior peso a palavras informativas que aparecem com frequência em um documento, mas não em toda a coleção.

A aplicação de técnicas de PLN sobre os títulos é central neste trabalho, pois parte da hipótese de que o título, mesmo isoladamente, pode conter informações suficientes para prever o impacto de um artigo científico. Estudos anteriores apontam que elementos textuais como o estilo, a estrutura e o vocabulário de um título estão relacionados ao seu potencial de engajamento e citação (ROSSI; BRAND, 2020; PAIVA; LIMA; PAIVA, 2012; YAN; DING, 2011).

2.3 Aprendizado Supervisionado

O aprendizado supervisionado é uma das principais abordagens dentro do campo do aprendizado de máquina (NASTESKI, 2017). Seu objetivo é construir modelos capazes de

prever um resultado (ou rótulo) com base em um conjunto de dados de entrada previamente rotulado. Em outras palavras, o modelo aprende a partir de exemplos onde a resposta correta é conhecida, e utiliza esse conhecimento para generalizar e fazer previsões em novos dados.

Formalmente, dado um conjunto de treinamento $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, onde $\mathbf{x}_i \in \mathbb{R}^d$ representa o vetor de atributos da amostra i e $y_i \in \mathcal{Y}$ é seu respectivo rótulo, o objetivo do aprendizado supervisionado é encontrar uma função $f : \mathbb{R}^d \to \mathcal{Y}$ que minimize o erro de predição em novos exemplos.

Essa abordagem é dividida em duas principais categorias:

- **Classificação**: quando os rótulos y_i pertencem a um conjunto discreto (por exemplo, 0 ou 1). Este é o caso deste trabalho, em que o objetivo é classificar se um artigo será *altamente citado* (classe positiva) ou não (classe negativa).
- **Regressão**: quando os rótulos y_i são valores contínuos, como a previsão de temperatura ou preço.

Durante o treinamento, o modelo ajusta seus parâmetros internos de forma a minimizar uma função de perda (*loss function*), que quantifica o erro entre a predição do modelo $\hat{y}_i = f(\mathbf{x}_i)$ e o valor real y_i. No caso da classificação binária, uma função de perda comum é a entropia cruzada:

$$\mathcal{L}(y, \hat{y}) = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right]. \tag{2.1}$$

Uma vez treinado, o modelo pode ser aplicado a novos dados não vistos para realizar previsões. Para avaliar seu desempenho, utilizam-se métricas como acurácia, precisão, *recall*, F1-Score e área sob a curva *Receiver Operating Characteristic* (ROC), conforme discutido nos capítulos posteriores.

Este trabalho emprega o aprendizado supervisionado para abordar o problema de prever se um artigo científico será altamente citado com base apenas em seu título (ou título + metadados), formulando essa tarefa como um problema de classificação binária. A seguir, são apresentados os principais algoritmos de aprendizado supervisionado empregados neste estudo, detalhando seu funcionamento, características e justificativas para sua escolha.

2.3.1 Regressão Logística

A Regressão Logística (de Menezes *et al.*, 2017) é um modelo linear probabilístico amplamente empregado em classificação binária. Diferentemente de métodos de regressão tradicional, este modelo estima a probabilidade $P(y = 1|\mathbf{x})$ de uma instância $\mathbf{x}$ pertencer à classe positiva por meio da **função sigmoide**:

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}} \tag{2.2}$$

onde:

- $\mathbf{w} \in \mathbb{R}^d$ é o vetor de pesos aprendidos
- $\mathbf{x} \in \mathbb{R}^d$ representa as *features* da instância
- $b \in \mathbb{R}$ é o termo de viés (*bias*)
- $\sigma(\cdot)$ mapeia valores reais para o intervalo $(0, 1)$

O treinamento do modelo otimiza os parâmetros $\mathbf{w}$ e b mediante a minimização da **função de entropia cruzada binária**:

$$\mathcal{L}(\mathbf{w}, b) = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log \sigma(\mathbf{w}^T\mathbf{x}_i + b) + (1 - y_i)\log(1 - \sigma(\mathbf{w}^T\mathbf{x}_i + b))\Big] \tag{2.3}$$

Destacam-se três vantagens fundamentais:

1. **Interpretabilidade**: Os coeficientes $\mathbf{w}$ indicam a influência relativa de cada *feature*;
2. **Eficiência**: Custo computacional reduzido durante treino/inferência;
3. **Versatilidade**: Integração eficaz com representações textuais como TF-IDF e *word embeddings*.

2.3.2 SGDClassifier

O *SGDClassifier* (KABIR *et al.*, 2015) é um classificador linear que utiliza o algoritmo de *Stochastic Gradient Descent* (SGDC) para otimizar funções de perda. Esse algoritmo é particularmente eficiente em problemas de alta dimensionalidade e grandes volumes de dados, como aqueles derivados da vetorização de textos com TF-IDF.

O objetivo do SGDC é encontrar os parâmetros $\mathbf{w}$ (vetor de pesos) que minimizam uma função de perda $L(\mathbf{w})$ associada ao modelo. Em vez de calcular o gradiente da função de perda em relação a todo o conjunto de dados (como no gradiente descendente tradicional), o SGD atualiza os pesos iterativamente com base em uma única amostra ou um pequeno lote de exemplos a cada passo. Isso torna o treinamento muito mais rápido e escalável.

A atualização dos pesos $\mathbf{w}$ em cada iteração t segue a seguinte equação:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \nabla L(\mathbf{w}_t, x_i, y_i) \tag{2.4}$$

Onde:

- $\mathbf{w}_t$ é o vetor de pesos na iteração t,
- η é a taxa de aprendizado (learning rate),
- $\nabla L(\mathbf{w}_t, x_i, y_i)$ é o gradiente da função de perda em relação aos pesos, computado com base na amostra (x_i, y_i).

O *SGDClassifier* do *scikit-learn* permite utilizar várias funções de perda, como:

- `hinge`: correspondente à SVM linear,
- `log`: correspondente à regressão logística,
- `modified_huber`, `squared_hinge`, entre outras.

O desempenho do *SGDClassifier* pode variar de acordo com a escolha da função de perda, regularização, taxa de aprendizado e número de iterações.

2.3.3 Random Forest

O *Random Forest* (PAL, 2005) é um algoritmo de aprendizado supervisionado baseado em um conjunto de árvores de decisão (*Decision Trees*) treinadas de forma independente e combinadas por meio de uma técnica chamada *bagging* (*Bootstrap Aggregating*). Esse método foi proposto por Breiman (BREIMAN, 2001) e tem como objetivo melhorar a capacidade preditiva das árvores individuais, reduzindo a variância e aumentando a generalização.

A ideia central do *Random Forest* é construir múltiplas árvores de decisão sobre subconjuntos aleatórios dos dados de treinamento e, em seguida, combinar suas predições por votação (classificação) ou média (regressão). Cada árvore é treinada com:

- Um subconjunto de instâncias obtido via amostragem com reposição (*bootstrap*),
- Um subconjunto aleatório dos atributos selecionado em cada divisão do nó (isso introduz diversidade entre as árvores).

Na tarefa de classificação, cada árvore vota em uma classe, e a predição final do *Random Forest* é a classe mais votada entre as árvores:

$$\hat{y} = \text{mode}\{T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_M(\mathbf{x})\} \tag{2.5}$$

Onde:

- $T_m(\mathbf{x})$ é a predição da m-ésima árvore,
- M é o número total de árvores na floresta,
- $\hat{y}$ é a classe prevista final.

