UNIVERSIDADE FEDERAL DO MARANHÃO

Curso de Ciência da Computação

Karla Felícia Carvalho da Silva

Um Método Computacional para Identificação de Propagação de Vieses usando Redes Neurais

São Luís - MA

2022

Karla Felícia Carvalho da Silva

Um Método Computacional para Identificação de Propagação de Vieses usando Redes Neurais

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

Curso de Ciência da Computação

Universidade Federal do Maranhão

Orientador: Prof. Dr. Antonio de Abreu Batista Jr

São Luís - MA

2022

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

Silva, Karla Felícia Carvalho da.
Um Método Computacional para Identificação de Propagação de Vieses usando Redes Neurais / Karla Felícia Carvalho da Silva. – 2022.
30 f.

Orientador(a): Antônio de Abreu Batista-Jr.
Monografia (Graduação) – Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís, 2022.

1. Efeito Mateus. 2. Gradient x input. 3. Redes Neurais. I. Batista-Jr, Antônio de Abreu. II. Título.

Karla Felícia Carvalho da Silva

Um Método Computacional para Identificação de Propagação de Vieses usando Redes Neurais

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

Trabalho Aprovado, em 24 de Agosto de 2022:

Prof. Dr. Antonio de Abreu Batista Jr
Orientador
Universidade Federal do Maranhão

Prof. Dr. Luciano Reis Coutinho
Examinador Interno
Universidade Federal do Maranhão

Prof. Dr. Jesús Pascual Mena Chalco
Examinador Externo
Universidade Federal do ABC

São Luís - MA
2022

Dedico à maior doutora de que já conheci: minha mãe, doutora em me amar.

Agradecimentos

Agradeço a Deus, por ter me dado forças para não desistir desse processo que foi tão árduo.

A minha mãe, que mesmo sozinha teve força e coragem para lutar pelo melhor para mim, pelo seu trabalho incessante que tornou possível eu ser a primeira pessoa da família a entrar em uma Universidade.

Agradeço ao professor Antonio, meu orientador, pela oportunidade, paciência e confiança que depositou em mim, eu não poderia ter escolhido uma pessoa melhor.

Gratidão ao meu querido José Emanuel, por ter me acompanhado desde o primeiro trabalho de implementação, pelo seu amor e companhia, companheiro de ônibus, de disciplina, de estágio, de estudo, de tudo que fizemos nos últimos anos na UFMA e fora dela.

Aos meus colegas de turma, por compartilharem comigo tantos momentos de aprendizado e por todo o companheirismo ao longo deste percurso, Samuel Silva, Alan Marques, Marcos Leite, Bruno Santos e Renan Souza.

Agradeço aos meus grandes amigos, Gustavo Oliveira, Pablo Costa, Alefe Brian por sempre me ajudarem quando foi preciso. Em especial, Gustavo pelas várias ajudas que me deu em programação.

Gratidão ao Frederic Menezes por tirar dúvidas sobre artigo cientifico, o que foi crucial para finalizar este trabalho.

Por fim, agradeço a todos que me ajudaram ao longo dessa jornada.

"Por vezes sentimos que aquilo que fazemos não é senão uma gota de água no mar. Mas o mar seria menor se lhe faltasse uma gota"

(Madre Teresa de Calcutá)

Resumo

O efeito Mateus é um fenômeno social descrevendo as recompensas desproporcionais colhidas por aqueles em posições privilegiadas. Apesar disso agravar as disparidades de impacto já existentes entre cientistas, poucos pesquisadores têm estudado o problema. Este trabalho propõe um novo método para identificar este viés. Usando o método do *Gradient x input*, computamos as contribuições de um conjunto de características do cientista para seu índice-h futuro, e comparamos suas intensidades encontradas para cientistas eminentes contra os valores encontrados para cientistas anônimos. As descobertas desse estudo indicam que as diferenças de impacto entre cientistas vão aumentar.

Palavras-chave: Efeito Mateus, Gradient x input, Redes Neurais.

Abstract

The Matthew effect is a social phenomenon describing the disproportionate rewards reaped by scientists in privileged positions. Despite the negative impacts of this, few researchers have addressed the problem. This work proposes a novel method to identify this bias. Using the *Gradient x input* method, we compute the contributions of a set of characteristics linked to an eminent scientist to his future h-index and compare their intensities against the values found in an anonymous scientist. The findings of this study indicate that the contrasts between scientists will increase.

