UNIVERSIDADE FEDERAL DO MARANHÃO

Curso de Ciência da Computação

Inez Cavalcanti Dantas

A Janela de Tempo Ideal: Otimizando Indicadores Bibliométricos para Prever o Sucesso de Curto Prazo de Pesquisadores

São Luís - MA

2025

Inez Cavalcanti Dantas

A Janela de Tempo Ideal: Otimizando Indicadores Bibliométricos para Prever o Sucesso de Curto Prazo de Pesquisadores

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

Curso de Ciência da Computação

Universidade Federal do Maranhão

Orientador: Prof. Dr. Luciano Reis Coutinho

São Luís - MA

2025

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

Cavalcanti Dantas, Inêz.
A Janela de Tempo Ideal: Otimizando Indicadores Bibliométricos para Prever o Sucesso de Curto Prazo de Pesquisadores / Inêz Cavalcanti Dantas. - 2025.
54 f.

Coorientador(a) 1: Antônio de Abreu Batista Júnior.
Orientador(a): Luciano Reis Coutinho.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, São Luis, 2025.

1. Aprendizado de Máquina. 2. Perceptron Multicamadas. 3. Cientometria, Cientométricas. 4. Redes Neurais. 5. Predição de Sucesso Científico. I. Abreu Batista Júnior, Antônio de. II. Reis Coutinho, Luciano. III. Título.

Inez Cavalcanti Dantas

A Janela de Tempo Ideal: Otimizando Indicadores Bibliométricos para Prever o Sucesso de Curto Prazo de Pesquisadores

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

Trabalho aprovado em 30 de julho de 2024.

Prof. Dr. Luciano Reis Coutinho
Orientador
Universidade Federal do Maranhão

Prof. Dr. Antônio de Abreu Batista-Jr
Coorientador
Universidade Federal do Maranhão

Prof. Dr. Francisco José da Silva e Silva
Examinador Interno
Universidade Federal do Maranhão

Prof. Dr. Ricardo de Andrade Lira Rabelo
Examinador Externo
Universidade Federal do Piauí

São Luís - MA
2025

Agradecimentos

Agradeço à Deus, por tudo em minha vida. Pela saúde e proteção, por me capacitar e por guiar meu passos nos caminhos da vida.

Agradeço em especial aos meus pais, Ulisses Cavalcanti da Silva e Eurides Cavalcanti da Silva, à quem devo tudo em minha vida. Agradeço pelo dom da vida, pela educação, pelos exemplos, pelos conselhos, por toda a dedicação e por todos os esforços feitos para garantir a minha educação e por todo suporte dado em toda minha vida.

Agradeço também a toda minha família, por todo apoio dado em momentos que precisei. Em especial aos meus filhos Aída, Aline e Amanda por toda ajuda e auxílio dados em minha formação acadêmica. Agradeço também ao meu esposo Arlindo Dantas Júnior pelo companheirismo.

Agradeço aos professores Luciano Reis Coutinho e Antônio de Abreu Batista Júnior da Universidade Federal do Maranhão, por todo suporte dado durante o período do mestrado. Agradeço por fim aos meus amigos, pela ajuda dada e pelas experiências compartilhadas durante a jornada, em especial a Bruno Moraes, Carlos Eduardo Portela, Magno e a Maria do Rosário. A todos os professores e funcionários do DEINF, por todos os ensinamentos e experiências compartilhadas.

"A escultura já está completa dentro do bloco de mármore, antes de eu começar meu trabalho. Já está lá, só tenho que esmiuçar o material supérfluo."

(Michelangelo)

Resumo

Prever o sucesso de um pesquisador por meio da avaliação do sucesso de suas publicações é um tópico importante que vem atraindo cada vez mais atenção na comunidade científica. No entanto, nesse contexto, não existem estudos abrangentes que forneçam o tamanho ideal da amplitude da janela para a captura de publicações como base para o cálculo de preditores. Neste estudo, estamos interessados em como janelas temporais curtas ou longas usadas como base para o cálculo de índices bibliométricos preditivos afetam a precisão do classificador que categoriza os pesquisadores entre aqueles com publicações recentes bem-sucedidas e outros pesquisadores. Usando o conjunto de dados da American Physical Society (APS), comparamos o desempenho dos classificadores. Usamos validação cruzada de 10 *folds* para determinar o classificador mais preciso em ambos os cenários descritos acima. Nós encontramos que uma avaliação mais abrangente dos cientistas é necessária sugerindo que uma janela longa supera uma janela curta.

Palavras-chave: : Aprendizado de máquina. Perceptron multicamadas. Redes neurais. Predição de sucesso. Cientometria. Cientométricas. Previsão de sucesso científico.

Abstract

Predicting the success of a researcher by evaluating the success of their publications is an important topic that is attracting more and more attention in the scientific community. However, in this context, there are no comprehensive studies that provide the ideal window size for capturing publications as a basis for calculating predictors. In this study, we are interested in how short or long time windows used as the basis for calculating predictive bibliometric indices affect the accuracy of the classifier that categorizes researchers into those with successful recent publications and other researchers. Using the American Physical Society (APS) dataset, we compare the performance of the classifiers. We use 10-fold cross-validation to determine the most accurate classifier in both scenarios described above. We found that a more comprehensive evaluation of scientists is necessary, suggesting that a long window is superior to a short window.

Keywords: Machine learning. Multilayer perceptron. Neural networks. Success prediction. Scientometrics. Scientometrics. Scientific success prediction.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

APS	*American Physical Society*
CNNs	*Redes Neurais Convolucionais*
CNPq	*Conselho Nacional de Desenvolvimento Científico e Tecnológico*
FAPESP	*Fundação de Amparo à Pesquisa do Estado de São Paulo*
F-SCORE	*Métrica de Avaliação de Modelos de Classificação*
FFNNs - FNNs	*Redes Neurais Feedforward*
FN	*False Negatives - Falsos Negativos*
FP	*False Positive - Falso Positivo*
IA	*Inteligência Artificial*
ISI	*Institute for Scientific Information*
LDA	*Latent Dirichlet Allocation*
LIME	*Local Interpretable Model-Agnostic Explanations*
LM	*Aprendizagem de Máquina - machine learning*
MCC	*Matthews correlation coefficient - Coeficiente de Correlação de Matthews*
MLPs	*Propagação Direta - Perceptrons Multicamadas*
NIH	*National Institutes of Health - Institutos Nacionais de Saúde*
NSF	*National Science Fecundativo - Fundação Nacional de Ciências*
PNL	*Processamento de Linguagem Natural*
RN	*Redes Neurais*
RNAs	*Redes Neurais Artificais*
ReLU	*Unidade Linear Retificada - Rectified Linear Unit*
RNNS	*Redes Neurais Recorrentes*
TN	*True Negative - Verdadeiro Negativo*
TP	*True Positives - Verdadeiros Positivos*

Sumário

1 Introdução

Os indicadores bibliométricos (número de citações, índice h) têm sido amplamente utilizados por governos, agências governamentais e outros atores (por exemplo, universidades e os próprios cientistas) para medir o desempenho pesquisadores, com o objetivo de orientar as decisões de aprovação na fase de teste de novos membros do corpo docente de programas de pós-graduação e como critério para a progressão profissional, aprovação de financiamento para investigação, seleção de membros de conselhos editoriais, entre outras aplicações. A racionalidade subjacente ao uso desses indicadores como ferramentas de apoio à tomada de decisões, nesses contextos, é a força preditiva que lhes é atribuída (BATISTA-JR; GOUVEIA; MENA-CHALCO, 2021), para essas e outras aplicações, o potencial de impacto futuro da pessoa avaliada é a preocupação central. O impacto científico é crucial para avaliar publicações, acadêmicos e instituições.

Embora a avaliação geralmente se baseie no desempenho passado, é mais relevante prever a influência futura de entidades acadêmicas. A previsão do impacto científico é importante para a recomendação de recursos, ajudando pesquisadores a encontrar artigos rapidamente, e para identificar novos talentos, facilitando recomendações de especialistas e colaborações. Além disso, previsões precisas fornecem dados importantes para a gestão de pesquisadores e instituições, auxiliando em decisões como contratação, promoção, financiamento de projetos e candidatura a prêmios. Prever o impacto de futuros cientistas é um desafio crescente, tradicionalmente abordado pela análise de preditores como metadados do artigo, métricas de autores e reputação do periódico (BAI; ZHANG; LEE, 2019; ACUNA; ALLESINA; KORDING, 2012; AYAZ; MASOOD, 2020).

Neste contexto, uma área negligenciada é a questão de saber se os índices bibliométricos calculados apenas com dados recentes fornecem melhores previsões do que com dados de toda a carreira do cientista. Para preencher esta lacuna, neste estudo pretendemos descobrir qual é o intervalo de tempo ideal para prever o sucesso a curto prazo de pesquisadores. A nossa hipótese é que os classificadores treinados com o primeiro intervalo de tempo obterão melhores resultados nesta tarefa do que os treinados com o segundo. Neste artigo, o sucesso dos físicos é medido pelo número de novas publicações futuras, cada uma com pelo menos o mesmo número de citações, num curto período de tempo determinado.

1.1 Motivação

Com o rápido avanço da pesquisa científica e o crescimento exponencial do número de cientistas, todos os anos são publicados inúmeros artigos.

Neste contexto, os sistemas de recomendação de artigos, que recomendam artigos aos cientistas de acordo com os seus interesses de investigação, estão a perder eficácia, uma vez que são feitas inúmeras sugestões sobre praticamente todos os temas, o que sobrecarrega o cientista. O sistema de recomendação poderia priorizar a lista de artigos sugeridos com base em estimativas do impacto esperado dos artigos publicados recentemente e dar maior prioridade àqueles com um impacto estimado mais elevado.

Diante desse cenário, a questão de como prever eficazmente a influência futura de artigos científicos e de seus autores emergiu como um problema de pesquisa crucial.

Além disso, este tópico atrai o interesse de pesquisadores de diversas áreas e possui grande relevância para otimizar a eficiência da pesquisa, além de fundamentar a tomada de decisões e a avaliação científica (XIA; LI; LI, 2023).

1.2 Hipótese inicial

Classificadores treinados com indicadores calculados a partir de uma janela de tempo longa (toda a produção do Cientista) não terão melhor desempenho na previsão do sucesso científico de curto prazo de físicos do que aqueles treinados com uma janela curta (produção recente).

1.3 Objetivo Geral e Específicos

O objetivo geral é determinar o intervalo de tempo para a coleta de publicações que sirvam de base para o cálculo dos índices preditivos do sucesso científico futuro do pesquisador, aquele que aumenta a precisão do classificador preditivo.

Objetivos Específicos:

1. Comparar o desempenho de classificadores treinados com diferentes janelas de tempo, utilizando o conjunto de dados da American Physical Society (APS).
2. Determinar se índices bibliométricos baseados em dados recentes fornecem previsões mais precisas do que o uso de dados de longo prazo.
3. Validar os resultados através de validação cruzada *k-fold* (k=10) para determinar o classificador mais preciso e o efeito de cada janela de tempo sobre o desempenho do classificador.

1.4 Relevância do estudo

Este estudo contribui significativamente para o debate atual sobre a bibliometria preditiva, destacando a importância de utilizar métricas de avaliação abrangentes. Ao identificar as condições ideais para prever o sucesso acadêmico, abrimos caminho para otimizar a alocação de recursos, o planejamento de carreira e a avaliação institucional.

Em última análise, o aprimoramento desses modelos preditivos trará mais transparência à avaliação científica, ajudará a reduzir vieses na análise de pesquisas e, por fim, promoverá um ambiente acadêmico mais imparcial.

1.5 Contribuição Principal do Estudo

Este estudo demonstrou que, ao contrário da hipótese original, o uso de um período de tempo curto para calcular os índices bibliométricos preditores leva a uma redução significativa no desempenho do classificador. Os nossos resultados demonstram que a precisão da previsão melhora quando se utiliza uma janela de tempo mais longa para recolher dados de publicação do cientista. Este resultado desafia a suposição comum de que a atualidade dos dados conduz a uma maior precisão. Demonstra que a cobertura temporal é um fator crucial para a solidez dos modelos de previsão.

1.6 Estrutura do trabalho

A Dissertação está organizada em cinco capítulos.

Além desta introdução, capítulo 1, o presente trabalho está estruturado em mais 5 capítulos.

No capítulo 2, será apresentada conceitos básicos sobre o trabalho já apresentado na introdução. O capítulo 2 apresenta os fundamentos teóricos e conceituais do estudo. Ele inicia com uma explanação sobre os fundamentos teóricos da cientometria, estabelecendo a base da área de estudo. Em seguida, As métricas tradicionais de impacto científico serão discutidas e detalhando como o impacto da produção científica é convencionalmente medido. Os métodos preditivos aplicados na cientometria, preparando o terreno para a discussão das abordagens de previsão que serão utilizadas. Uma contextualização histórica e evolutiva do campo da física no Brasil será feita, fornecendo o cenário para a aplicação das análises. No Capítulo 2 apresentará as Redes Neurais Artificiais (RNAs): Uma Visão Geral e culminando na detalhada explicação da escolha das Redes Neurais Feedforward (FFNNs) no trabalho e o funcionamento da FFNNs, fornecendo o suporte técnico para a metodologia.