A construção de cada árvore utiliza o critério de impureza, sendo o mais comum o índice de Gini. Dado um nó com proporção p_i de exemplos da classe i, a impureza de Gini é definida por:

$$G = 1 - \sum_{i=1}^{C} p_i^2 \tag{2.6}$$

Onde:

- C é o número de classes possíveis,
- p_i é a fração de amostras da classe i no nó.

A divisão ideal em cada nó da árvore é aquela que resulta na maior redução de impureza (também chamada de ganho de Gini). O *Random Forest* explora diversas divisões, avaliando combinações diferentes de atributos e thresholds para maximizar esse ganho.

Uma das principais vantagens do *Random Forest* é sua robustez contra o sobreajuste, especialmente quando o número de árvores é suficientemente grande. Ele também lida bem com dados de alta dimensionalidade e fornece estimativas de importância dos atributos com base na frequência com que são usados para dividir os dados ao longo das árvores.

2.3.4 LightGBM

O *Light Gradient Boosting Machine* (LightGBM) (KE *et al.*, 2017) é um algoritmo de aprendizado supervisionado baseado em *Gradient Boosting Decision Trees* (GBDT), otimizado para alta eficiência e escalabilidade. Ele foi desenvolvido pela Microsoft e é amplamente utilizado devido à sua rapidez, baixo consumo de memória e bom desempenho preditivo, especialmente em tarefas com grandes volumes de dados e alta dimensionalidade.

A técnica de *boosting* consiste em combinar múltiplos modelos fracos (tipicamente, árvores de decisão de pouca profundidade) de forma sequencial, onde cada novo modelo é treinado para corrigir os erros cometidos pelos anteriores. O LightGBM implementa essa

ideia com importantes melhorias, como a técnica de crescimento de árvore baseada em folha e estratégias de amostragem mais eficientes. Suponha que temos um conjunto de dados de treinamento $\{(x_i, y_i)\}_{i=1}^{n}$ e queremos minimizar uma função de perda diferenciável $\mathcal{L}$ entre a predição $F(x)$ e o rótulo real y. O modelo é atualizado iterativamente da seguinte forma:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{2.7}$$

Onde:

- $F_m(x)$ é a predição após m iterações,
- η é a taxa de aprendizado (*learning rate*),
- $h_m(x)$ é o novo modelo treinado na iteração m para ajustar os resíduos (gradientes negativos da função de perda).

A função $h_m(x)$ é construída para minimizar a perda residual usando gradiente descendente. O gradiente da função de perda em relação às predições do modelo é:

$$g_i = \frac{\partial \mathcal{L}(y_i, F(x_i))}{\partial F(x_i)} \tag{2.8}$$

O objetivo é ajustar $h_m(x)$ de forma que aproxime $-g_i$.

Diferente de outros métodos que crescem árvores de forma nível a nível (nivelado), o LightGBM cresce árvores de forma orientada a folhas, isto é, ele expande o nó com maior ganho de perda:

- Isso permite que o modelo encontre regiões com maior erro residual mais rapidamente,
- Porém, também aumenta o risco de sobreajuste — o que pode ser controlado por hiperparâmetros como *max_depth* e *min_data_in_leaf*.

2.3.5 XGBoost

O *eXtreme Gradient Boosting* (XGBoost) (AYDIN; OZTURK, 2021) é uma biblioteca de *boosting* baseada em gradiente que foi projetada para ser altamente eficiente, escalável e precisa. Ele segue o princípio do *gradient boosting*, no qual modelos fracos, tipicamente árvores de decisão, são adicionados sequencialmente para corrigir os erros dos modelos anteriores. No XGBoost, isso é feito minimizando uma função de perda com regularização, o que melhora a capacidade de generalização do modelo.

A função objetivo do XGBoost é composta por dois termos: a função de perda que mede o erro entre as previsões e os valores reais, e um termo de regularização que penaliza a complexidade do modelo:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{2.9}$$

onde:

- l é a função de perda (por exemplo, erro quadrático médio),
- y_i é o rótulo real e $\hat{y}_i^{(t)}$ é a previsão na iteração t,
- f_k representa a k-ésima árvore de decisão,
- $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ é o termo de regularização, com T sendo o número de folhas na árvore, w os pesos das folhas, γ e λ parâmetros de regularização.

Uma das principais inovações do XGBoost é o uso de uma aproximação de segunda ordem da função de perda via a expansão de Taylor:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{2.10}$$

onde g_i e h_i são, respectivamente, o gradiente e o hessiano da função de perda em relação à previsão anterior $\hat{y}_i^{(t-1)}$.

O XGBoost também implementa várias otimizações práticas, como:

- Processamento paralelo durante o treinamento,
- Suporte a dados esparsos,
- Poda de árvores durante o crescimento,
- Cache de blocos de estrutura para acelerar o aprendizado.

Essas características fazem do XGBoost uma das escolhas mais populares em competições de ciência de dados e aplicações práticas com dados tabulares.

2.3.6 Gradient Boosting

O *Gradient Boosting* (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021) é uma técnica de aprendizado de máquina que constrói um modelo preditivo forte a partir da combinação sequencial de vários modelos fracos, geralmente árvores de decisão rasas. A

ideia central é treinar cada novo modelo para corrigir os erros residuais dos modelos anteriores, minimizando uma função de perda por meio de um processo iterativo baseado no gradiente descendente.

A previsão final do modelo após m iterações é dada por:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x), \tag{2.11}$$

onde:

- $F_m(x)$ é a previsão acumulada após a m-ésima iteração,
- $h_m(x)$ é o modelo fraco treinado para ajustar os resíduos da iteração anterior,
- ν é a taxa de aprendizado (*learning rate*), que controla o impacto de cada novo modelo.

O processo de treinamento consiste em encontrar, em cada passo, o modelo $h_m(x)$ que melhor aproxima o gradiente negativo da função de perda $L(y, F(x))$ em relação à predição atual:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \tag{2.12}$$

onde r_{im} são os resíduos (gradientes negativos) para a amostra i na iteração m.

O novo modelo fraco h_m é ajustado para prever esses resíduos, aproximando a direção de maior descida da função de perda. A predição é então atualizada adicionando o ajuste ponderado pelo *learning rate*.

As principais vantagens do *Gradient Boosting* incluem a capacidade de lidar com diferentes tipos de funções de perda, flexibilidade na escolha do modelo base e a capacidade de produzir modelos altamente precisos. No entanto, a técnica pode ser suscetível ao sobreajuste se o número de iterações for muito grande ou se a taxa de aprendizado não for adequadamente regulada.

Métodos derivados, como o XGBoost e o LightGBM, implementam otimizações e melhorias para acelerar o treinamento e melhorar a generalização.

2.3.7 Multinomial Naive Bayes

O *Multinomial Naive Bayes* (MNB) (XU; LI; WANG, 2017) é um algoritmo probabilístico amplamente utilizado para tarefas de classificação de texto, especialmente em modelos de representação como *Bag-of-Words* (BGW) e TF-IDF. Ele se baseia no

teorema de Bayes e na suposição de independência condicional entre as características (palavras) dado a classe.

O objetivo é calcular a probabilidade posterior da classe C_k dado um vetor de atributos (tokens) $x = (x_1, x_2, ..., x_n)$, usando a fórmula de Bayes:

$$P(C_k \mid x) = \frac{P(C_k) \cdot P(x \mid C_k)}{P(x)}. \tag{2.13}$$

Como o denominador $P(x)$ é comum a todas as classes, pode-se utilizar apenas o numerador para fins de classificação:

$$\hat{y} = \arg\max_{C_k} P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k). \tag{2.14}$$

No caso do MNB, assume-se que as ocorrências dos termos seguem uma distribuição multinomial. Assim, a probabilidade condicional $P(x_i \mid C_k)$ é estimada pela frequência relativa do termo x_i na classe C_k:

$$P(x_i \mid C_k) = \frac{N_{ik} + \alpha}{N_k + \alpha \cdot V}, \tag{2.15}$$

onde:

- N_{ik} é o número de vezes que o termo x_i aparece em documentos da classe C_k,
- N_k é o número total de palavras em documentos da classe C_k,
- V é o tamanho do vocabulário (número de palavras únicas),
- α é o parâmetro de suavização de Laplace (tipicamente $\alpha = 1$).