Keywords: Matthew effect, Gradient x input, Neural Network

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AM	*Aprendizado de Máquina*
MSE	*Mean Squared Error*
RMSE	*Root Mean Squared Error*
RELU	*Rectified Linear Unit*

Sumário

1 Introdução

Aprendizado de máquina tem muitas aplicações críticas, mas a renuncia do controle e da supervisão humana para estas máquinas tem causado desconforto para nós humanos, principalmente devido as suspeitas de discriminação contra pessoas em posições de menor destaque (KÖCHLING; WEHNER, 2020). A próxima década é provável testemunhar um aumento considerável nas pesquisas relacionadas a temática.

No contexto da ciência, diversos estudos (BOL; VAAN; RIJT, 2018; KÖCHLING; WEHNER, 2020; BIANCHINI; MüLLER; PELLETIER, 2022) têm concluído um alargamento das distâncias entre cientistas respeitados e desconhecidos. A causa disso pode ser muitos fatores, e um deles pode ser o efeito Mateus (MERTON, 1968).

O impacto disso em uma carreira científica ainda é pouco entendido. A fim de preencher essa lacuna, neste trabalho, propõe-se um novo método para identificar a intensidade desse viés presente nas explicações do método do Gradient x input (SHRIKUMAR; GREENSIDE; KUNDAJE, 2017), um método de atribuição que usa a informação do gradiente de uma função para demonstrar se uma variável de entrada é relevante.

Nossa hipótese é que o método do Gradient x input, provendo o entendimento por trás de uma decisão de uma rede neural, involuntariamente, discrimina entre cientistas respeitados e cientistas pouco conhecidos.

A abordagem proposta compara a relevância de cada característica do cientista provida pelo método, para cientistas eminentes e anônimos. Este trabalho tem o potencial de contribui para a diminuição de vieses em aprendizado de máquina.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo geral deste trabalho consiste em propor um novo método para identificar a presença do efeito Mateus em explicações de predições de redes neurais.

1.1.2 Objetivos Específicos

Especificamente, este trabalho busca os seguintes objetivos menores:

- Demostrar que o método do Gradient x input discrimina entre cientistas respeitados e não respeitados;
- Validar o método proposto;

1.2 Contribuições

Destacam-se como principais contribuições:

- A introdução da IA explicável na Cientometria.
- Melhorias no entendimento do que mais contribui (qual das características do cientista) para o índice-h futuro do cientista.
- Desenvolvimento de um método para identificação de modelos enviesados.

2 Trabalhos Relacionados

O efeito Mateus implica que o cientista mais respeitado recebe mais crédito do que o cientista com menor prestígio por uma contribuição científica, embora a contribuição seja da mesma qualidade técnica.

Um estudo anterior por Bornmann et al. (2020) concluiu que o crédito não é atribuído de forma justa com base na contribuição realizada, mas (injustamente) com base em contribuições anteriores.

Para Silva (2021) o efeito Mateus impacta a ciência e a publicação acadêmica ao amplificar preferencialmente citações, métricas e status, e concluiu que a fama ou status de um acadêmico pode não ser construído exclusivamente com base no mérito da pesquisa.

Por isso, enviesamentos de modelos têm sido uma preocupação constante da comunidade científica (ACUNA; ALLESINA; KORDING, 2012; ZENG et al., 2022). No entanto, falta ainda métodos claros para identificar esse viés.

Diferentemente dos estudos anteriores, o foco deste trabalho é mostrar que métodos de explicação (oriundos da IA explicável) destinados a explicar modelos de aprendizado de máquina perpetuarão o efeito Mateus, ao favorecer cientistas em posições privilegiadas. E, adicionalmente, introduzir/avaliar o uso destes métodos para a identificação de vieses em modelos.

3 Fundamentação Teórica

Este capítulo apresenta os conceitos explorados para o desenvolvimento do estudo, bem como a base teórica utilizada para os experimentos práticos. O conteúdo a seguir discorre sobre o Aprendizado de Máquina, especificando posteriormente o camppo do Aprendizado de Máquina Explicável. Além disso, traz sobre os métodos de atribuição baseados em gradiente, sendo um deles escolhido para as práticas: Gradient x Input.

3.1 Efeito Mateus

Este é um fenômeno social descrevendo as recompensas desproporcionais colhidas por aqueles em posições privilegiadas (WANG; BARABáSI, 2021). O termo foi utilizado pela primeira vez em 1968 por Robert K Merton em estudos de sociologia da ciência (MERTON, 1968).