O capítulo 3 abordará os estudos anteriores relevantes para o tema da pesquisa, revisitando trabalhos que já exploraram tópicos semelhantes e que servem de base para a presente investigação.

No capítulo 4 detalhará a metodologia utilizada na pesquisa, incluindo o tratamento e análise de dados, métricas utilizadas, algoritmos de previsão, avaliação e validação de modelos e a comparação de desempenho.

O capítulo 5 vamos apresentar os resultados da pesquisa. Os resultados obtidos a partir da aplicação da metodologia serão apresentados de forma clara e objetiva, utilizando gráficos, tabelas e outras representações visuais quando apropriado. A discussão irá contextualizar esses resultados em relação à revisão da literatura, comparando-os com estudos anteriores e destacando as contribuições originais deste trabalho. Serão exploradas as implicações dos achados, as limitações do estudo e as possíveis explicações para os padrões observados.

Finalmente, o capítulo 6 trará as considerações finais do trabalho. Nesse último capítulo, vamos fazer uma síntese dos principais pontos abordados, revisitando os objetivos e os resultados mais importantes que conseguimos alcançar. Também vamos destacar as contribuições relevantes do estudo, tanto na teoria quanto na prática. Além disso, serão apresentadas sugestões para futuras pesquisas, apontando possíveis caminhos e áreas que ainda podem ser exploradas, ajudando a expandir o conhecimento na área.

2 Fundamentação Teórica

2.1 Cientometria

A Ciência da Informação é um campo amplo e multidisciplinar que estuda tudo sobre a informação: como ela é criada, organizada, armazenada, recuperada e utilizada, e as tecnologias envolvidas nesse processo. Seu principal objetivo é melhorar o acesso e a aplicação do conhecimento em diversas áreas. É um campo em constante evolução que busca entender como a informação funciona e como pode ser melhor gerenciada para atender às necessidades humanas (VIEIRA; SILVA, 2023). Na Comunicação Científica estudos que empregam métodos matemáticos e estatísticos são cruciais para entender a comunicação e a atividade científica, especialmente quando o volume de dados se torna muito grande para ser processado por humanos(ARAúJO, 2006). A Bibliometria se destaca como uma ferramenta essencial para a análise quantitativa da informação. Embora frequentemente associada à avaliação de publicações científicas, a Bibliometria se ramifica em diversas subáreas veja Figura 1, cada uma com um foco específico na medição e análise de diferentes aspectos da produção e disseminação do conhecimento.

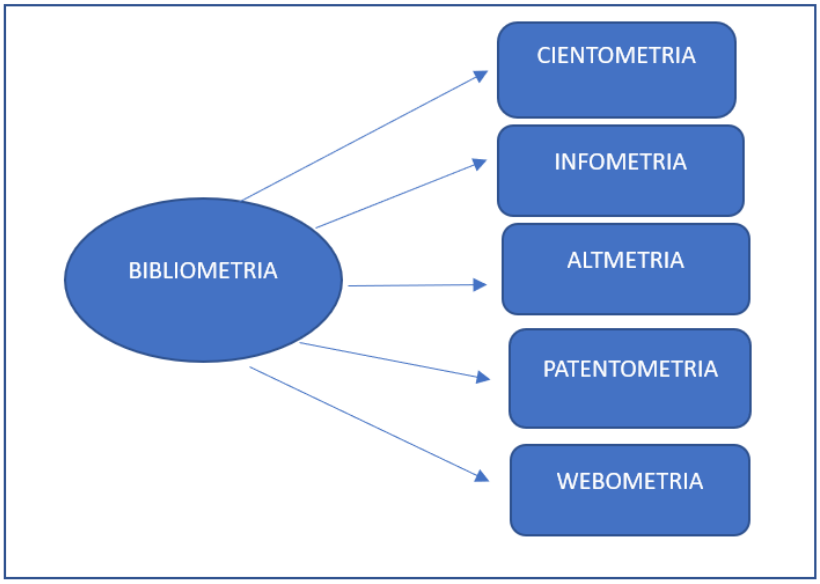


Figura 1 – Subáreas da Bibliometria,'Criada pelo autor'

Entre as Subáreas mais conhecidas estão:

- Cientometria: Voltado para a medição da ciência como um todo. A Cientometria explora aspectos da produção, disseminação e uso do conhecimento científico em

um contexto mais abrangente, buscando entender a dinâmica da pesquisa, as tendências em diferentes disciplinas e a performance de instituições e pesquisadores. A Bibliometria, focada em documentos bibliográficos, pode ser considerada uma parte integrante da Cientometria.

- Informetria: Esta subárea se dedica à "medição de sistemas de informação e modelos matemáticos"relacionados à produção, disseminação e uso de todo tipo de informação, não apenas a científica. A Informetria pode analisar, por exemplo, o fluxo de informações em redes sociais, a eficiência de bancos de dados ou o impacto de patentes na inovação tecnológica.

- Patentometria: Com um foco específico, a Patentometria se dedica à "análise de patentes". Ela explora aspectos relacionados à produção intelectual industrial e à inovação tecnológica. Ao analisar patentes, é possível entender o cenário de desenvolvimento tecnológico de um setor ou país, identificar inventores-chave e mapear tendências de pesquisa e desenvolvimento.

- Webometria: Esta vertente aplica a análise quantitativa à *World Wide Web*. A Webometria mede recursos, acessos e a utilidade da informação publicada na internet, incluindo a análise de links entre sites. É uma ferramenta importante para entender a visibilidade online de instituições, a popularidade de tópicos e o fluxo de informação digital.

- Altmetria: Uma área mais recente e em crescimento, a Altmetria estuda as ações, interações e menções de pesquisadores e não-pesquisadores relacionadas à produção acadêmica em redes e mídias sociais. Ela busca medir o impacto da pesquisa para além das citações tradicionais em periódicos científicos, avaliando o engajamento em plataformas como Twitter, blogs, Wikipédia e outros repositórios de dados.

Essas subáreas, juntamente com a Bibliometria descritiva — que avalia o panorama atual da produção científica — e a Bibliometria preditiva — voltada para a antecipação de tendências e impactos futuros — constituem um conjunto sólido de métodos para entender a complexa dinâmica de geração e disseminação do conhecimento. Ao quantificar diversos aspectos do universo informacional, essas abordagens oferecem *insights* valiosos para a pesquisa acadêmica, a gestão de informações e a formulação de decisões estratégicas. A seguir é abordado a cientometria, foco do estudo.

A cientometria é uma área de estudo que aplica-se ao estudo e análise quantitativa da ciência. A cientometria pesquisa mensurar e avaliar o progresso científico através de indicadores bibliométricos, como número de publicações, citações, fatores de impacto, entre outros (SILVA; BIANCHI, 2001).

A cientometria surgiu no início do século XX, com o objetivo de medir e avaliar a produção científica. No início, ela se concentrava principalmente na análise de publicações científicas, mas, passou a incluir também outras formas de comunicação científica, como apresentações em conferências e patentes.

Exemplos de aplicações da cientometria:

- Avaliação de desempenho: a cientometria pode ser usada para avaliar o impacto e a relevância da pesquisa científica, o que pode ajudar a identificar áreas que requerem mais financiamento ou atenção.
- Previsão de tendências: a cientometria pode ser usada para identificar padrões e tendências nos dados científicos, o que pode ajudar a prever futuras áreas de pesquisa e desenvolvimento.
- Identificação de oportunidades de colaboração: a cientometria pode ser usada para identificar pesquisadores que estão trabalhando em áreas relacionadas, o que pode ajudar a promover a colaboração e a inovação.

Na previsão de tendências a cientometria pode ter o propósitos de prever futuros trabalhos científicos. As redes neurais podem ser treinadas para identificar padrões nos dados que possam indicar possíveis áreas de pesquisa futuras. Por exemplo, uma rede neural poderia ser treinada para identificar áreas em que o número de publicações ou número de citações estão crescendo rapidamente. Essas áreas seriam então identificadas como potenciais áreas de pesquisa futura.(RUAN et al., 2020) Outro exemplo para prever futuros trabalhos científicos é o uso de aprendizado de máquina para analisar dados de financiamento de pesquisa. O aprendizado de máquina pode ser usado para identificar padrões nos dados que possam indicar quais áreas de pesquisa estão recebendo mais financiamento. Por exemplo, um modelo de aprendizado de máquina poderia ser treinado para identificar áreas em que o financiamento está aumentando rapidamente. Essas áreas seriam então identificadas como potenciais áreas de pesquisa futura. A importancia de identificar os cientistas que têm potencial de sucesso para que eles possam receber o apoio necessário para realizar suas pesquisas com confiabilidade.

A IA está sendo cada vez mais usada no campo da cientometria para automatizar tarefas, melhorar a precisão das análises e identificar novos padrões e tendências.

2.1.1 Métricas tradicionais de impacto científico

Cientometria e bibliometria são campos que permitem quantificar e compreender a produção científica. A cientometria é o estudo da ciência por meio de métodos numéricos e analíticos. Já a bibliometria atua especificamente no tratamento e na análise estatística

dos dados resultantes dessa produção, observados em diversas publicações científicas, como artigos, livros e revistas.

Os pesquisadores que investigam a "ciência da ciência"empregam uma variedade de indicadores bibliométricos. Alguns exemplos desses indicadores são: o volume de titulações acadêmicas e científicas concedidas, o número de patentes registradas por pesquisadores, a quantidade de artigos científicos publicados, o número de cientistas que publicam, a frequência de referências bibliográficas citadas em artigos, as citações que um artigo recebe, o volume de auxílios à pesquisa obtidos por cientistas, e a verba destinada às atividades de pesquisa pelas agências de fomento. As tradicionais métricas de impacto científico, a exemplo do índice h e das citações, representam ferramentas valiosas na avaliação da influência e da produtividade tanto de pesquisadores quanto de suas publicações. Contudo, é crucial reconhecer que nenhuma métrica isolada oferece uma avaliação abrangente do impacto científico. Embora úteis, essas métricas demandam interpretação cautelosa e devem ser consideradas em conjunto com outros indicadores relevantes, como a qualidade intrínseca da pesquisa, o reconhecimento da comunidade científica, a janela de tempo decorrida desde a publicação e o impacto mais amplo na sociedade (h-index, citações, etc.) (SILVA; BIANCHI, 2001).

Métricas:

- Publicações Científicas

  Elas mostram a capacidade do pesquisador de gerar novas ideias e compartilhá-las com o mundo. Como avaliamos as publicações:

  1. Quantidade de artigos: o número de trabalhos publicados é um indicador direto da produtividade. Uma produção constante de artigos geralmente é vista de forma positiva.

  2. Periódicos revisados por pares: publicar em revistas que passam por revisão por pares significa que o trabalho foi examinado e aprovado por outros especialistas da área, garantindo sua qualidade e validade.

  3. Posição de autoria: a ordem dos nomes dos autores importa. O primeiro autor geralmente é o principal responsável pela pesquisa. Já o último autor (ou autor sênior) costuma ser o líder da equipe ou o supervisor do projeto. Ambas as posições indicam liderança e uma contribuição significativa.

  4. Fator de impacto do periódico: publicar em periódicos com alto fator de impacto (uma medida da frequência com que seus artigos são citados) sugere que a pesquisa é de alta qualidade e visível. Isso mostra que o trabalho alcança um público maior e mais influente.

5. Conferências importantes: em certas áreas, como ciência da computação e engenharia, as publicações em anais de conferências de alto nível são tão valorizadas quanto, ou até mais, que os artigos em periódicos.

- Citações

  O número de citações que o trabalho de um cientista recebe é fundamental para entender a influência e o alcance de sua pesquisa na comunidade científica. Métricas principais:

  1. Número total de citações: essa é a contagem bruta de quantas vezes o trabalho de um cientista foi referenciado por outros pesquisadores. É um bom indicador da frequência de uso de sua pesquisa.

  2. Índice H (h-index): esta métrica busca equilibrar produtividade e impacto. Um cientista tem um h-index de h se h de suas publicações receberam, cada uma, pelo menos h citações. Por exemplo, se um pesquisador tem 20 artigos citados pelo menos 20 vezes cada, seu h-index é 20.

  3. Índice G (g-index): uma alternativa ao índice H, o g-index valoriza mais os artigos altamente citados, dando maior peso a eles na avaliação.

  4. Citações por artigo: calcular a média de citações por publicação pode oferecer uma visão mais clara do impacto individual de cada trabalho realizado pelo cientista.

- Concessão de bolsas e financiamento

  A capacidade de um cientista em conseguir financiamento competitivo demonstra claramente o reconhecimento de seu trabalho e a percepção de que suas propostas de pesquisa são relevantes e viáveis.

  1. Concessão de bolsas de pesquisa: indica diretamente a capacidade do pesquisador de atrair recursos para o desenvolvimento de seu trabalho.

  2. Financiamento de agências de fomento (nacionais e internacionais): receber verbas de agências respeitadas — como CNPq, FAPESP, NSF, NIH, ou Horizon Europe — funciona como um selo de aprovação tanto da comunidade científica quanto das políticas de pesquisa globais.