O MNB é particularmente eficaz em tarefas de classificação de texto devido à sua simplicidade, baixo custo computacional e bom desempenho em domínios com alto número de atributos esparsos, como é o caso de representações textuais vetorizadas.

2.3.8 Linear SVC

O *Linear Support Vector Classifier* (Linear SVC) (YANG; LI; YANG, 2015) é uma variação do algoritmo de SVM, utilizado para tarefas de classificação binária. A ideia central da SVM é encontrar um hiperplano que maximize a margem entre duas classes no espaço de características. No caso da Linear SVC, assume-se que os dados são linearmente separáveis (ou aproximadamente separáveis), e o objetivo é resolver o seguinte problema de otimização:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{2.16}$$

sujeito a:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n, \tag{2.17}$$

onde:

- $\mathbf{w}$ é o vetor de pesos que define o hiperplano,
- b é o termo de bias (viés),
- C é o parâmetro de regularização que controla o trade-off entre maximizar a margem e minimizar os erros de classificação,
- ξ_i são variáveis de folga que permitem a penalização de erros,
- $y_i \in \{-1, 1\}$ são os rótulos das classes,
- $\mathbf{x}_i$ são os vetores de características.

A função objetivo combina a maximização da margem (minimização de $\|\mathbf{w}\|^2$) com a penalização por classificações incorretas (via ξ_i).

Na prática, o Linear SVC da biblioteca `scikit-learn` utiliza uma versão otimizada baseada na função de perda de hinge (perda da SVM), definida como:

$$\text{Loss} = \sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) + \frac{\lambda}{2}\|\mathbf{w}\|^2, \tag{2.18}$$

onde $\lambda = 1/C$ é o parâmetro de regularização inversa.

O Linear SVC é particularmente eficaz em conjuntos de dados com grande dimensionalidade (como vetores de texto), sendo uma escolha comum para classificação de textos vetorizados com técnicas como TF-IDF.

2.3.9 MLPClassifier (Rede Neural)

O *MLPClassifier* (MLP) (HARIBABU *et al.*, 2021) é um tipo de rede neural do tipo *feedforward*, composta por camadas totalmente conectadas. Ele é treinado com o algoritmo de retropropagação (*backpropagation*) e utiliza funções de ativação não lineares, como a *ReLU*.

A saída de uma camada oculta é dada por:

$$\mathbf{h} = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1), \tag{2.19}$$

e a predição final por:

$$\hat{y} = \phi(\mathbf{W}_2\mathbf{h} + \mathbf{b}_2), \tag{2.20}$$

onde σ é geralmente a ReLU e ϕ pode ser a função *sigmoid*:

$$\phi(z) = \frac{1}{1 + e^{-z}}. \tag{2.21}$$

A função de perda mais comum para classificação binária é a entropia cruzada:

$$\mathcal{L}(y, \hat{y}) = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right]. \tag{2.22}$$

A otimização dos pesos é feita via gradiente descendente:

$$\theta := \theta - \eta \frac{\partial \mathcal{L}}{\partial \theta}. \tag{2.23}$$

O MLP é eficaz para capturar relações não lineares e possui diversos hiperparâmetros ajustáveis, como número de camadas, neurônios, taxa de aprendizado e função de ativação.

3 Trabalhos Relacionados

Nas últimas décadas, a produção científica tem crescido exponencialmente, transformando a avaliação do impacto da pesquisa em um desafio complexo. Instituições de investigação, agências de financiamento e os próprios pesquisadores enfrentam a difícil tarefa de medir a relevância e a influência de trabalhos acadêmicos em meio a um volume massivo de publicações.

Nesse cenário, as citações se consolidaram como uma das métricas mais usadas para avaliar o impacto acadêmico. A importância e a influência das citações foram primeiramente discutidas por Garfield (1955), que estabeleceu as bases da cienciometria moderna. Posteriormente, Tahamtan e Bornmann (2019) reforçaram o papel crucial das citações, evidenciando sua influência direta em decisões como promoções de carreira, obtenção de financiamento e classificação de revistas científicas.

No entanto, o uso de citações como métrica de impacto só é possível meses ou anos após a publicação, quando é possível observar a sua influência na literatura. Diante desse cenário, a capacidade de prever o impacto de uma publicação científica utilizando apenas informações de metadados (no momento da publicação) tornou-se uma tarefa de grande relevância para a comunidade científica. Isso permite uma avaliação mais precoce do potencial de influência de novos estudos.

Ao mesmo tempo, o avanço de técnicas de aprendizado de máquina e processamento de linguagem natural ampliou significativamente a capacidade de prever citações com base em informações como títulos, resumos e metadados bibliográficos. Por exemplo, Yuan *et al.* (2022) utilizaram redes recorrentes com *Long Short-Term Memory* (LSTM) para modelar padrões de citações, levando em conta efeitos temporais e intrínsecos da publicação.

Muitos trabalhos estudaram este tema. Estudos recentes (JR. *et al.*, 2024) demonstram que modelos pré-treinados, como *Generative Pre-Trained Transformer* (GPT) ou XLNet, combinados com *embeddings* semânticos, podem se aproximar ou superar as representações clássicas baseadas em TF-IDF, especialmente quando aplicados à classificação de impacto baseada em texto.

Por outro lado, Newman (2014) propôs um modelo baseado na vantagem de ser o primeiro a chegar (*first-mover advantage*), segundo o qual os primeiros artigos que exploram um novo campo tendem a receber mais citações ao longo do tempo. O modelo foi validado cinco anos após a sua publicação e demonstrou que os artigos previstos como altamente citados chegaram a receber até 23 vezes mais citações do que a média.

Stegehuis, Litvak e Waltman (2015b) utilizaram a regressão para prever o número

de citações futuras, levando em consideração fatores como o fator de impacto da revista e o número de citações iniciais. Este trabalho demonstrou que, embora exista uma forte correlação com as métricas iniciais, prever o impacto a longo prazo requer informações adicionais. E, Hirako, Sasano e Takeda (2024) propuseram o modelo *CiMaTe*, uma abordagem baseada em modelos de linguagem como o BERT, capaz de prever a contagem futura de citações de um artigo. A diferença do estudo foi o uso do texto completo do artigo dividido em seções estruturadas, o que permitiu melhorar o desempenho preditivo em diferentes domínios científicos.

Também, Suzen *et al.* (2021) exploraram o uso de metadados semânticos, como o título e o resumo, para prever o impacto de citações. Por meio de representações vetoriais e aprendizado supervisionado, alcançaram uma acurácia próxima de 80% na identificação de artigos altamente citados, reforçando a relevância de atributos textuais simples na tarefa de predição. E, Baba, Baba e Ikeda (2019) investigaram a capacidade preditiva de características textuais extraídas apenas do resumo e de dados estruturais. Seus resultados indicaram que informações limitadas, quando bem processadas, já são suficientes para distinguir artigos com alto potencial de citação, aproximando-se do objetivo deste trabalho de explorar o potencial do título isoladamente.