Merton deu a este fenômeno este nome inspirado em uma passagem bíblica que diz: “para aquele que tem, tudo lhe será dado e terá em abundância; mas para aquele que não tem, até o que tem lhe será tirado “ — Evangelho de Mateus, capítulo 13, versículo 12.

Merton observou que entre seus pares era frequente a manifestação do sentimento de que cientistas eminentes recebiam crédito desproporcionalmente elevado por suas contribuições à ciência, ao passo que cientistas pouco conhecidos recebiam pouco ou nenhum crédito por suas contribuições, mesmo que estas tivessem sido comparativamente mais relevantes.

Stephen Stigler propôs um fenômeno denominado lei de Stigler (GIERYN, 1980), que se refere ao Efeito Mateus, no qual ele diz que nenhuma descoberta científica recebe o nome de seu descobridor original, pois o autor mais eminente receberia maior crédito. Em vista disso, ele pondera que no âmbito cientifico o Efeito Mateus abrange mais do que um simples reconhecimento, pois configura um sistema desproporcional de recompensa por quantidade de contribuição, no qual autores conhecidos tem maiores visibilidade em seus artigos do que autores desconhecidos com propostas igualmente válidas, o que gera desigualdade na distribuição de recursos e nos processos de eleição social numa estrutura coletiva de vantagem acumulada.

Os estudos sobre o Efeito Mateus ainda podem ser aplicados atualmente, ao perceber que as agências de fomento à pesquisa priorizam pesquisadores que já estão bem situados no meio acadêmico, levando a questionar o desmerecimento de pesquisadores iniciantes que precisam de apoio para o desenvolvimento de sua carreira.

Assim, o estudo do efeito Mateus na ciência pode auxiliar em um processo

de mudança na ciência, a tornando mais acessível e democrática, interrompendo a discriminação que atualmente ocorre.

3.2 Aprendizado de Máquina

Aprendizado de Máquina (AM) é um subcampo da Ciência da Computação interessado na criação de programas de computadores que podem melhorar através da experiência (MITCHELL, 1997)- a partir de dados. Um programa de computador é dito ter aprendido alguma classe de tarefas T, a partir da experiência E e medida de desempenho P, se o desempenho do programa nas tarefas em T, como medida por P, melhora com a experiência E.

Dito de forma prática, é dado um conjunto de dados $D = \{(\mathbf{x}_1, y_1); (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, em que y_i é a saída para a entrada $\mathbf{x}_i = \{x_{i1}; x_{i2}, \cdots, x_{im}\}$. Para avaliar o desempenho do modelo inferido (preditor) f, compara-se a predição do modelo $f(\mathbf{x}_i)$ contra o valor de fato y_i, em dado não visto. O objetivo durante o aprendizado é minimizar este erro em dado diferente do anterior, denominado de conjunto de treino.

Comumente, usa-se o MSE (Mean Squared Error):

$$MSE = M(f, D) = \frac{1}{m}\sum_{1}^{m}(y_i - \mathrm{f}(\mathbf{x}_i))^2$$

como métrica de desempenho, mas outra medida também é muito usada: o $RMSE = \sqrt{MSE}$ e o

3.3 Aprendizado de Máquina Explicável

O Aprendizado de Máquina Explicável é um sub-campo de Aprendizado de Máquina preocupado em explicar as decisões de modelos de aprendizado de máquina. A seguir dar-se mais detalhes sobre o campo.

Métodos de aprendizado de máquina, especialmente com o surgimento das redes neurais (RNs), são hoje amplamente utilizados em aplicações comerciais. Esse sucesso levou a uma aceitação considerável do aprendizado de máquina (ML) em muitas áreas científicas. Normalmente, esses modelos são treinados no que diz respeito à alta precisão, mas há uma alta demanda recente e contínua para entender a maneira como um modelo específico opera e as razões subjacentes às decisões tomadas pelo modelo (ROSCHER et al., 2020). Para construir modelos de ML dignos de confiança humana, os pesquisadores propuseram uma variedade de técnicas para explicar os modelos de ML às partes interessadas. Considerada "explicabilidade", este conjunto de trabalhos anteriores tenta iluminar o raciocínio usado pelos modelos de ML.

"Explicabilidade" refere-se vagamente a qualquer técnica que ajude o usuário ou desenvolvedor de modelos de ML a entender por que os modelos se comportam da maneira como se comportam (LUNDBERG et al., 2018). As explicações podem vir de várias formas: desde dizer aos pacientes quais sintomas eram indicativos de um diagnóstico específico até ajudar os trabalhadores da fábrica a analisar ineficiências em um pipeline de produção.