  3. Participação em projetos colaborativos financiados: isso sugere a competência do cientista em trabalhar em equipe e contribuir efetivamente para iniciativas de pesquisa de grande porte.

- Apresentações e convites para palestrar

  A participação ativa em conferências e eventos científicos mostra o quanto um pesquisador está engajado com sua área e sua habilidade em compartilhar conhecimento.

1. Apresentações em conferências (orais e pôsteres): isso significa que o cientista participa ativamente das discussões e da divulgação de seus resultados de pesquisa.

2. Palestras convidadas (Keynote Speeches, Seminários): ser convidado para palestrar em outras instituições ou em eventos importantes é um grande sinal de reconhecimento e liderança no seu campo de atuação.

- Supervisão de alunos e mentoria

  A capacidade de formar novos pesquisadores é uma contribuição significativa para o campo científico.

  1. Teses de doutorado e mestrado orientadas: este ponto mostra a habilidade do pesquisador em moldar e guiar a próxima geração de cientistas, transmitindo conhecimento e metodologia.

  2. Pós-doutorados supervisionados: indicia a liderança e a mentoria de cientistas que já estão em fases avançadas de suas carreiras, auxiliando-os a consolidar sua autonomia e expertise.

- Prêmios e reconhecimentos

  Prêmios, honrarias e afiliações as sociedades científicas de prestígio são reconhecimentos formais da excelência e contribuição de um cientista.

  1. Prêmios nacionais e internacionais: induzem o reconhecimento significativo por suas contribuições.

  2. Afiliação a academias de ciências: a eleição para academias de ciências (como a academia brasileira de ciências, a national academy of sciences nos EUA) é uma das maiores honras na carreira científica.

- Participação em comitês e revisão por pares

  O envolvimento em atividades de serviço à comunidade científica demonstra engajamento e reconhecimento pelos pares.

  1. Membro de comitês editoriais de periódicos: indica que o cientista é valorizado por sua expertise para moldar a direção de publicações importantes.

  2. Revisor por pares para periódicos e agências de fomento: sugere que o cientista é considerado um especialista confiável para avaliar a pesquisa de outros.

  3. Membro de comitês de conferências: contribui para a organização e seleção de trabalhos em eventos importantes.

A natureza multidisciplinar da ciência moderna e a ênfase crescente na ciência aberta e na ciência cidadã estão levando ao desenvolvimento de novas métricas e abordagens para avaliar o desempenho científico.

Segundo (SILVA; BIANCHI, 2001), indicadores de impacto - o número de publicação é um indicador meramente quantitativo, que não leva em conta a qualidade e ou a importância do conteúdo do trabalho realizado. É claro o que nem todas as publicações desperta o mesmo interesse e nem contribuem de maneira semelhante para o progresso o científico da área. Para a valorização de alguma forma a qualidade dos trabalhos publicados foram introduzidos dois outros indicadores: o número de citações que um artigo recebe na literatura e o fator de impacto da revista ou periódico em que a publicação é feita.

2.1.2 Métodos preditivos aplicados na cientometria

Os métodos preditivos representam uma ferramenta poderosa para a cientometria, permitindo uma análise mais profunda e orientada para o futuro do quadro científico. Serão apresentados exemplos de aplicações de métodos preditivos no campo da cientometria.

- Previsão de citações: modelos de regressão podem ser empregados para estimar o número futuro de citações de um artigo com base em variáveis como o número de autores, o periódico em que foi publicado, o tema abordado e as citações recebidas nos estágios iniciais. (BECKER et al., 2018) (JUNIOR, 2021)
- Detecção de tendências de pesquisa: técnicas de análise de séries temporais e modelos preditivos auxiliam na identificação do crescimento ou declínio de determinadas áreas do conhecimento ao longo do tempo.
- Previsão de colaborações científicas: modelos baseados em redes sociais permitem antecipar possíveis colaborações entre pesquisadores ou instituições, com base em padrões anteriores de cooperação.
- Avaliação de políticas científicas: abordagens causais e preditivas podem ser utilizadas para estimar os efeitos de políticas de financiamento ou outras intervenções sobre o ecossistema científico.

A incorporação de métodos preditivos na cientometria revela-se particularmente vantajosa, oferecendo perspectivas enriquecedoras para diversos atores do cenário científico. Tais métodos atuam como um forte apoio à tomada de decisões. Ao fornecerem informações prognósticas, instrumentalizam agências de fomento, instituições de pesquisa e os próprios pesquisadores com subsídios cruciais. Essa capacidade preditiva permite deliberações mais estratégicas no que concerne à alocação eficiente de recursos financeiros, à definição de prioridades investigativas e ao estabelecimento de colaborações frutíferas. Outrossim, a utilização de métodos preditivos proporciona a antecipação de oportunidades em ascensão. A habilidade de discernir tendências científicas em ascensão possibilita um investimento em domínios de pesquisa com alto potencial de impacto futuro, melhorando o desenvolvimento

científico. Por fim, a cientometria preditiva contribui significativamente para a avaliação prospectiva do impacto científico. Complementando as métricas de análise retrospectiva, esses métodos oferecem uma visão mais dinâmica e voltada para o futuro da influência da pesquisa, permitindo uma compreensão mais abrangente de seu alcance e relevância potencial.

Os principais tipos de métodos preditivos e como eles são aplicados na cientometria. (LOUPPE, 2015) (LIVINGSTONE, 2008. v. 458.) (ZHAO; FENG, 2022)

1. Métodos de regressão

Os métodos de regressão modelam a relação entre uma variável dependente (o que você quer prever) e uma ou mais variáveis independentes (o que você usa para prever).

- Prever citações: modelos de regressão conseguem estimar o número futuro de citações de um artigo, considerando fatores como o periódico de publicação, a quantidade de autores, a área de pesquisa e as citações iniciais.
- Estimar o impacto de periódicos: o fator de impacto de uma revista pode ser previsto com base em métricas como o número de artigos publicados, sua idade e a rede de colaboração dos autores.
- Analisar a produtividade de pesquisadores: prever a produção futura de artigos ou o índice H de um pesquisador com base em seu histórico de publicações, colaborações e o financiamento recebido.
- Modelar o crescimento de áreas de pesquisa: prever a evolução de determinadas áreas de pesquisa em termos de volume de publicações ou número de pesquisadores envolvidos.

2. Métodos de classificação

Métodos de classificação servem para categorizar dados em grupos ou classes que já existem.

- Classificação de artigos por área de pesquisa: automaticamente classificar novos artigos em áreas temáticas específicas com base em seus títulos, resumos e palavras-chave.
- Identificação de pesquisas de alto impacto: é possível classificar artigos ou autores como "de alto impacto"ou "de baixo impacto", usando suas características bibliométricas. isso ajuda a identificar potenciais *hot topics* ou pesquisadores de destaque.

- Detecção de plágio ou comportamento anômalo: você consegue identificar padrões em publicações que podem indicar plágio ou outras irregularidades éticas, classificando documentos como "suspeitos"ou "legítimos".
- Predição de sucesso em bolsas de pesquisa: é possível classificar propostas de pesquisa como "provável de ser financiada"ou "improvável de ser financiada", utilizando as características do projeto e dos pesquisadores como base.

3. Métodos de agrupamento *(clustering)*

Embora não prevejam diretamente um valor futuro, os métodos de agrupamento são essenciais na cientometria. Eles identificam padrões e estruturas intrínsecas nos dados, o que é um passo crucial para futuras previsões. Basicamente, eles agrupam itens semelhantes sem precisar de categorias predefinidas.

- Identificação de comunidades científicas: agrupar pesquisadores com base em seus padrões de coautoria ou co-citação para descobrir comunidades de pesquisa e redes de colaboração.
- Detecção de tópicos emergentes: agrupar artigos com base em similaridade de conteúdo para identificar novos tópicos de pesquisa ou subáreas.
- Agrupamento de periódicos: agrupar periódicos com base em similaridades temáticas ou padrões de citação.

4. Métodos de previsão de séries temporais

Estes métodos são feitos sob medida para prever valores futuros, usando como base dados históricos que seguem uma sequência (as chamadas séries temporais).

- Previsão do número de publicações: prever o número de artigos ou livros que serão publicados em uma determinada área ou por uma instituição nos próximos anos.
- Estimativa do crescimento de citações ao longo do tempo: prever a curva de citação de um artigo após sua publicação inicial.
- Projeção da evolução de palavras-chave: analisar e prever a frequência de uso de determinadas palavras-chave em publicações ao longo do tempo, indicando tendências de pesquisa.

5. Mineração de texto e processamento de linguagem natural

Embora não sejam métodos preditivos sozinhos, a PNL (Processamento de Linguagem Natural) e a mineração de texto são ferramentas cruciais. Elas abastecem os modelos

preditivos na cientometria ao extrair informações importantes de textos científicos, como resumos e artigos completos.

- Extração de entidades: identificar autores, instituições, palavras-chave e tópicos.
- Modelagem de tópicos: usar técnicas como *Latent Dirichlet Allocation* (LDA) para identificar tópicos latentes em grandes coleções de documentos. Os resultados podem então ser usados como características para modelos de classificação ou regressão.
- Análise de sentimentos: é possível avaliar o "sentimento"(positivo, negativo, neutro) em revisões por pares ou discussões científicas, mas isso é menos comum em predições diretas.

Apesar do potencial, usar métodos preditivos na cientometria vem com desafios. Primeiro, a disponibilidade e qualidade dos dados são essenciais; a precisão dos resultados depende diretamente de bancos de dados bibliográficos completos e confiáveis. Outro ponto é o viés: os modelos podem, sem querer, reforçar preconceitos dos dados históricos, mantendo ou até aumentando desigualdades (JUNIOR, 2021). Além disso, o fenômeno científico é complexo e tem muitas camadas, e nem tudo pode ser previsto por esses modelos. Por último, a interpretabilidade é um problema, especialmente com modelos mais avançados como as redes neurais profundas, que podem dificultar a compreensão do porquê de certas previsões.

Em suma, os métodos preditivos são ferramentas poderosas para a cientometria. Eles não só nos ajudam a entender o passado e o presente da ciência, mas também a antecipar seu futuro.

2.1.3 História e desenvolvimento do campo científico da física no Brasil

A física é a ciência que estuda os fenômenos naturais que ocorrem no universo. A presença da física e sua contribuição é inestimável para o desenvolvimento de toda a tecnologia moderna desde a prensa até os computadores quânticos. O conhecimento em ciência e engenharia da computação é aplicado como ferramenta para os avanços tanto em física teórica como experimental, ao mesmo tempo em que conceitos da física são aplicados à teoria da computação, criando uma relação entre as duas áreas (COTINGUIBA, 2023).

A história nos mostra que árabes, egípcios e outros desenvolveram o atual sistema de numeração, a geometria primitiva e a matemática básica. Na Tabela 1 apresentamos uma breve cronologia dos físicos. Isso vai nos ajudar a visualizar um pouco da história desta ciência. (SANTOS; GALLETTI, 2023)

Tabela 1 – Cronologia breve dos físicos

Época	**Físico / Contribuição**
525 a.C.	- Pitágoras obtém uma síntese do misticismo e da matemática, desviando-se dos mitos para os números na busca da fonte da verdade.
335 a.C.	- Aristóteles formula modelo de cosmo cujo centro é a Terra, imóvel.
295 a.C.	- Euclides publica os Elementos, codificando a geometria clássica.
240 a.C.	- Arquimedes desenvolve a mecânica clássica e física elementar.
100	- Cláudio Ptolomeu elabora complexo modelo do universo centrado na Terra que é base da astronomia por mais de 1.400 anos.
1543	- Nicolau Copérnico publica De revolutionibus, postulando um universo centrado no Sol.
1619	- Johannes Kepler demonstra que as órbitas dos planetas são elípticas e desenvolve leis do movimento planetário.
1687	- Isaac Newton as Leis de Newton. (Princípio da Inércia,Princípio Fundamental da Dinâmica e Princípio da ação e reação).
1799	- Pierre-Simon Laplace consolidou as bases matemáticas da hipótese da gravitação de Newton, desenvolveu a teoria da probabilidade e contribuiu para a criação do sistema métrico.
1824	- Christian Doppler descobre que, para um observador estacionário, emissões (luz ou som) de uma fonte em movimento parecerão ter frequência mais alta se o objeto estiver se aproximando, mas mais baixas se ele estiver se afastando – o "Desvio Doppler".
1898	- Marie e Pierre Curie identificam os elementos radioativos rádio e polônio.
1900	- Max Planck postula a teoria quântica da radiação; desenvolve a base da física quântica.
1905	- Albert Einstein publica artigos sobre a relatividade restrita e o efeito fotoelétrico.