Apesar de avanços recentes, ainda não foram encontrados trabalhos que investiguem exclusivamente o uso do título como variável de entrada principal para prever se um artigo será muito citado. Dessa forma, o presente estudo busca preencher essa lacuna, testando essa hipótese com diferentes configurações de entrada e algoritmos.

4 Procedimentos Metodológicos

Este capítulo descreve detalhadamente os procedimentos adotados para o desenvolvimento e avaliação do experimento proposto, cujo objetivo é investigar a viabilidade de prever, com base apenas no título de um artigo científico, se ele será altamente citado. O processo envolve desde a construção dos conjuntos de dados até a aplicação e análise de diversos algoritmos de aprendizado supervisionado.

4.1 Definição do Problema

A tarefa foi tratada como um problema de classificação binária, no qual cada artigo é rotulado como "altamente citado" ou "pouco citado", a partir da comparação entre o número de citações recebidas e a média da base. A hipótese central é que o título de um artigo carrega informações suficientes para predizer seu impacto em termos de citações.

4.2 Coleta e Pré-processamento dos Dados

Os dados utilizados foram extraídos de [coloque a fonte aqui, ex.: um repositório acadêmico ou base própria]. Cada entrada da base contém o título do artigo, número de autores, ano de publicação e total de citações recebidas.

O pré-processamento incluiu:

- Remoção de duplicatas;
- Tokenização e vetorização via TF-IDF;
- Definição da variável alvo (citações acima da média = 1, caso contrário = 0).

Além disso, foram criadas duas versões principais dos conjuntos de dados:

- **Título somente:** contendo apenas o texto do título como entrada;
- **Título com metadados:** contendo o título e variáveis adicionais, como ano, número de autores e presença de pontuação.

Cada conjunto de dados utilizado neste trabalho foi gerado em duas versões: uma versão desbalanceada, contendo os dados originais, e uma versão balanceada. Para criar o conjunto balanceado, inicialmente selecionou-se todas as amostras da classe positiva do *dataset* original. Em seguida, foram escolhidas aleatoriamente amostras da classe negativa

em quantidade igual ao número de amostras positivas, resultando em um conjunto com classes perfeitamente equilibradas.

4.3 Modelos de Classificação Utilizados

Foram utilizados modelos clássicos de aprendizado supervisionado:

- **Regressão Logística;**
- **SGDClassifier;**
- **Random Forest;**
- **LightGBM;**
- **XGBoost;**
- **Gradient Boosting;**
- **Linear SVC;**
- **Multinomial Naive Bayes;**
- **MLPClassifier (Rede Neural).**

Todos os modelos foram treinados utilizando vetores TF-IDF extraídos dos textos dos títulos. Para os modelos que também utilizaram metadados, as variáveis numéricas e categóricas foram padronizadas ou codificadas conforme necessário.

4.4 Avaliação dos Modelos

A avaliação dos modelos foi feita com 80% dos dados para treino e 20% para teste. As seguintes métricas foram calculadas:

- Acurácia;
- Precisão;
- Recall;
- F1-Score;
- Área sob a curva ROC (AUC).

Para os conjuntos desbalanceados, a atenção foi especialmente voltada ao *recall* e ao *F1-Score*, por serem métricas mais sensíveis à detecção da minoria (artigos altamente citados).

4.5 Ferramentas e Ambiente de Desenvolvimento

O experimento foi conduzido em ambiente Python 3.x com uso das bibliotecas:

- `scikit-learn`;
- `xgboost`, `lightgbm`;
- `pandas`, `numpy`;
- `matplotlib`, `seaborn` (para visualização);
- `imbalanced-learn`.

5 Avaliação Experimental

Este capítulo apresenta a avaliação experimental conduzida com o objetivo central de investigar se é possível prever, com base apenas no título de um artigo científico, se ele será altamente citado no futuro. Essa hipótese fundamenta a ideia de que o título, por si só, pode carregar informações semânticas suficientes para inferir o impacto de uma publicação. A tarefa é tratada como um problema de classificação binária, em que os artigos são rotulados com base na média de citações recebidas no conjunto de dados.

Para testar essa hipótese de forma estruturada, foram construídas diferentes versões do conjunto de dados. A principal delas contém exclusivamente os títulos dos artigos e suas respectivas classes (muito citado ou pouco citado). Outras versões incluem metadados adicionais, como número de autores, ano de publicação e presença de pontuação no título. Essa variação foi projetada propositalmente para avaliar se o acréscimo de informações estruturadas melhora significativamente o desempenho preditivo em relação ao uso exclusivo do título.

Além disso, os *datasets* foram organizados em versões balanceadas e desbalanceadas, com o intuito de medir o impacto da distribuição das classes sobre o desempenho dos modelos. Essa estratégia experimental foi adotada para garantir uma avaliação completa e realista dos algoritmos testados.

O processo experimental contemplou a aplicação de múltiplos modelos clássicos de aprendizado supervisionado, incluindo regressão logística, máquinas de vetores de suporte, árvores de decisão, *ensembles* como *Random Forest*, XGBoost e LightGBM, além de redes neurais simples (*MLPClassifier*). Todos os modelos foram treinados com textos vetorizados utilizando a técnica TF-IDF, e as métricas de avaliação foram analisadas de forma criteriosa para garantir comparabilidade entre abordagens.

A seguir, são descritos os dados utilizados, a configuração dos experimentos, as métricas aplicadas, os resultados obtidos e uma análise crítica sobre o desempenho dos modelos testados.

5.1 Conjunto de Dados

Os dados utilizados neste trabalho foram obtidos a partir do repositório disponibilizado pela *American Physical Society* (APS). Esse repositório reúne metadados de artigos científicos publicados nos periódicos da APS, abrangendo informações como título, autores, instituição de afiliação, ano de publicação, periódico, referências e número de citações recebidas.

A partir dos metadados coletados, foi realizada uma etapa de pré-processamento para organizar e padronizar as informações relevantes. Como parte desse processo, foi calculada a média do número de citações recebidas pelos artigos do conjunto de dados. Essa média resultou em aproximadamente 8 citações por artigo. Com base nesse valor, foi definida uma variável binária de classificação: artigos que receberam mais de 8 citações foram considerados **muito citados** (classe 1), enquanto os que receberam 8 ou menos citações foram rotulados como **pouco citados** (classe 0).

Em seguida, construíram-se diferentes versões do conjunto de dados com o objetivo de testar abordagens variadas de classificação. As principais variações desenvolvidas foram:

- **Dataset com título e label**: composto apenas pelos títulos dos artigos e uma variável binária indicando se o artigo é muito citado ou não.
- **Dataset com título, metadados e label**: além do título, inclui ano de publicação, número de autores e uma indicação se o título possui pontuação ou não, além da variável binária.
- **Versões balanceadas e desbalanceadas dos datasets**: foram geradas versões ajustadas para conter proporções semelhantes (balanceadas) ou desiguais (desbalanceadas) entre artigos muito citados e pouco citados.

Tanto para os conjuntos **desbalanceados** quanto para os **balanceados**, foram elaboradas duas variações principais, totalizando quatro datasets distintos, conforme ilustrado nas Figuras 1 e 2. É importante destacar que todos esses conjuntos possuem a mesma quantidade de exemplos, o que varia entre eles é a forma como as informações estão organizadas na coluna `text`, ou seja, com ou sem metadados adicionais:

- **Versão com título e label**: contém apenas o título do artigo na coluna `text` e a respectiva classe (0 para pouco citado, 1 para muito citado).
- **Versão com título, metadados e label**: a coluna `text` é composta pela combinação de informações como o título, o ano de publicação, o número de autores e uma indicação sobre a presença de pontuação no título, além da respectiva label.