2.2 Redes Neurais Artificiais (RNAs)

Redes neurais são sistemas computacionais inspirados no cérebro humano, isto é, imitam a forma como o cérebro humano processa informações. Elas usam algoritmos para encontrar padrões e conexões em dados, são capazes de identificar padrões ocultos e correlações em dados brutos, agrupá-los e classificá-los, aprimorando-se continuamente com novos dados, fazem previsões e tomam decisões de forma autônoma. O alto desempenho, para eles, não é um ponto de chegada, mas sim um processo constante de crescimento e

aprimoramento. Na figura 2 mostra visão geral das características e aplicações de redes neurais. (CHEN et al., 2019)

Redes Neurais Artificiais: Detalhando o Conceito e Aplicações As Redes Neurais Artificiais (RNAs), frequentemente chamadas apenas de redes neurais, representam uma classe sofisticada de sistemas e métodos computacionais. Elas são a espinha dorsal de muitas inovações no campo do aprendizado de máquina, permitindo a representação de conhecimento e a otimização de respostas em sistemas complexos (CHEN et al., 2019). A característica mais notável das redes neurais é sua capacidade de adquirir conhecimento a partir de exemplos e dados. Isso as capacita a discernir padrões e relações complexas de forma automatizada, um aprendizado que pode ser aplicado em inúmeras áreas, maximizando os resultados de sistemas intrincados e impulsionando avanços em pesquisa e desenvolvimento (HAYKIN, 2001) . No cerne da Rede Neural Artificial (RNA) está um modelo de processamento de dados inspirado diretamente no funcionamento do sistema nervoso biológico, mais especificamente no cérebro (LIVINGSTONE, 2008. v. 458.). A premissa fundamental é replicar a dinâmica cerebral para processar dados e informações, facilitando o aprendizado e a indução de conhecimento a partir de conjuntos de dados. Essa concepção é solidamente embasada na forma como o sistema neural biológico opera: bilhões de neurônios interconectados processam informações de entrada e saída. Esse processo intrincado possibilita a interpretação e o processamento eficiente de dados, de maneira notavelmente semelhante ao que ocorre no cérebro humano. (GERVEN; BOHTE, 2017)

Estrutura e aplicações das redes neurais em sua essência, as redes neurais podem ser descritas como conjuntos de unidades de processamento, denominadas neurônios, que são interligadas por sinapses artificiais (LIVINGSTONE, 2008. v. 458.). Essa estrutura permite que as Redes Neurais Artificiais (RNAs) sejam amplamente empregadas em diversas aplicações de reconhecimento de padrões. Dentre as mais notáveis, destacam-se o reconhecimento de voz, a detecção de objetos, a identificação de tumores e uma vasta gama de outras tarefas complexas. A capacidade das redes neurais de aprender e generalizar a partir de dados as torna ferramentas poderosas em campos que exigem a análise e interpretação de grandes volumes de informações. (KOVáCS, 2002)

Existem vários tipos de Redes Neurais Artificiais (RNAs), cada uma com sua própria estrutura e propósito. Cada tipo otimizada com uma arquitetura e funcionalidades especificas no campo da inteligência artificial. destacam-se: (HAYKIN, 2001)

- **Redes neurais *feedforward*:** as Redes Neurais Feedforward (FFNNs), também conhecidas como Redes Neurais de Propagação Direta ou Perceptrons Multicamadas (MLPs), são a arquitetura mais básica das redes neurais artificiais. A informação flui numa única direção, da camada de entrada para a camada de saída, sem

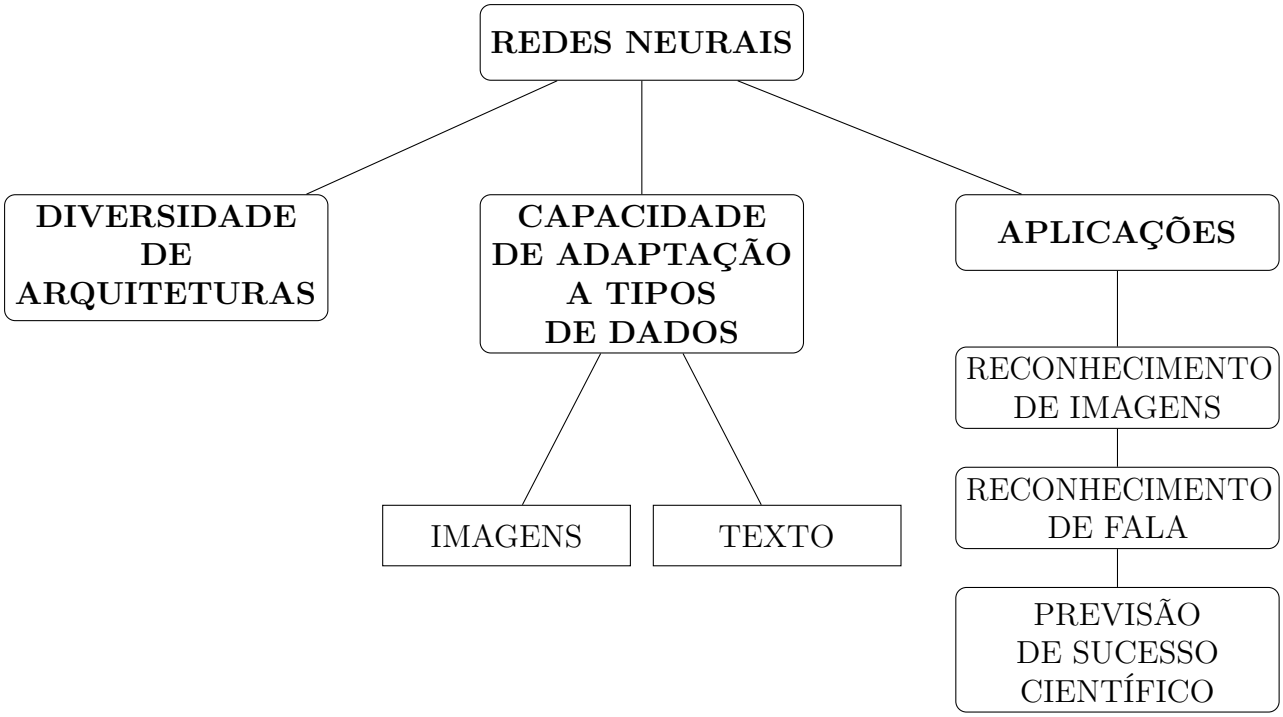


Figura 2 – Visão geral das características e aplicações de redes neurais 'Criada pelo autor'

ciclos ou loops. São utilizadas para tarefas de classificação e regressão, onde o objetivo é aprender uma relação entre um conjunto de entradas e uma saída. Para melhor entendimento a classificação é uma tarefa de aprendizado de máquina que se concentra em identificar a categoria ou classe de uma entrada específica, como uma imagem ou um texto. Para fazer isso, ela emprega algoritmos que aprendem a partir de dados de treinamento rotulados. Esse aprendizado permite que os modelos de classificação façam previsões sobre dados novos e não vistos, atribuindo-os a uma das classes predefinidas.

- **Redes neurais recorrentes (RNNs):** projetadas para lidar com sequências de dados, possuindo "memória"de entrada anteriores, ou seja, representam um tipo especial de arquitetura de rede neural artificial, criada com um propósito de lidar com dados sequenciais. Ao contrário das redes neurais tradicionais (*feedforward*), onde o fluxo de informação segue uma única direção, as RNNs se destacam por suas conexões de *feedback*. Essas conexões permitem que a rede utilize informações de etapas anteriores da sequência, realimentando-as em seus cálculos. É exatamente essa capacidade de "relembrar"entradas passadas que confere às RNNs sua característica de "memória".

- **Redes neurais convolucionais (CNNs):** especialmente eficazes para processamento de imagens e vídeos, utilizando camadas de convolução. As Redes Neurais Convolucionais (CNNs), ou ConvNets, são um tipo de rede neural artificial que revolucionou o processamento de imagens e vídeos. A sua arquitetura, que inclui camadas de

convolução que operam diferente das redes neurais comuns, que encaram uma imagem como um vetor gigante de pixels, as CNNs trabalham de outra forma. As camadas de convolução aplicam filtros *(ou kernels)* a pequenas partes da imagem. Pense nisso como um *"scanner"*que desliza pela imagem, encontrando e destacando características específicas.

- ***Autoencoders*:** usadas para aprender representações eficientes de dados (codificação), muitas vezes para redução de dimensionalidade ou geração de dados, ou seja, *autoencoders* são redes neurais artificiais criadas para aprender representações eficientes dos dados, que chamamos de codificações. Embora a meta pareça simples – reproduzir a própria entrada na saída –, o segredo está na sua arquitetura interna. Essa estrutura força o autoencoder a criar uma representação compacta e cheia de significado dos dados originais. Resumindo, autoencoders são ferramentas versáteis para extrair *insights* e estrutura dos dados de forma não supervisionada, servindo como uma base para muitas tarefas complexas em aprendizado de máquina.

2.2.1 A Escolha das Redes Neurais Feedforward (FNNs) no trabalho e o funcionamento das FNNs.

A Escolha das Redes Neurais Feedforward (FNNs) - propagação direta. A previsão de impacto é uma técnica que nos permite antecipar valores futuros com base em dados históricos. Dentro deste campo, a análise de séries temporais se destaca por sua capacidade de identificar tendências e sazonalidades nos dados, utilizando modelos para projetar previsão. Assim, a escolha das Redes Neurais Feedforward, apesar de sua simplicidade, são ferramentas extremamente relevantes e eficazes. Sua interpretabilidade e eficiência as tornam versáteis para diversas aplicações, como classificação e regressão, além de outras tarefas preditivas.

A arquitetura de uma FNNs é definida pelas suas camadas e pela forma como os neurônios estão conectados: as redes neurais feedforward processam dados em uma direção, do nó de entrada para o nó de saída. Cada nó de uma camada está conectado a todos os nós da próxima camada. Com o passar do tempo, uma rede feedforward usa o processo de feedback para aprimorar as previsões. (ZHAO; FENG, 2022) (SUNDARARAJAN; TALY; YAN, 2017) (RUAN et al., 2020)

2.2.2 Estrutura Fundamental

Na Figura 3 é apresentado um diagrama ilustrativo de uma Rede Neural Feedforward simples, com uma camada de entrada, uma camada oculta e uma camada de saída.

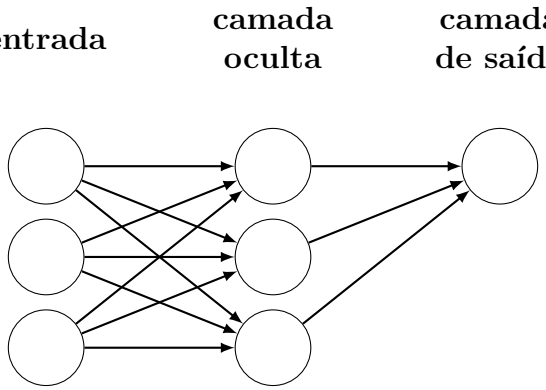


Figura 3 – Diagrama de uma Rede Neural Feedforward (FFNN)

Camada de Entrada: é onde os dados brutos chegam. O número de neurônios aqui é igual ao número de características *(features)* no seu conjunto de dados.

Camadas Ocultas: são as camadas intermediárias, e pode haver uma ou mais. É nelas que a maior parte do processamento ocorre. Cada neurônio numa camada oculta recebe informações de todos os neurônios da camada anterior, faz uma transformação e passa o resultado adiante.

Camada de Saída: esta camada entrega o resultado final da rede, seja uma previsão, uma classificação, ou outro tipo de saída. O número de neurônios aqui varia conforme o problema — por exemplo, um para regressão ou 'n' para classificação com múltiplas classes.

2.2.3 Funcionamento das FNNs - Fundamentos Matemáticos

O funcionamento de uma FNN se baseia em duas operações fundamentais realizadas por cada neurônio: o cálculo da soma ponderada das entradas e a aplicação de uma função de ativação.

Considerando uma rede genérica com L camadas, onde a camada de entrada é a camada *0* e a camada de saída é a camada L:

1 - Entrada dos dados

Cada neurônio da camada de entrada recebe uma característica (*feature*) do vetor de entrada $X = (x_1, x_2, x_3, ..., x_n)$. Esses valores são repassados diretamente para a próxima camada.

2 - Cálculo da entrada líquida (soma ponderada)

Para um neurônio j em uma camada k qualquer, a entrada líquida (ou net input), denotada por:

Para um neurônio j em uma camada qualquer (digamos, camada k), a entrada

líquida (ou *net input*), denotada por $z_j^{(k)}$, é calculada como a soma ponderada das saídas dos neurônios da camada anterior *(k - 1)*, mais um termo de viés.

Seja:

- $w^{(k)}$: matriz de pesos que conecta a camada *(k - 1)* à camada *k*. Cada elemento $w_{ji}^k(k)$ é o peso da conexão do neurônio i na camada *(k - 1)* para o neurônio j na camada k;
- $y_i^{(k-1)}$: é a saída do neurônio i na camada $k-1$;
- $b_j^{(k)}$: é o viés do neurônio j na camada k;
- $z_j^{(k)}$: é a entrada líquida do neurônio j na camada k;
- $\sum_{i=1}^{n}$: O somatório de todas as contribuições dos neurônios da camada anterior.