Dessa forma, foram obtidos dois *datasets* com apenas o título (um balanceado e um desbalanceado) e dois com o título acrescido de metadados (também um balanceado e um desbalanceado). Essa estratégia permitiu avaliar o desempenho dos modelos em diferentes contextos, desde situações com entrada textual mínima até configurações mais completas com dados estruturados. O foco principal foi investigar até que ponto informações disponíveis, especialmente o título, são suficientes para prever o impacto futuro de um artigo em termos de citações.

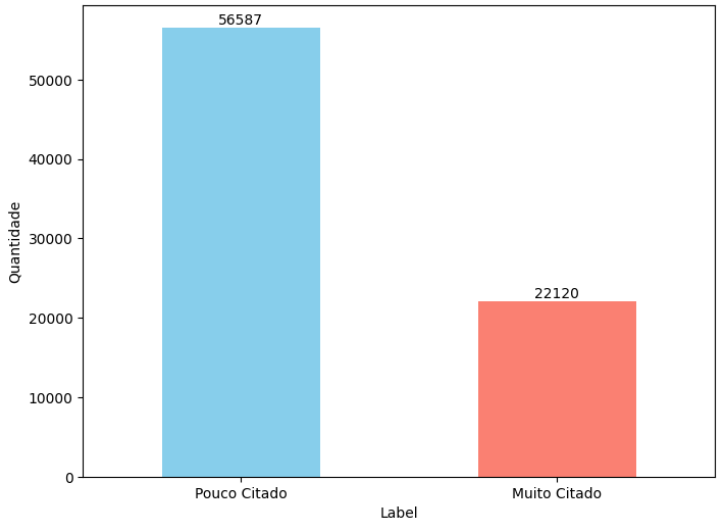

Figura 1 – Versões desbalanceadas do conjunto de dados.

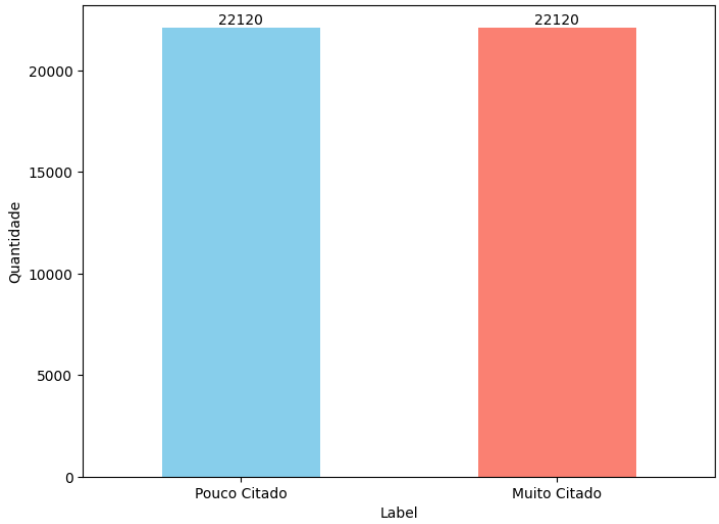

Figura 2 – Versões balanceadas do conjunto de dados.

5.2 Configuração experimental

Todos os experimentos utilizaram uma divisão padrão de 80% para treino e 20% para teste, com estratificação das classes para preservar a proporção original de artigos muito e pouco citados em cada conjunto.

Foram testados nove algoritmos de aprendizado de máquina, abrangendo desde modelos lineares até *ensembles* e redes neurais. Dentre os modelos lineares, destacam-se a *Logistic Regression*, a *Linear SVC* e o *SGDClassifier*. Nos conjuntos desbalanceados, foi ativado o parâmetro *class_weight='balanced'* para compensar o desnível entre as classes. Já para os conjuntos balanceados, essa configuração não foi necessária. A *Logistic Regression* foi executada com *max_iter=1000* e processamento paralelo (*n_jobs=-1*). O *SGDClassifier* utilizou a função de perda *log loss* com o mesmo limite de iterações.

Entre os modelos baseados em árvores de decisão, foram utilizados o *Random Forest*, o *Gradient Boosting*, o *XGBoost* e o *LightGBM*. O *Random Forest* foi configurado com 200 árvores e execução paralela habilitada. O *XGBoost* foi ajustado com a métrica *logloss*, desabilitando o codificador automático de rótulo (*use_label_encoder=False*) e com paralelismo (*n_jobs=-1*). O *LightGBM* também utilizou múltiplos núcleos.

Modelos probabilísticos, como o *Multinomial Naive Bayes*, também foram incluídos por sua simplicidade e eficiência em tarefas de classificação de texto. Por fim, foi testado um modelo de rede neural do tipo *Multilayer Perceptron (MLPClassifier)*, configurado com uma única camada oculta contendo 100 neurônios e *max_iter=300*.

Para todos os experimentos, os textos foram transformados utilizando a técnica de vetorização TF-IDF, com um limite de 10.000 *features* e *n-grams* de 1 a 2 palavras (*ngram_range=(1, 2)*). Essa abordagem permitiu capturar tanto termos isolados quanto combinações frequentes de palavras.

Ao final do treinamento, os modelos foram avaliados com base nas previsões realizadas sobre o conjunto de teste. Para isso, utilizaram-se as probabilidades (*predict_proba*) ou, quando não disponíveis, os *scores* de decisão, devidamente normalizados. Com esses valores, foram calculadas as métricas descritas anteriormente, bem como geradas curvas ROC e matrizes de confusão. Os experimentos foram realizados em uma máquina com sistema operacional *Linux* 6.8.0-59-generic, arquitetura *x86_64*, e processador compatível com essa arquitetura. A máquina dispõe de 125,62 GB de memória *RAM* e uma *GPU NVIDIA GeForce RTX 4090* com 23,54 GB de memória dedicada.

Tabela 1 – Principais hiperparâmetros utilizados nos modelos avaliados

Modelo	**Hiperparâmetros principais**
Logistic Regression	`max_iter=1000`, `class_weight='balanced'`*
Linear SVC	`class_weight='balanced'`*
SGDClassifier	`loss='log_loss'`, `max_iter=1000`, `class_weight='balanced'`*
Multinomial Naive Bayes	Parâmetros padrão
Random Forest	`n_estimators=200`, `random_state=42`
Gradient Boosting	Parâmetros padrão
XGBoost	`use_label_encoder=False`, `eval_metric='logloss'`
LightGBM	Parâmetros padrão
MLPClassifier	`hidden_layer_sizes=(100,)`, `max_iter=300`

*O hiperparâmetro `class_weight='balanced'` foi utilizado apenas nos conjuntos de dados desbalanceados.

5.3 Métricas de Avaliação

Para avaliar o desempenho dos modelos de classificação utilizados neste trabalho, foram adotadas diversas métricas amplamente utilizadas na literatura, especialmente em tarefas de classificação binária (KOYEJO *et al.*, 2014). A seguir, são descritas as métricas empregadas, bem como seus respectivos cálculos.

5.3.1 Acurácia

A acurácia mede a proporção de previsões corretas em relação ao total de previsões realizadas. É definida como:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

onde:

- TP = verdadeiros positivos,
- TN = verdadeiros negativos,
- FP = falsos positivos,
- FN = falsos negativos.

5.3.2 Precisão

A precisão indica a proporção de exemplos classificados como positivos que são realmente positivos. É útil quando o custo de um falso positivo é alto.

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{5.2}$$

5.3.3 Recall

Também chamada de sensibilidade ou taxa de verdadeiros positivos, o *recall* mede a proporção de exemplos positivos corretamente identificados pelo modelo.

$$\textit{Recall} = \frac{TP}{TP + FN} \tag{5.3}$$

5.3.4 F1-Score

O *F1-score* é a média harmônica entre precisão e *recall*, sendo especialmente útil quando se busca um equilíbrio entre ambas:

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \textit{Recall}}{\text{Precisão} + \textit{Recall}} \tag{5.4}$$

5.3.5 Área sob a Curva ROC (AUC)

A métrica *Area Under the Curve* (AUC) refere-se à área sob a curva ROC, que representa a capacidade do modelo em distinguir entre as classes positiva e negativa.

Valores de AUC variam de 0 a 1, onde:

- AUC = 0,5 indica um modelo com desempenho equivalente ao acaso;
- AUC entre 0,6 e 0,7 indica desempenho modesto, mas melhor que aleatório;
- AUC acima de 0,7 pode ser considerado um desempenho razoável;
- AUC = 1,0 representa separação perfeita entre as classes.

5.3.6 Curva ROC

A curva ROC é um gráfico que representa a relação entre a taxa de verdadeiros positivos (TPR) e a taxa de falsos positivos (FPR), definidos como:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN} \tag{5.5}$$

O gráfico é construído variando-se o limiar de decisão e plotando-se os valores de TPR versus FPR. Quanto mais próxima a curva estiver do canto superior esquerdo, melhor o desempenho do modelo.

5.3.7 Matriz de Confusão

A matriz de confusão é uma ferramenta visual para avaliar o desempenho da classificação. Ela mostra a distribuição dos acertos e erros do modelo, permitindo identificar padrões de erro. A estrutura típica é:

	Previsto Positivo	**Previsto Negativo**
Real Positivo	TP	FN
Real Negativo	FP	TN

Essa matriz permite analisar quais tipos de erros são mais comuns e ajustar os modelos conforme a sensibilidade desejada.

5.4 Resultados

As Tabelas 2 e 3 apresentam o desempenho dos modelos nos conjuntos balanceados e desbalanceados, respectivamente, considerando duas abordagens distintas: uma baseada exclusivamente no título do artigo e outra que combina o título com metadados adicionais.

Na Tabela 2, os modelos treinados apenas com o título apresentam AUC em torno de 0,66, com destaque para o *SGDClassifier* e a *Logistic Regression*. A inclusão de metadados proporcionou uma leve melhoria nas métricas, elevando o AUC máximo para cerca de 0,68, observado na *Logistic Regression*.

Já na Tabela 3, os resultados exibem maior variabilidade entre os modelos. No conjunto limitado ao título, os melhores desempenhos foram alcançados pelo *Multinomial Naive Bayes* e pelo *Random Forest*, ambos com *Acurácia* próxima a 0,72 e AUC superiores a 0,65. Na abordagem que combinou título com metadados, o *Random Forest* e o *LightGBM* obtiveram *Acurácia* ao redor de 0,72 e AUC próximos a 0,67.

De forma geral, as métricas indicam que, independentemente da configuração do conjunto de dados (balanceado ou desbalanceado), os modelos treinados exclusivamente com o título do artigo apresentaram desempenho comparável aos que incorporaram metadados adicionais. Essa proximidade sugere que, no contexto analisado, o título do artigo carrega informações preditivas suficientemente relevantes para a tarefa, de modo que a adição de metadados como número de autores ou ano de publicação não contribuiu de forma expressiva para o aumento da capacidade preditiva dos modelos avaliados.

5.5 Discussão

Os resultados obtidos nas Tabelas 2 e 3 revelam padrões importantes sobre o desempenho dos modelos de aprendizado supervisionado aplicados à tarefa de previsão de artigos altamente citados, a partir de diferentes níveis de informação textual e estruturada.

5.5.1 Principais Achados

Diferentemente de trabalhos anteriores, como os de Hirako, Sasano e Takeda (2024), que utilizaram o texto completo dos artigos, e Suzen *et al.* (2021) e Baba, Baba e Ikeda (2019), que se basearam principalmente em resumos e metadados estruturados, o presente estudo foca exclusivamente no potencial preditivo dos títulos. Essa abordagem minimalista representa uma contribuição inédita à literatura, ao demonstrar empiricamente que, mesmo com entradas extremamente reduzidas, é possível obter resultados competitivos na tarefa de previsão de citações. Os resultados evidenciam que o título, por ser uma síntese concisa e cuidadosamente elaborada do conteúdo do artigo, carrega sinais suficientes para prever o seu impacto futuro, o que não havia sido diretamente investigado nos trabalhos anteriores.

Tabela 2 – Desempenho dos modelos nos conjuntos *balanceados*, utilizando apenas título e título com metadados.

Conjunto	Modelo	Acurácia	Precisão	Recall	F1-Score	AUC
Apenas Título - Balanceado						
	SGDClassifier	**0.6184**	**0.6203**	0.6105	0.6154	**0.6618**
	Logistic Regression	**0.6184**	0.6182	0.6196	0.6189	0.6609
	Random Forest	0.6129	0.6189	0.5877	0.6029	0.6500
	MultinomialNB	0.6110	0.6148	0.5943	0.6044	0.6486
	LightGBM	0.6028	0.6011	0.6117	0.6063	0.6415
	Linear SVC	0.5914	0.5625	**0.8228**	**0.6682**	0.6386
	MLPClassifier	0.5966	0.5969	0.5952	0.5960	0.6349
	XGBoost	0.5879	0.5852	0.6042	0.5945	0.6291
	Gradient Boosting	0.5796	0.5785	0.5861	0.5823	0.6163
Título com Metadados - Balanceado						
	Logistic Regression	**0.6319**	0.6298	0.6399	0.6348	**0.6815**
	SGDClassifier	0.6291	**0.6304**	0.6239	0.6271	0.6794
	LightGBM	0.6126	0.6058	0.6447	0.6246	0.6624
	Random Forest	0.6172	0.6138	0.6320	0.6228	0.6658
	MultinomialNB	0.6223	0.6276	0.6015	0.6143	0.6625
	Linear SVC	0.5897	0.5557	**0.8951**	**0.6857**	0.6599
	MLPClassifier	0.6043	0.6112	0.5732	0.5916	0.6573
	XGBoost	0.6044	0.5978	0.6381	0.6173	0.6503
	Gradient Boosting	0.5837	0.5684	0.6960	0.6257	0.6333

Além disso, este estudo contribui ao explorar comparativamente diferentes configurações (balanceadas e desbalanceadas) e algoritmos, oferecendo uma visão abrangente sobre a robustez da hipótese proposta.

Os resultados obtidos reforçam a hipótese central deste trabalho: é possível prever, com desempenho competitivo, se um artigo será altamente citado utilizando apenas o seu título. Esta suposição orientou o desenho experimental e foi validada principalmente pelos resultados obtidos com os modelos treinados exclusivamente com títulos.

No cenário com dados balanceados (Tabela 2), os modelos que utilizaram apenas o título do artigo apresentaram desempenhos sólidos. O *SGDClassifier* e a *Logistic Regression* se destacaram com **acurácia de 0.6184** e AUC próximos a 0.66, evidenciando que o título, por si só, contém informações preditivas relevantes para inferir o impacto futuro da publicação.

Ao enriquecer os dados com metadados, como número de autores, ano de publicação e presença de pontuação no título, houve uma melhora modesta nos resultados. A melhor performance foi da *Logistic Regression*, com **acurácia de 0.6319**, **F1-Score de 0.6348** e AUC de 0.6815. Esses ganhos, embora presentes, não foram significativamente superiores aos modelos baseados somente no título, o que reforça a viabilidade da hipótese de que o

Tabela 3 – Desempenho dos modelos nos conjuntos *desbalanceados*, utilizando apenas título e título com metadados.