A fórmula para a entrada líquida $z_j^{(k)}$ do neurônio j na camada k é:

$$z_j^{(k)} = \sum_{i=1}^{n} w_{ji}^{(k)}\, y_i^{(k-1)} + b_j^{(k)} \tag{2.1}$$

Em notação vetorial/matricial (para todos os neurônios da camada k simultaneamente):

$$z^{(k)} = \mathbf{W}^{(k)}\, y^{(k-1)} + b^{(k)} \tag{2.2}$$

Onde:

- $W^{(k)}$: Matriz de pesos que conecta a camada $k-1$ à camada *k*. Cada elemento $w_{ji}^{(k)}$ é o peso da conexão do neurônio i da camada $k-1$ para o neurônio j da camada *k*.

3 - Aplicação da função de ativação em uma FNN

Aplicação da Função de Ativação após calcular a entrada líquida, aplica-se uma função de ativação para obter a saída do neurônio. A função de ativação, denotada por

$$f(.) \tag{2.3}$$

pode ser, por exemplo, a função *sigmoide*, ReLU (*Rectified Linear Unit*), tangente hiperbólica, entre outras. A escolha dessa função influencia diretamente a capacidade de aprendizado e a não-linearidade da rede. Veja Tabela 2 principais funções de ativação para redes neurais *feedforward* (FNN)

O valor de saída $y_j^{(k)}$ do neurônio j na camada k é calculado aplicando-se a função de ativação à sua entrada líquida $z_j^{(k)}$:

$$y_j^{(k)} = f\left(z_j^{(k)}\right). \tag{2.4}$$

A $f(\cdot)$ representa a função de ativação escolhida, como a função sigmoide, ReLU (Rectified Linear Unit), tangente hiperbólica, entre outras. A escolha da função de ativação influencia diretamente a capacidade de aprendizado da rede. Veja a Tabela 2.

Expressão Geral para um Neurônio em uma FNN, para um neurônio j na camada k:

$$z_j^{(k)} = \sum_{i=1}^{n_{k-1}} w_{ij}^{(k)}\, y_i^{(k-1)} + b_j^{(k)}$$

$$y_j^{(k)} = f\left(z_j^{(k)}\right)$$

onde:

- n_{k-1} é o número de neurônios na camada anterior,
- $w_{ij}^{(k)}$ é o peso da conexão entre o neurônio i da camada $k-1$ e o neurônio j da camada k,
- $b_j^{(k)}$ é o termo de viés do neurônio j na camada k,
- $y_i^{(k-1)}$ é a saída do neurônio i na camada anterior.

O valor de saída $y_j^{(k)}$ do neurônio j na camada k é calculado aplicando-se a função de ativação à sua entrada líquida $z_j^{(k)}$:

$$y_j^{(k)} = f\left(z_j^{(k)}\right). \tag{2.5}$$

4. Propagação para a Próxima Camada

Os resultados (ativações) de uma camada são passados adiante para a camada seguinte. O processo de soma ponderada (onde cada entrada é multiplicada por um peso e somada) e ativação (a aplicação de uma função não linear) se repete, camada após camada. É como um revezamento, onde a saída de uma etapa se torna a entrada da próxima, impulsionando a informação através da rede.

5. Produção da Saída Final

Eventualmente, os sinais chegam à camada de saída. Aqui, as ativações finais representam a resposta da rede. Dependendo do que a rede foi projetada para fazer, essa saída pode ser: a probabilidade de algo pertencer a uma categoria específica. Um valor numérico contínuo (como a previsão do preço de uma casa). Ou qualquer outra estimativa que a tarefa exija. É neste ponto que a rede neural entrega sua "decisão"ou "previsão".

O processo de soma ponderada seguido de ativação é repetido camada a camada, até que os sinais alcancem a camada de saída.

Tabela 2 – Principais funções de ativação para redes neurais *feedforward* (FFNN) 'Criada pelo autor'

Função de Ativação	**Expressão Matemática**	**Intervalo de Saída**	**Características Principais**
Sigmoide (Logística)	$f(x) = \frac{1}{1+e^{-x}}$	(0, 1)	Saída suave entre 0 e 1; útil para probabilidades; pode sofrer *vanishing gradient*.
Tangente Hiperbólica (Tanh)	$f(x) = \tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$	(-1, 1)	Similar à sigmoide, mas centrada em zero; também suscetível a *vanishing gradient*.
ReLU (Rectified Linear Unit)	$f(x) = \max(0, x)$	[0, ∞)	Simples e eficiente; ajuda a mitigar *vanishing gradient*; pode sofrer *dead neurons*.
Leaky ReLU	$f(x) = \begin{cases} x, & x \geq 0 \\ \alpha x, & x < 0 \end{cases}$	(-∞, ∞)	Variante da ReLU que permite gradiente pequeno para $x < 0$, evitando neurônios mortos.
ELU (Exponential Linear Unit)	$f(x) = \begin{cases} x, & x \geq 0 \\ \alpha(e^{x} - 1), & x < 0 \end{cases}$	(-α, ∞)	Suaviza a ReLU, com saída negativa controlada; pode melhorar o aprendizado.
Softmax	$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$	(0,1), soma = 1	Normaliza vetor de saídas em probabilidades; usada na camada de saída para classificação multiclasse.
Swish	$f(x) = x \cdot \text{sigmoid}(x)$	(-0.28, ∞)	Função não monotônica proposta pelo Google; pode melhorar a performance em alguns modelos.

2.3 Classificação binária

Machine learning ou apredizagem de máquina é um campo abrangente dentro da inteligência artificial. Uma sub-área de apredizagem de máquina é o *Deep Learning* (ou Redes Neurais Profundas). Temos várias arquiteturas de *Deep Learning*, dentre elas vamos detalhar classificação binária. A classificação binária prevê se a entrada pertence a uma determinada categoria de interesse ou não: fraude ou não-fraude, gato ou não-gato.

Em aprendizado de máquina, muitos métodos utilizam classificação binária. Os mais comuns são: Máquinas de Vetores de Suporte; Naive Bayes; Vizinho mais próximo; Árvores de Decisão; Regressão Logística e Redes Neurais. (Bansilal; THUKARAM; KASHYAP, 2003) (LI et al., 2012) (GOLAB et al., 2022)

2.3.1 Arquitetura de uma Rede Neural para Classificação Binária

Uma rede neural típica para classificação binária é composta por:

- Camada de entrada: recebe os dados de entrada, com cada neurônio representando uma característica (ou feature).
- Camadas ocultas: realizam o processamento não linear por meio de pesos, somas ponderadas e funções de ativação (como ReLU ou tangente hiperbólica).
- Camada de saída: composta por um único neurônio, geralmente com a função de ativação sigmoide, que retorna um valor entre 0 e 1, interpretado como a probabilidade de pertencimento a uma das classes.

2.3.2 Funcionamento

O funcionamento da rede neural em um classificador binário envolve duas etapas principais:

- Propagação direta (feedforward): os dados percorrem a rede da camada de entrada até a camada de saída. Em cada camada, os neurônios recebem esses dados. Eles aplicam uma transformação linear, que é basicamente uma operação matemática simples. Em seguida, usam uma função de ativação para introduzir não-linearidade, o que permite que a rede aprenda padrões mais complexos. O resultado final dessa jornada é a previsão ou saída da rede.
- Treinamento (ajuste dos pesos): o objetivo é que a rede ajuste seus pesos (que são como os "parâmetros"que ela usa para fazer as transformações) para que suas previsões fiquem cada vez mais precisas.

Cálculo do erro: a rede compara sua previsão com o resultado correto (o que ela deveria ter previsto). A diferença entre os dois é o erro.

Função de custo: esse erro é quantificado por uma função de custo (ou função de perda), que indica o quão "ruim"a previsão da rede foi. Uma função comum para tarefas de classificação binária é a entropia cruzada binária. Quanto menor o valor da função de custo, melhor a rede está se saindo.

Retropropagação do erro *(backpropagation)*: este é o algoritmo central. O erro é propagado de volta através da rede (do fim para o começo). Isso permite que a rede descubra quais pesos contribuíram mais para o erro e como ajustá-los.

Otimizador (gradiente descendente): com base nas informações da retropropagação, um otimizador (como o famoso gradiente descendente) entra em ação. Ele usa essa informação para ajustar os pesos da rede de forma gradual, passo a passo, na direção que minimiza a função de custo.

2.3.3 Saída e decisão

A saída do neurônio final, após a função sigmoide, representa a probabilidade de a entrada pertencer à classe positiva (geralmente rotulada como "1"). Para tomar a decisão final, um limiar de decisão é aplicado (por exemplo, 0,5):

Se a saída $>= 0{,}5 \rightarrow$ classe 1 (positiva)

Se a saída $< 0{,}5 \rightarrow$ classe 0 (negativa)

Para um bom desempenho, é essencial cuidar de aspectos como a normalização dos dados, o balanceamento entre as classes (para evitar viés), e a escolha adequada da arquitetura e dos parâmetros da rede.

3 Trabalhos Relacionados

A predição do impacto científico (do autor, do artigo, da instituição, do país) tem sido, por muito tempo, um desafio fundamental na cientometria, com estudos iniciais que introduziram medidas baseadas em citações para avaliar a influência dos investigadores. Uma das contribuições mais notáveis neste campo foi o desenvolvimento do índice-H por (HIRSCH, 2005), uma métrica que integra tanto a contagem de publicações como o impacto das citações, conforme citado em. Desde então, vários índices alternativos foram propostos, como o índice G e o índice M, para abordar as limitações do índice H, incorporando fatores como a longevidade da publicação e a ponderação da coautoria (KOLTUN; HAFNER, 2021).

Ao longo dos anos, várias abordagens foram exploradas para melhorar as previsões bibliométricas. Os modelos tradicionais basearam-se predominantemente em contagens de citações e atributos baseados no autor, como demonstrado por estudos que utilizam dados históricos de citações para prever o impacto futuro (ABRAMO; D'ANGELO; FELICI, 2019). Paralelamente, metodologias baseadas em redes ganharam força, utilizando redes de coautoria e citações para inferir pontuações de influência e prever o sucesso acadêmico (SINATRA; AL., 2016). Mais recentemente, foram aplicadas técnicas de aprendizagem de máquina neste domínio, integrando características como o crescimento inicial das citações, a reputação do autor e os fatores de impacto das revistas para melhorar a precisão preditiva.

Os avanços na predição do impacto também foram impulsionados por modelos de *deep learning* e indicadores altmétricos. Pesquisas indicam que a combinação de padrões iniciais de citações com algoritmos de aprendizagem automática pode melhorar significativamente as previsões de impacto a longo prazo (AKELLA et al., 2021). Além disso, a altmetria — incluindo menções nas redes sociais, downloads e cobertura da mídia — tem sido explorada como indicadores complementares da influência acadêmica, refletindo uma mudança mais ampla das métricas tradicionais baseadas em citações para estruturas de avaliação mais diversificadas e holísticas (MARQUES-CRUZ; AL., 2024).

Apesar destas inovações, continuam a existir vários desafios críticos. Uma lacuna fundamental na literatura é a falta de consenso sobre o intervalo de tempo ideal para calcular índices bibliométricos preditivos. Embora muitos estudos assumam que as publicações recentes são os preditores mais sólidos do sucesso futuro, as provas empíricas sugerem o contrário (TEPLITSKIY et al., 2022). Além disso, embora os modelos de aprendizagem automática tenham melhorado a precisão preditiva, a sua natureza de «caixa preta» limita a interpretabilidade e a confiança nas suas previsões (KOLTUN;

HAFNER, 2021).

Este estudo pretende abordar esses desafios avaliando sistematicamente diferentes janelas temporais para previsões bibliométricas e medindo sua eficácia empírica na previsão do sucesso científico. Ao aproveitar um conjunto de dados abrangente e empregar métodos de avaliação rigorosos, esta investigação fornece novos *insights* sobre a dinâmica temporal da avaliação do impacto acadêmico.

4 Procedimento Metodológico

4.1 Predição de sucesso de pesquisadores

O sucesso de físicos no futuro próximo é medido pelo sucesso de suas publicações científicas após o ano marcado como ano de predição (Figura 4). Já o sucesso de físicos no passado é mensurado pelo sucesso de suas publicações científicas antes do ano marcado como ano de predição. Utilizamos essa abordagem para criar um conjunto de dados de exemplos de físicos bem-sucedidos e não bem-sucedidos, a fim de treinar uma rede neural artificial capaz de prever novos exemplos nunca antes vistos. Definimos isso como um **problema de classificação binária**, onde um físico bem-sucedido é aquele que publicou pelo menos um determinado número de artigos, cada um com este mesmo número de citações, nos anos seguintes ao ano de predição. Por outro lado, um físico não bem-sucedido é o oposto. A nossa abordagem baseia-se na aprendizagem supervisionada, e todo o processo está resumido na Figura 5.