Conjunto	Modelo	Acurácia	Precisão	Recall	F1-Score	AUC
Apenas Título - Desbalanceado						
	Random Forest	**0.7202**	0.5112	0.0976	0.1640	0.6506
	Gradient Boosting	0.7196	**0.5781**	0.0084	0.0165	0.6186
	MultinomialNB	0.7183	0.4959	0.1517	0.2323	0.6556
	LightGBM	0.7190	0.5014	0.0396	0.0733	0.6493
	XGBoost	0.7189	0.4990	0.0554	0.0997	0.6337
	MLPClassifier	0.6680	0.4003	0.3642	0.3814	0.6331
	SGDClassifier	0.6295	0.3946	0.5961	**0.4749**	**0.6659**
	Logistic Regression	0.6273	0.3921	0.5925	0.4719	0.6624
	Linear SVC	0.4019	0.3100	**0.9204**	0.4638	0.6403
Título com Metadados - Desbalanceado						
	XGBoost	**0.7235**	0.5583	0.0769	0.1351	0.6580
	Random Forest	0.7233	0.5511	0.0830	0.1442	0.6729
	LightGBM	0.7232	**0.5688**	0.0626	0.1128	0.6713
	MultinomialNB	0.7220	0.5183	0.1508	0.2336	0.6710
	Gradient Boosting	0.7203	0.6019	0.0140	0.0274	0.6402
	MLPClassifier	0.6938	0.4444	0.3580	0.3966	0.6665
	Logistic Regression	0.6408	0.4089	0.6243	**0.4942**	**0.6881**
	SGDClassifier	0.6344	0.4034	0.6284	0.4913	0.6869
	Linear SVC	0.4173	0.3176	**0.9342**	0.4740	0.6676

título é um forte indicativo do potencial de citação de um artigo.

Nos conjuntos desbalanceados (Tabela 3), a diferença entre os modelos foi mais acentuada. O *Random Forest* alcançou a maior **acurácia (0.7202)** usando apenas o título, porém com **baixo recall (0.0976)**, o que é esperado em contextos de classes desiguais. O *SGDClassifier*, apesar de menor acurácia (0.6295), apresentou **o melhor equilíbrio entre precisão e recall**, com **F1-Score de 0.4749** e AUC de 0.6659.

Por fim, ao considerar os dados desbalanceados com metadados, o melhor resultado em **acurácia (0.7235)** foi alcançado pelo *XGBoost*. Contudo, mais uma vez, a *Logistic Regression* obteve o maior **F1-Score (0.4942)** e AUC (0.6881), demonstrando robustez e consistência, mesmo em contextos com classes desproporcionais.

5.5.2 Limitações

Apesar dos resultados promissores, algumas limitações devem ser consideradas. Primeiramente, o uso de metadados básicos pode não ser suficiente para capturar aspectos mais contextuais do impacto de uma publicação, como área do conhecimento, tipo de periódico ou engajamento em redes acadêmicas. Além disso, os modelos testados são essencialmente clássicos e lineares, limitando sua capacidade de capturar interações complexas

entre variáveis textuais e estruturadas.

Outro ponto relevante é o impacto do desbalanceamento nos resultados: mesmo com boas acurácias, muitos modelos apresentaram baixos valores de *recall*, o que é particularmente crítico em cenários onde a identificação correta de artigos altamente citados (minoria) é prioritária.

5.5.3 Trabalhos Futuros

Para trabalhos futuros, propõe-se a adoção de modelos baseados em aprendizado profundo, como redes neurais recorrentes, *transformers* (por exemplo, BERT, SciBERT) e, especialmente, *Large Language Models* (LLMs), como os oferecidos pela biblioteca *Unsloth*, que permitem o ajuste fino eficiente de modelos modernos mesmo em ambientes com recursos computacionais limitados. Tais modelos são promissores para tarefas de classificação textual, dada sua capacidade de capturar nuances semânticas com alta precisão.

Outra direção relevante é o uso de métodos de aprendizado semi-supervisionado, que têm se mostrado eficazes na melhoria das métricas em contextos de dados rotulados limitados. Trabalhos como o de (AZEVEDO; OLIVEIRA; TELES, 2024) demonstram que o uso do BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) com técnicas semi-supervisionadas pode aumentar significativamente o desempenho em tarefas de classificação textual, incluindo domínios sensíveis como o de saúde mental.

Além disso, futuras abordagens podem investigar novas formas de engenharia de atributos, como a inclusão de variáveis contextuais (por exemplo, fator de impacto da revista, número de colaborações institucionais) e o uso de estratégias mais sofisticadas de balanceamento de dados, como *SMOTE*, *focal loss* e subamostragem informada.

Recomenda-se aprofundar os estudos sobre interpretabilidade dos modelos, por meio de técnicas como *LIME*, *SHAP* ou geração de explicações via LLMs, com o objetivo de fornecer maior transparência sobre os critérios utilizados para prever o impacto de publicações científicas. Esse aspecto é especialmente importante em contextos de avaliação editorial, financiamento de pesquisa e políticas públicas baseadas em evidência.

6 Considerações Finais

Este trabalho teve como objetivo principal desenvolver e avaliar modelos preditivos capazes de classificar artigos científicos de acordo com seu potencial de citação a curto prazo, utilizando como principal fonte de informação os títulos dos artigos.

A motivação para esta pesquisa surgiu da crescente demanda por métodos que permitam antecipar o impacto de publicações científicas, dado que as citações, tradicionalmente utilizadas como métrica de relevância, só se acumulam com o tempo. Ao explorar o potencial informativo de elementos disponíveis no momento da submissão, buscamos contribuir com estratégias mais rápidas e eficientes para a identificação de artigos promissores.

Os resultados obtidos demonstraram que é possível, com níveis razoáveis de acurácia, prever se um artigo estará entre os mais citados utilizando apenas o seu título. Esse achado corrobora a hipótese central deste trabalho, de que informações presentes exclusivamente no título, mesmo sem acesso ao conteúdo completo do artigo, já carregam elementos suficientes para que modelos preditivos identifiquem padrões associados a um maior potencial de citação.

Além disso, o estudo mostrou que, embora o uso exclusivo do título já forneça sinais relevantes, a inclusão de metadados adicionais, como o ano de publicação, autoria e presença de pontuação no título, pode contribuir para melhorar a capacidade preditiva dos modelos, especialmente em contextos com desequilíbrio entre classes. Isso sugere que, apesar da simplicidade do dado textual principal, há valor agregado em considerar características complementares para refinar as previsões.

Limitações e Trabalhos Futuros

Entre as principais limitações deste trabalho, destaca-se a restrição ao uso de informações textuais básicas, sem considerar o conteúdo completo dos artigos, resumos ou redes de citação. Além disso, a análise foi baseada em um recorte específico de dados, o que pode limitar a generalização dos resultados para outras áreas do conhecimento ou bases bibliográficas.

Como direções futuras, propõe-se:

- A incorporação de dados mais ricos, como resumos, palavras-chave e redes de coautoria, a fim de melhorar o desempenho dos modelos;

- A experimentação com técnicas mais avançadas de representação textual, como embeddings gerados por modelos de linguagem pré-treinados;
- A análise do impacto de diferentes estilos de escrita de títulos sobre a atratividade e a disseminação dos artigos;
- A aplicação da metodologia proposta em diferentes domínios científicos, para avaliar sua robustez e capacidade de generalização.

Por fim, espera-se que os achados aqui apresentados possam contribuir para um melhor entendimento sobre os fatores que influenciam o impacto de publicações científicas e inspirem novas pesquisas voltadas à ciência de dados aplicada à bibliometria.

Referências

AGARWAL, A. *et al.* Bibliometrics: tracking research impact by selecting the appropriate metrics. *Asian journal of andrology*, Medknow, v. 18, n. 2, p. 296–309, 2016. Citado na página 16.

AYDIN, Z. E.; OZTURK, Z. K. Performance analysis of xgboost classifier with missing data. *Manchester Journal of Artificial Intelligence and Applied Sciences (MJAIAS)*, v. 2, n. 02, p. 2021, 2021. Citado na página 25.

AZEVEDO, J. P. C.; OLIVEIRA, A. C. de; TELES, A. S. Identificação de ideação suicida em textos usando aprendizado semi-supervisionado. *Journal of Health Informatics*, v. 16, n. Especial, nov. 2024. Citado na página 45.

BABA, T.; BABA, K.; IKEDA, D. Citation count prediction using abstracts. *Journal of Web Engineering*, River Publishers, v. 18, n. 1-3, p. 207–228, 2019. Citado 2 vezes nas páginas 32 e 42.

BENTÉJAC, C.; CSÖRGŐ, A.; MARTÍNEZ-MUÑOZ, G. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, v. 54, n. 3, p. 1937–1967, Mar 2021. ISSN 1573-7462. Citado na página 26.

BORNMANN, L. Is collaboration among scientists related to the citation impact of papers because their quality increases with collaboration? an analysis based on data from f1000prime and normalized citation scores. *Journal of the Association for Information Science and Technology*, v. 68, n. 4, p. 1036–1047, 2017. Citado 2 vezes nas páginas 15 e 19.

BORNMANN, L.; DANIEL, H.-D. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, Emerald Group Publishing Limited, v. 64, n. 1, p. 45–80, 2008. Citado na página 16.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 23.

CAMBRIA, E. *et al.* New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, v. 28, n. 2, p. 15–21, 2013. Citado na página 20.

de Menezes, F. S. *et al.* Data classification with binary response through the boosting algorithm and logistic regression. *Expert Systems with Applications*, v. 69, p. 62–73, 2017. ISSN 0957-4174. Citado na página 21.

FU, L.; ALIFERIS, C. Models for predicting and explaining citation count of biomedical articles. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, v. 6, p. 222–6, 02 2008. Citado na página 15.

GARFIELD, E. Citation indexes for science: a new dimension in documentation through association of ideas. *Science*, American Association for the Advancement of Science, v. 122, n. 3159, p. 108–111, 1955. Citado 2 vezes nas páginas 19 e 31.

GOODWIN, J. Citation indexing—its theory and application in science, technology, and humanities by eugene garfield. *Technology and Culture*, Johns Hopkins University Press, v. 21, n. 4, p. 714–715, 1980. Citado na página 15.

HARIBABU, S. *et al.* Prediction of flood by rainf all using mlp classifier of neural network model. In: *2021 6th International Conference on Communication and Electronics Systems (ICCES)*. [S.l.: s.n.], 2021. p. 1360–1365. Citado na página 29.

HIRAKO, J.; SASANO, R.; TAKEDA, K. *CiMaTe: Citation Count Prediction Effectively Leveraging the Main Text*. 2024. Citado 2 vezes nas páginas 32 e 42.

JR., A. V. *et al. Predicting citation impact of research papers using GPT and other text embeddings*. 2024. Citado na página 31.

KABIR, F. *et al.* Bangla text document categorization using stochastic gradient descent (sgd) classifier. In: *2015 International Conference on Cognitive Computing and Information Processing(CCIP)*. [S.l.: s.n.], 2015. p. 1–4. Citado na página 22.

KE, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 24.

KOYEJO, O. *et al.* Consistent binary classification with generalized performance metrics. In: GHAHRAMANI, Z. *et al.* (Ed.). *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2014. v. 27. Citado na página 40.

NASTESKI, V. An overview of the supervised machine learning methods. *Horizons. b*, v. 4, n. 51-62, p. 56, 2017. Citado na página 20.

NEWMAN, M. E. J. Prediction of highly cited papers. *EPL (Europhysics Letters)*, IOP Publishing, v. 105, n. 2, p. 28002, jan. 2014. ISSN 1286-4854. Citado na página 31.

PAIVA, C. E.; LIMA, J. P. da S. N.; PAIVA, B. S. R. Articles with short titles describing the results are cited more often. *Clinics*, v. 67, n. 5, p. 509–513, 2012. ISSN 1807-5932. Citado na página 20.

PAL, M. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, Taylor & Francis, v. 26, n. 1, p. 217–222, 2005. Citado na página 23.

ROSSI, M. J.; BRAND, J. C. Journal article titles impact their citation rates. *Arthroscopy: The Journal of Arthroscopic Related Surgery*, v. 36, n. 7, p. 2025–2029, 2020. ISSN 0749-8063. Citado na página 20.

SADEQI-ARANI, Z.; KADKHODAIE, A. A bibliometric analysis of the application of machine learning methods in the petroleum industry. *Results in Engineering*, v. 20, p. 101518, 2023. ISSN 2590-1230. Citado na página 15.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). *Intelligent Systems*. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8. Citado na página 45.

STEGEHUIS, C.; LITVAK, N.; WALTMAN, L. Predicting the long-term citation impact of recent publications. *Journal of informetrics*, Elsevier, v. 9, n. 3, p. 642–657, 2015. Citado na página 15.

STEGEHUIS, C.; LITVAK, N.; WALTMAN, L. Predicting the long-term citation impact of recent publications. *Journal of Informetrics*, v. 9, n. 3, p. 642–657, 2015. ISSN 1751-1577. Citado na página 31.

SUZEN, N. *et al. Semantic Analysis for Automated Evaluation of the Potential Impact of Research Articles*. 2021. Citado 2 vezes nas páginas 32 e 42.

TAHAMTAN, I.; AFSHAR, A. S.; AHAMDZADEH, K. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, Springer-Verlag, Berlin, Heidelberg, v. 107, n. 3, p. 1195–1225, jun. 2016. ISSN 0138-9130. Citado 2 vezes nas páginas 16 e 19.

TAHAMTAN, I.; BORNMANN, L. What do citation counts measure? an updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics*, Springer, v. 121, n. 3, p. 1635–1684, 2019. Citado 2 vezes nas páginas 19 e 31.

WANG, D.; SONG, C.; BARABÁSI, A.-L. Quantifying long-term scientific impact. *Science*, American Association for the Advancement of Science, v. 342, n. 6154, p. 127–132, 2013. Citado na página 15.

XU, S.; LI, Y.; WANG, Z. Bayesian multinomial naïve bayes classifier to text classification. In: SPRINGER. *International Conference on Multimedia and Ubiquitous Engineering*. [S.l.], 2017. p. 347–352. Citado na página 27.

YAN, E.; DING, Y. Discovering author impact: A pagerank perspective. *Information Processing Management*, v. 47, n. 1, p. 125–134, 2011. ISSN 0306-4573. Citado na página 20.

YANG, Y.; LI, J.; YANG, Y. The research of the fast svm classifier method. In: *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. [S.l.: s.n.], 2015. p. 121–124. Citado na página 28.

YUAN, S. *et al. Modeling and Predicting Citation Count via Recurrent Neural Network with Long Short-Term Memory*. 2022. Citado na página 31.

ZHANG, C. *et al.* A survey of citation recommendation and prediction. In: *IJCAI*. [S.l.: s.n.], 2020. p. 5197–5203. Citado na página 16.

ZHANG, F.; WU, S. Predicting future influence of papers, researchers, and venues in a dynamic academic network. *Journal of Informetrics*, v. 14, n. 2, p. 101035, 2020. ISSN 1751-1577. Citado na página 15.