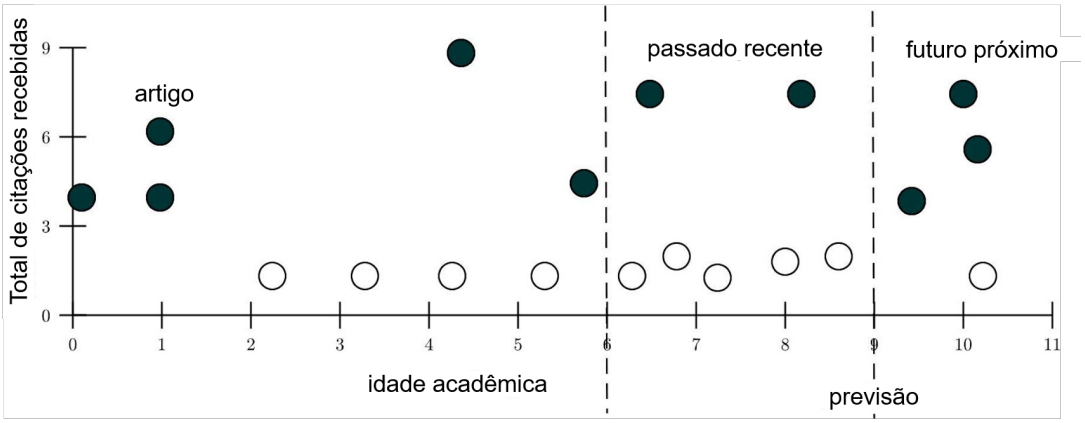


Figura 4 – A lista de publicações de um cientista bem-sucedido, porque ele tem pelo menos três artigos com três citações cada um no futuro próximo.

Utilizamos os seguintes **índices bibliométricos** do cientista como features preditoras:

- **H**: o índice H do cientista u desde o ano específico y de sua carreira até o ano de predição;
- **co**: o número de coautores distintos do cientista u desde o ano y de sua carreira até o ano de predição;
- **c**: o número total de citações do cientista u desde o ano y de sua carreira até o ano de predição;

- I_{iu}: o número total de artigos do cientista u com i citações cada, desde o ano y de sua carreira até o ano de predição;
- **v**: o número total de publicações nas quais o cientista u publicou, desde o ano y de sua carreira até o ano de predição.

Comparamos os efeitos de uma **janela de tempo curta** para a coleta de publicações – que serviu de base para o cálculo dos índices – com uma **janela de tempo longa** – que representa todas as publicações do cientista – no poder preditivo dos preditores H, co e c sobre o desempenho do classificador.

Utilizamos **múltiplas redes neurais *feedforward*** com duas camadas ocultas como classificadores, onde k e L representam o número de neurônios na primeira e segunda camadas ocultas, respectivamente. A camada de saída das redes utiliza uma **função logística**, enquanto as saídas das camadas ocultas empregam **funções ReLU (*Rectified Linear Unit*)**.

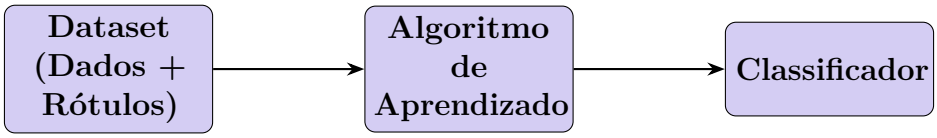


Figura 5 – Fluxo simplificado de geração de modelo.

4.2 Engenharia de características

Utilizamos os seguintes índices bibliométricos como preditores:

- Considere A_{np} uma matriz em que uma entrada a_{ij} indica o número de citações que o artigo i recebeu no ano j. n é o número de artigos pertencentes ao conjunto de dados, e p é o número total de anos após o ano de publicação.
- O **total de citações do artigo** i no conjunto de dados é:

$$c_i = \sum_{j=0}^{p} a_{ij}$$

  onde p é o número máximo de anos.

- O **total de citações de um cientista** u no conjunto de dados é:

$$c_u = \sum_{i \in L_u} c_i$$

  onde L_u é a lista de publicações de u.

- O **índice H de um cientista** u é calculado com base em sua lista de publicações L_u. A fórmula baseia-se no número de artigos (H) que foram citados e com que frequência, em comparação com aqueles que não foram citados (ou não foram citados com tanta frequência). Por exemplo, um pesquisador tem um índice H de 6 porque possui 6 publicações que foram citadas pelo menos 6 vezes. Os artigos restantes ou aqueles que ainda não foram citados 6 vezes são deixados de lado.

- O **número de coautores do cientista** u foi calculado usando a matriz de coautoria C_{nm}, onde uma entrada $c_{ui} = 1$ indica que o autor u escreveu o artigo i, e $c_{ui} = 0$ caso contrário.

4.3 Tratamento do desbalanceamento de classes

Desequilíbrio de classe se refere à distribuição desigual de classes, ou seja, há muitos exemplos em uma classe e poucos em outras classes. Isso afeta o desempenho de algoritmos de aprendizado padrão que exigem uma distribuição balanceada de classes.

Para lidar com o desequilíbrio de classes, existem diferentes estratégias. Uma delas atua diretamente nos dados de treinamento, buscando o equilíbrio através da inclusão ou exclusão de exemplos. Outra linha de ação se concentra no algoritmo de classificação, alterando sua função de aprendizado para que a classe minoritária tenha maior influência. Na nossa estratégia para mitigar o problema de desequilíbrio de classes, adotamos uma abordagem dual, intervindo tanto no nível dos dados de treinamento quanto no nível do algoritmo de aprendizado.

No nível dos dados, implementamos uma técnica de subamostragem da classe majoritária. O desequilíbrio de classes, caracterizado por uma representação significativamente menor da classe minoritária em comparação com a majoritária, pode levar modelos de aprendizado de máquina a serem tendenciosos, favorecendo a classe dominante. Para contrabalancear essa disparidade, a subamostragem envolve a seleção aleatória de um subconjunto das instâncias da classe majoritária e a remoção das demais. Esse processo visa reduzir a predominância da classe majoritária no conjunto de treinamento, tornando a distribuição das classes mais equilibrada e, consequentemente, permitindo que o modelo aprenda de forma mais eficaz os padrões da classe minoritária.

No nível do algoritmo, empregamos uma metodologia de aprendizado sensível ao custo. Essa abordagem reconhece inerentemente a importância diferenciada das classes durante o processo de treinamento do modelo. Especificamente, definimos uma função de perda modificada que atribui um custo maior aos erros de classificação cometidos em exemplos da classe positiva (que, neste contexto, é a classe minoritária). Ao penalizar mais severamente as previsões incorretas da classe minoritária, forçamos o algoritmo a

dar maior atenção a esses exemplos durante o aprendizado. Essa sensibilidade ao custo direciona o modelo a otimizar seus parâmetros de forma a minimizar os erros na classe minoritária, resultando em um classificador mais robusto e com melhor capacidade de generalização para ambas as classes.

Resumo do Fluxo

- **Entrada:** dados altamente desequilibrados (classe majoritária domina).
- **Etapa 1 (Dados):** subamostragem para reduzir exemplos da classe majoritária → equilíbrio na distribuição de classes.

  (- Seleção aleatória de parte da classe majoritária - Remoção de excesso de exemplos => Resultado: classes com distribuição mais equilibrada).
- **Etapa 2 (Algoritmo):** função de perda sensível ao custo → penaliza mais os erros na classe minoritária.

  (- Modificação da função de perda - Maior penalidade para erros na classe minoritária (positiva) => Resultado: Modelo foca mais na classe minoritária, melhorando generalização).
- **Saída:** modelo treinado com maior atenção à classe minoritária → melhor desempenho em ambas as classes.

Em resumo, abordamos o desequilíbrio de classes nos níveis de dados e algoritmos. No nível de dados, usamos a subamostragem da classe majoritária e removemos instâncias do conjunto de dados de treinamento para restaurar o equilíbrio. No nível de algoritmo, usamos uma abordagem de aprendizado sensível ao custo que especifica uma função de perda que considera um peso maior para exemplos positivos (minoritários).

5 Experimentação

5.1 Configuração do experimento

Para verificar a nossa hipótese de que os índices bibliométricos calculados a partir das publicações recentes de um cientista são melhores indicadores do sucesso dos cientistas do que aqueles calculados a partir das publicações de toda a sua carreira, avaliamos o desempenho dos classificadores preditivos (Tabela 3). O objetivo era classificar os cientistas entre aqueles com e sem um número determinado de artigos de sucesso (3 esse número) a curto prazo no conjunto de dados da APS.

Tabela 3 – Hiperparâmetros Utilizados nos Classificadores

Configurações dos Hiperparâmetros				
	NET	**Camada *j***	**Camada *k***	**Peso *w***
Aprendizado Sensível ao Custo	1	25	25	4
	2	25	25	4,5
	3	25	25	5
	4	50	50	4
	5	50	50	4,5
	6	50	50	5
	7	100	100	4
	8	100	100	4,5
	9	100	100	5
Subamostragem da Classe Majoritária	10	25	25	1[1]
	11	50	25	1
	12	100	25	1
	13	25	50	1
	14	50	50	1
	15	100	50	1
	16	25	100	1
	17	50	100	1
	18	100	100	1

Avaliamos os índices de citações totais, o índice h e o número de coautores, comparando o efeito desses preditores calculados com as publicações do cientista nos últimos três anos no desempenho dos classificadores em comparação com quando o índice é calculado com todas as publicações do cientista. Os outros preditores (mistos) são idênticos em ambos os casos, para isolar a influência desse índice.

Utilizamos a média de uma validação cruzada k-fold (com k=10) para determinar o classificador mais preciso e, indiretamente, os efeitos destas duas abordagens. O número total de épocas e tamanhos de batch para o treino em cada fold foi de 300 épocas cada, e

o tamanho do batch foi de 100. Um peso de 4 (Tabela 3) significa que a função de perda se comporta como se o conjunto de dados contivesse 4 vezes mais exemplos positivos (da classe minoritária)

5.2 Conjunto de dados APS

Utilizamos o conjunto de dados gerenciado pela APS [2] como fonte de dados para nossas análises e utilizamos os nomes dos autores dos metadados dos artigos para identificá-los.

Definimos a idade acadêmica dos autores com base nos anos entre sua primeira e última publicação. Selecionamos autores que começaram a publicar após 1985 e tinham idade acadêmica de pelo menos 8 anos. Utilizamos suas publicações até três anos antes de sua idade máxima para calcular o valor dos preditores e suas publicações restantes como publicações futuras. Utilizamos todas as citações de um determinado artigo dentro do conjunto de dados e, portanto, não definimos uma janela temporal para o acúmulo de citações. O conjunto de dados é desbalanceado, aproximadamente 20 por cento (Figura 6) são exemplos de físicos bem-sucedidos (classe 1) e o restante são exemplos de físicos malsucedidos.

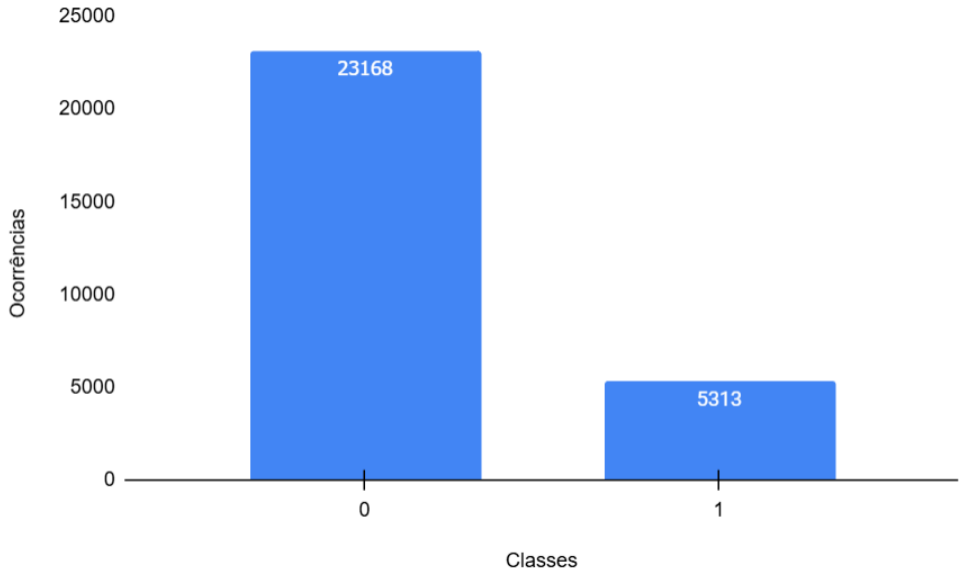


Figura 6 – A distribuição desigual de classes no conjunto de dados APS.

5.3 Métricas de avaliação de desempenho

O desempenho dos modelos é medido utilizando o **coeficiente de correlação de Matthews (MCC)** (THAI-NGHE; GANTNER; SCHMIDT-THIEME, 2010) e a

[2] https://journals.aps.org/datasets

métrica F1-score (HAND; CHRISTEN; KIRIELLE, 2021) em uma validação cruzada k-fold. Escolhemos essas métricas por serem as mais precisas ao avaliar classificadores treinados com muito mais exemplos de uma classe do que da outra. Apenas 20% dos nossos exemplos provêm de cientistas com sucesso a curto prazo.

Abordamos o desbalanceamento de classes com **técnicas de reamostragem** e **aprendizado sensível ao custo** (cost-sensitive learning). O primeiro foca em manipulações no nível dos dados, e o último na mudança interna do classificador.

O **coeficiente de correlação de Matthews (MCC)** é uma medida estatística que só resulta em uma pontuação alta se a predição tiver um bom desempenho em todas as quatro categorias da matriz de confusão: **verdadeiros positivos (TP)**, **falsos negativos (FN)**, **verdadeiros negativos (TN)** e **falsos positivos (FP)**. O MCC é a única métrica de desempenho para classificação binária que produz uma pontuação alta apenas quando o classificador binário é capaz de prever corretamente a maioria das instâncias da classe majoritária e a maioria da classe minoritária.

O cálculo do MCC é dado por:

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

No entanto, também utilizamos o **F1-score** como uma segunda, mas não a principal, medida de desempenho.

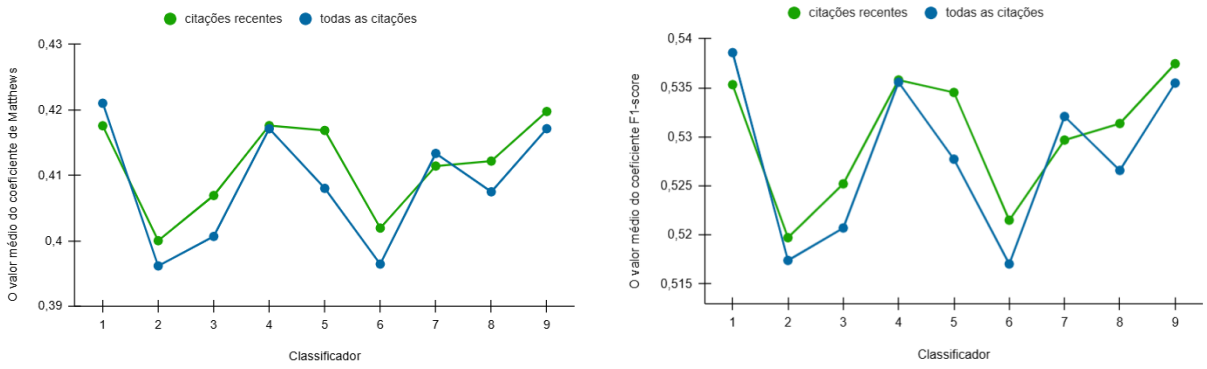
O cálculo do F1-score é dado por:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

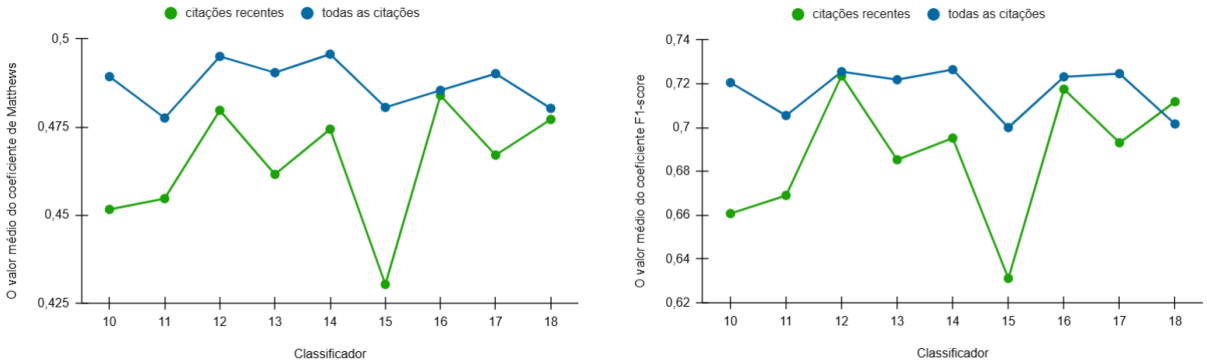
5.4 Resultados e discussão

Como pode ser observado na Figura 7, os resultados deste experimento sugerem que a nossa hipótese deve ser rejeitada. Os valores médios de desempenho numa validação cruzada k-fold (k=10) entre os classificadores treinados no conjunto de dados com a característica total de citações recebidas para todas as publicações do cientista são superiores aos dos outros classificadores treinados com a característica Total de citações recebida apenas para as publicações mais recentes. Um coeficiente de correlação de Matthews mais alto e um valor F1 mais elevado indicam um melhor desempenho de um classificador em comparação com um classificador com um valor inferior. Da mesma forma, as Figuras 8 e 9 confirmam uma clara tendência para rejeitar a nossa hipótese, independentemente do índice bibliométrico utilizado como preditor (coautoria, índice H, citações).

É importante notar que este resultado é independente do método escolhido para resolver o problema do desequilíbrio de classes (subamostragem da classe maioritária e abordagem de aprendizagem sensível ao custo). Isto sugere o contrário do que pensávamos, ou seja, que é melhor utilizar um intervalo de tempo curto para a coleta de publicações do que um intervalo de tempo completo como base para o cálculo dos indicadores preditivos. Os resultados deste estudo defendem um intervalo de tempo amplo que inclua todos os artigos de uma carreira.



a) abordagem de aprendizagem sensível ao custo



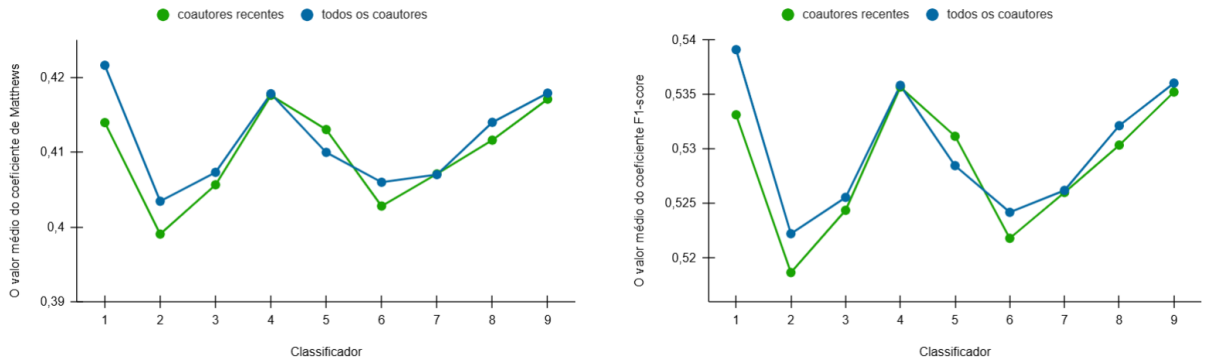
b) Subamostragem da classe majoritária

Figura 7 – Comparação dos valores médios entre os desempenhos dos classificadores individuais na validação cruzada de 10-folds, onde o indicador, citações, foi calculado com dados recentes e dados de toda a carreira.

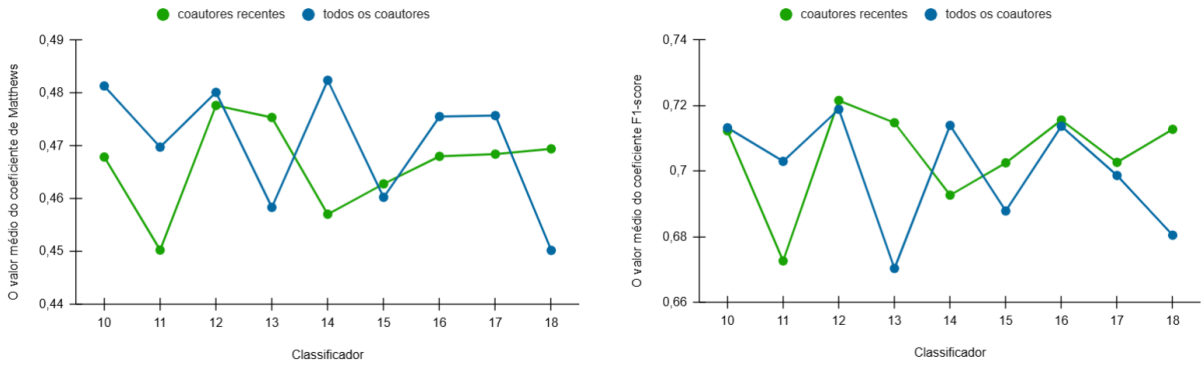
Este trabalho demonstrou que o tamanho da janela temporal para a coleta de publicações, utilizada para calcular os índices preditivos, é um fator crítico para a precisão dos preditores, mas até agora tem sido ignorado.

Como observam (MOMENI; MAYR; DIETZE, 2023), as evidências que encontramos sugerem que as métricas baseadas nos autores são bons preditores das próprias métricas futuras desses investigadores, mas os intervalos de tempo utilizados para calculá-las são cruciais.

A principal limitação deste estudo deriva do fato de se centrar num conjunto



a)Aprendizado Sensível ao Custo

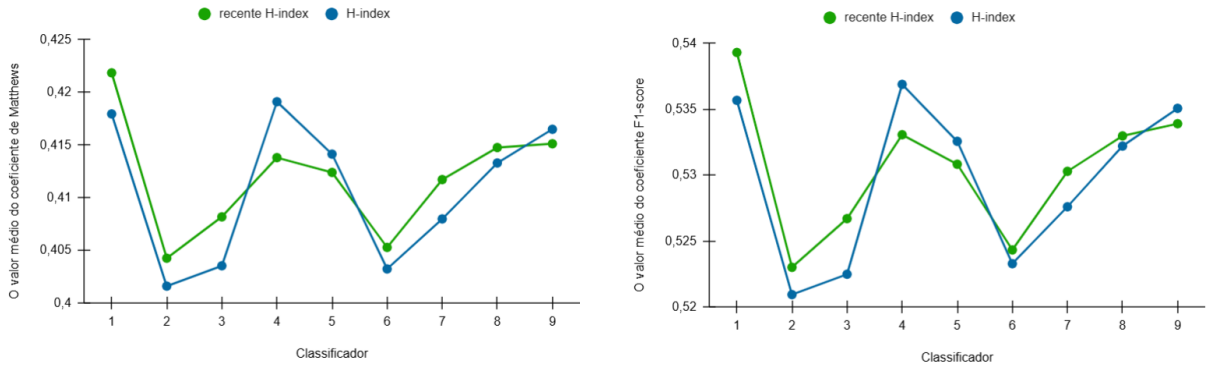


b)Subamostragem

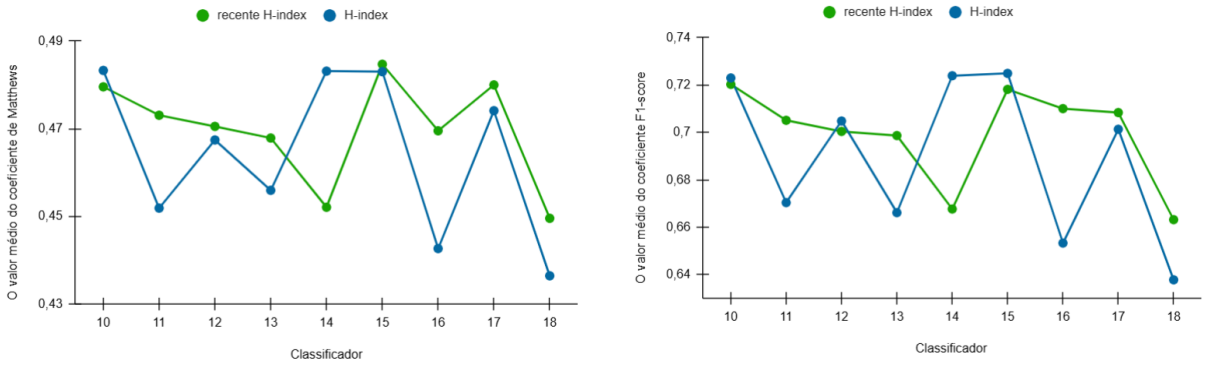
Figura 8 – Comparação dos valores médios entre as performances dos classificadores individuais na validação cruzada de 10-folds, onde o indicador, o número total de coautores (diferentes), foi calculado com dados recentes e dados de toda a carreira.

de dados específico de físicos, e as principais observações deste estudo podem não ser generalizáveis a outros campos. Uma outra grande fonte de incerteza reside na natureza dos índices utilizados como preditores, que se baseiam exclusivamente em citações de publicações. A avaliação com índices que se concentram mais nos metadados das publicações é um tema importante para pesquisas futuras.

Além disso, reconhecemos que há vários fatores adicionais que devem ser considerados na seleção do corpus de treino. Entre eles estão a afiliação geográfica dos autores, a diversidade de subcampos dentro da física, o alcance temático da literatura relevante, o impacto potencial das funções administrativas dos investigadores nas suas instituições, entre outras variáveis contextuais que podem influenciar os resultados da investigação. Essas limitações destacam a complexidade inerente à criação de conjuntos de dados que abranjam não apenas os metadados das publicações e as redes de citações, mas também informações contextuais mais ricas que muitas vezes são difíceis de obter.



a)Aprendizado Sensível ao Custo



b)Subamostragem

Figura 9 – Comparação dos valores médios entre os desempenhos dos classificadores individuais na validação cruzada de 10-folds, em que o indicador, índice H, foi calculado com dados recentes e dados de toda a carreira.

Portanto, utilizar uma janela temporal ampla, que inclua todas as publicações da carreira de um cientista, é mais eficaz para calcular indicadores preditivos de sucesso futuro do que se concentrar apenas nas publicações recentes. O estudo destaca a importância crucial do tamanho do intervalo temporal na precisão desses preditores, um fator que havia sido negligenciado em pesquisas anteriores. Essa conclusão se mantém independentemente do método utilizado para abordar o desequilíbrio de classes e é consistente para diferentes índices bibliométricos (citações totais, índice H e número de citações).

5.5 Implicações e limitações para cientistas e formuladores de políticas acadêmicas

O estudo implica que tanto cientistas quanto formuladores de políticas devem considerar a importância de uma perspectiva de longo prazo ao avaliar o impacto e o potencial futuro na ciência, especialmente no que se refere ao uso de índices bibliométricos. Além disso, ressalta a necessidade de explorar métricas mais abrangentes e de realizar

pesquisas em diferentes áreas para validar e refinar esses achados.

Implicações para Cientistas

Consideração da carreira completa: os cientistas devem reconhecer que a avaliação de seu impacto e potencial futuro pode ser mais precisamente estimada considerando o histórico completo de suas publicações, e não apenas a produção mais recente. Isso pode influenciar a forma como eles apresentam seu trabalho em currículos, solicitações de financiamento e processos de avaliação.

Consciência da janela temporal: os pesquisadores precisam estar cientes de que a janela temporal utilizada para calcular seus índices bibliométricos pode afetar a percepção de seu desempenho. Ao apresentar métricas, pode ser útil fornecer informações sobre o período considerado para o cálculo.

Relevância de citações ao longo do tempo: o estudo sugere que o impacto de um trabalho científico pode se manifestar e ser reconhecido ao longo de um período mais extenso. Isso pode encorajar os cientistas a valorizarem e acompanharem o impacto de suas publicações antigas, que continuam a influenciar a comunidade científica.

Abertura a métricas complementares: dada a limitação do estudo em focar apenas em citações, os cientistas podem se beneficiar de uma visão mais ampla de seu impacto, considerando também métricas baseadas em metadados, como colaborações e afiliações, que podem fornecer uma imagem mais rica de sua contribuição.

Implicações para Formuladores de Políticas Acadêmicas

Avaliação de desempenho: as instituições e agências de financiamento que utilizam índices bibliométricos para avaliar o desempenho de pesquisadores podem precisar reconsiderar a janela temporal padrão para o cálculo dessas métricas. O estudo sugere que uma visão de longo prazo, abrangendo a carreira completa, pode ser mais preditiva de sucesso futuro.

Alocação de recursos: se os índices baseados na carreira completa são melhores preditores de sucesso, as políticas de alocação de recursos podem se beneficiar ao darem peso a essa perspectiva, em vez de se focarem excessivamente na produção recente.

Planejamento de carreira e avaliação institucional: as instituições podem usar esses achados para desenvolver modelos de avaliação mais robustos e equitativos, que considerem a trajetória completa dos pesquisadores. Isso pode auxiliar no planejamento de carreira e na identificação de talentos com potencial de impacto a longo prazo.

Desenvolvimento de métricas mais abrangentes: os formuladores de políticas podem incentivar o desenvolvimento e a utilização de métricas de avaliação que vão além das citações, incorporando informações sobre colaborações, financiamento e outros aspectos do impacto acadêmico, conforme sugerido para pesquisas futuras no texto.

Cautela na generalização: é crucial reconhecer a limitação do estudo em se concentrar apenas na física. Os formuladores de políticas devem ser cautelosos ao generalizar esses resultados para outras disciplinas e podem precisar de estudos específicos em diferentes áreas para informar suas decisões.

6 Considerações Finais

Nosso estudo revela que o cálculo de índices preditivos com base em uma janela temporal curta para a coleta de publicações resulta em um desempenho inferior do classificador em comparação com o uso de uma janela temporal mais extensa. Essa descoberta refuta nossa expectativa inicial de que um período mais recente de publicações geraria previsões mais acuradas. Adicionalmente, nossos achados sugerem que a subamostragem se mostra uma tática mais eficiente do que a Aprendizagem Sensível ao Custo para lidar com o desequilíbrio de classes.

Embora esses resultados demonstrem robustez, sua aplicabilidade a outras áreas do saber ainda é uma questão em aberto. Faz-se necessário conduzir mais investigações experimentais para validar essas conclusões em diversos campos acadêmicos e para otimizar a determinação da janela temporal ideal para preditores fundamentados em citações.

Além do aprimoramento metodológico, nossa pesquisa sinaliza diversas avenidas promissoras para futuras explorações. A expansão do conjunto de preditores para incorporar índices baseados em metadados – como redes de coautoria, afiliações institucionais e fontes de financiamento – pode potencializar a precisão das previsões de sucesso. Outrossim, a integração de técnicas de aprendizado profundo, especialmente modelos de IA explicáveis, pode oferecer *insights* mais profundos sobre os intrincados fatores que moldam o impacto futuro de um pesquisador.

A discussão sugere a necessidade de investigar o desempenho de índices baseados em metadados das publicações, além das citações. Também ressalta a importância de considerar outros fatores contextuais que podem influenciar os resultados, como afiliação geográfica, diversidade de subcampos e funções administrativas. A dificuldade em obter conjuntos de dados abrangentes que incluam essas informações contextuais é reconhecida.

Em um panorama mais amplo, este estudo enriquece a discussão contemporânea sobre bibliometria preditiva, enfatizando a relevância de métricas de avaliação abrangentes. A identificação das condições ótimas para prever o sucesso acadêmico pode otimizar a alocação de recursos, o planejamento de carreira e a avaliação institucional. Em última instância, o refinamento desses modelos preditivos aumentará a transparência na avaliação científica, atenuará vieses na análise de pesquisas e fomentará um ambiente acadêmico mais justo.

Referências

ABRAMO, G.; D'ANGELO, C. A.; FELICI, G. Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics*, v. 13, n. 1, p. 32–49, 2019. Citado na página 36.

ACUNA, D. E.; ALLESINA, S.; KORDING, K. P. Predicting scientific success. *Nature*, v. 489, n. 7415, p. 201–202, set. 2012. Citado na página 12.

AKELLA, A. P.; ALHOORI, H.; KONDAMUDI, P. R.; FREEMAN, C.; ZHOU, H. Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics*, v. 15, n. 2, p. 101128, maio 2021. Citado na página 36.

ARAúJO, C. A. A. Bibliometria: evolução histórica e questões atuais. *Em Questão*, v. 12, n. 1, p. 11–32, dez. 2006. Citado na página 16.

AYAZ, S.; MASOOD, N. Comparison of researchers' impact indices. *PLOS ONE*, v. 15, p. e0233765, maio 2020. Citado na página 12.

BAI, X.; ZHANG, F.; LEE, I. Predicting the citations of scholarly paper. *Journal of Informetrics*, v. 13, n. 1, p. 407–418, fev. 2019. Citado na página 12.

Bansilal; THUKARAM, D.; KASHYAP, K. Artificial neural network application to power system voltage stability improvement. In: *TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region*. Bangalore, India: Allied Publishers Pvt. Ltd, 2003. p. 53–57. Citado na página 34.

BATISTA-JR, A. de A.; GOUVEIA, F. C.; MENA-CHALCO, J. P. Predicting the q of junior researchers using data from the first years of publication. *Journal of Informetrics*, v. 15, n. 2, p. 101130, 2021. Citado na página 12.

BECKER, D.; BREDA, W. V.; FUNK, B.; HOOGENDOORN, M.; RUWAARD, J.; RIPER, H. Predictive modeling in e-mental health: A common language framework. *Internet Interventions*, v. 12, p. 57–67, jun. 2018. Citado na página 22.

CHEN, M.; CHALLITA, U.; SAAD, W.; YIN, C.; DEBBAH, M. *Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial*. 2019. Citado na página 27.

COTINGUIBA, J. R. R. d. O. A ciência desde a antiguidade até o renascimento. *Cuadernos de Educación y Desarrollo*, v. 15, n. 9, p. 8822–8840, set. 2023. Citado na página 25.

GERVEN, M. V.; BOHTE, S. Editorial: Artificial Neural Networks as Models of Neural Information Processing. *Frontiers in Computational Neuroscience*, v. 11, p. 114, 2017. Citado na página 27.

GOLAB, A.; GOOYA, E. S.; FALOU, A. A.; CABON, M. A multilayer feed-forward neural network (MLFNN) for the resource-constrained project scheduling problem (RCPSP). *Decision Science Letters*, v. 11, n. 4, p. 407–418, 2022. Citado na página 34.

HAND, D. J.; CHRISTEN, P.; KIRIELLE, N. F*: an interpretable transformation of the F-measure. *Machine Learning*, v. 110, n. 3, p. 451–456, mar. 2021. Citado na página 44.

HAYKIN. *Redes neurais: princípios e prática. Porto Alegre, RS, Brasil.* [S.l.]: Bookman, 2001. Citado na página 27.

HIRSCH, J. E. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, v. 102, n. 46, p. 16569–16572, 2005. Citado na página 36.

JUNIOR, A. d. A. B. *Predição na Ciência da Ciência: Explicativas de Modelos para Predições de Impacto Futuro de Cientistas Júnior*. Tese (Doutorado) — Universidade Federal do ABC, Programa de Pós Graduação em Ciência da Computação, Santo André, 2021. Orientador: Jesús Pascual Mena Chalco. Citado 2 vezes nas páginas 22 e 25.

KOLTUN, V.; HAFNER, D. The h-index is no longer an effective correlate of scientific reputation. *PLOS ONE*, Public Library of Science, v. 16, n. 6, p. 1–16, 2021. Citado 2 vezes nas páginas 36 e 37.

KOVáCS. *Redes Neurais Artificiais*. [S.l.]: Editora Livraria da Fisica, 2002. Citado na página 27.

LI, J.; CHENG, J.-h.; SHI, J.-y.; HUANG, F. Brief Introduction of Back Propagation (BP) Neural Network Algorithm and Its Improvement. In: JIN, D.; LIN, S. (Ed.). *Advances in Computer Science and Information Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. v. 169, p. 553–558. Citado na página 34.

LIVINGSTONE, D. J. *Artificial neural networks: methods and applications*. [S.l.]: UK: Springer, 2008. v. 458. Citado 2 vezes nas páginas 23 e 27.

LOUPPE, G. *Understanding Random Forests: From Theory to Practice*. 2015. Citado na página 23.

MARQUES-CRUZ, M.; AL. et. Ten year citation prediction model for systematic reviews using early years citation data. *Scientometrics*, v. 129, n. 8, p. 4847–4862, 2024. Citado na página 36.

MOMENI, F.; MAYR, P.; DIETZE, S. Investigating the contribution of author- and publication-specific features to scholars' h-index prediction. *EPJ Data Science*, v. 12, n. 1, 2023. Citado na página 45.

RUAN, X.; ZHU, Y.; LI, J.; CHENG, Y. Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics*, v. 14, n. 3, p. 101039, 2020. Citado 2 vezes nas páginas 18 e 29.

SANTOS, W. R. D.; GALLETTI, R. C. A. F. História do Ensino de Ciências no Brasil: Do Período Colonial aos Dias Atuais. *Revista Brasileira de Pesquisa em Educação em Ciências*, p. e39233, maio 2023. Citado na página 25.

SILVA, J. A. D.; BIANCHI, M. D. L. P. Cientometria: a métrica da ciência. *Paidéia (Ribeirão Preto)*, v. 11, n. 21, p. 5–10, 2001. Citado 3 vezes nas páginas 17, 19 e 22.

SINATRA, R.; AL. et. Quantifying the evolution of individual scientific impact. *Science*, v. 354, n. 6312, p. aaf5239, 2016. Citado na página 36.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. *Axiomatic Attribution for Deep Networks*. 2017. Citado na página 29.

TEPLITSKIY, M.; DUEDE, E.; MENIETTI, M.; LAKHANI, K. R. How status of research papers affects the way they are read and cited. *Research Policy*, v. 51, n. 4, p. 104484, maio 2022. Citado na página 36.

THAI-NGHE, N.; GANTNER, Z.; SCHMIDT-THIEME, L. Cost-sensitive learning methods for imbalanced data. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. Barcelona, Spain: IEEE, 2010. p. 1–8. Citado na página 43.

VIEIRA, L. J. C.; SILVA, I. C. O. D. A produção científica sobre os estudos bibliométricos no Brasil: uma análise a partir da Brapci. *Em Questão*, v. 29, p. e–128160, 2023. Citado na página 16.

XIA, W.; LI, T.; LI, C. A review of scientific impact prediction: tasks, features and methods. *Scientometrics*, v. 128, n. 1, p. 543–585, 2023. Citado na página 13.

ZHAO, Q.; FENG, X. Utilizing citation network structure to predict paper citation counts: A Deep learning approach. *Journal of Informetrics*, v. 16, n. 1, p. 101235, fev. 2022. Citado 2 vezes nas páginas 23 e 29.