Os usuários, no entanto, geralmente não estão preparados para entender como os dados brutos e o código se traduzem em benefícios ou danos que podem afetá-los individualmente (DHURANDHAR et al., 2018). Ao fornecer uma explicação de como o modelo tomou uma decisão, as técnicas de explicabilidade buscam fornecer transparência direcionada diretamente aos usuários humanos, muitas vezes com o objetivo de aumentar a confiabilidade (O'NEILL, 2018). A importância da explicabilidade como um conceito foi refletida em diretrizes legais e éticas para dados e aprendizado de máquina.

Com o crescente interesse em fornecer explicações de modelos de ML para usuários humanos, a explicabilidade tornou-se um importante subcampo do ML (SELBST; BAROCAS, 2018). Apesar de uma literatura crescente, tem havido poucos trabalhos caracterizando como as explicações estão sendo implantadas pelas organizações no mundo real.

Os atuais sistemas de inteligência artificial (IA) baseados em aprendizado de máquina se destacam em muitos campos. Eles não apenas superam os humanos em tarefas visuais complexas, mas também se tornou uma parte indispensável de todos os nossos dia a dia, por exemplo, como câmeras de telefones celulares inteligentes que podem reconhecer e rastrear faces, como serviços online que podem analisar e traduzir textos escritos, ou como dispositivos de consumo que podem entender a fala e gerar respostas (SAMEK et al., 2019).

Porém, nem sempre a IA se encontra de forma que pode ser facilmente entendida pelos seres humanos, principalmente com a tendência do uso de bases de dados maiores e algoritmos mais complexos. Portanto, estudiosos da IA precisam praticar a criação de algoritmos que possam ser explicados e entendidos pelos humanos, surgindo assim a explainable AI - ou Inteligência Artificial Explicável.

Há uma dimensão social das explicações. Explicando a razão por trás das decisões de alguém é uma parte importante das interações humanas. Explicações ajudam a construir a confiança em um relacionamento entre os seres humanos e, portanto, devem também fazer parte das interações homem-máquina. As explicações não são apenas parte inevitável da aprendizagem e educação humana (por exemplo, o professor explica a solução ao aluno), mas também favorecem a aceitação de decisões difíceis e são importantes para consentimento informado (por exemplo, médico explicando a terapia ao paciente) (SAMEK et al., 2019).

Assim, explicações tornam algoritmos mais confiáveis e aumentam a sua praticidade,

além de melhorar na tomada de decisões e gerar segurança diante de eventuais mudanças necessárias por serem mais facilmente compreensíveis, portando, uma IA que interage principalmente com dados que precisam ser trabalhados com ética e transparência precisam ser explicáveis.

Percebe-se que explicações para predições de redes neurais se encontram cada vez mais necessárias. Com base nisso, este trabalho visa identificar enviesamentos utilizando um método de explicação do Gradient x input, ele será abordado a seguir.

3.3.1 Métodos de Atribuição baseados em Gradiente

São métodos de explicação destinados a explicar modelos de aprendizado de máquina, produzindo explicações chamadas de atribuições. As atribuições podem ser feitas a partir de informações de gradientes. Neste trabalho, foca-se no método Gradient x Input.

3.3.1.1 Gradient x Input

Gradient x Input (SHRIKUMAR; GREENSIDE; KUNDAJE, 2017) é um método de atribuição que usa a informação do gradiente de uma função para demonstrar se uma variável de entrada é relevante: se ela é diferente de zero e se o modelo reage positivamente a eles. Isso é importante pois o gradiente pode mudar rapidamente, mesmo após pequenas mudanças no espaço de entrada. As explicações são obtidas calculando o produto da derivada parcial avaliada localmente a ativação da entrada:

$$x_i \times \vec{\nabla} f(x_i)$$

em que f é a rede neural, i se refere ao cientista alvo e x_i é o vetor de entrada referente ao cientista i.

3.4 Redes Neurais Profundas

Computadores aprendem com a experiência e compreendem o mundo em termos de uma hierarquia de conceitos (GOODFELLOW; BENGIO; COURVILLE, 2016), nele a máquina aprende conceitos complicados construindo-os a partir de conceitos mais simples. Se desenharmos um gráfico mostrando como esses conceitos foram construídos uns sobre os outros, o gráfico é profundo, com muitas camadas. Por esse motivo, chamamos essa abordagem de aprendizado profundo de IA.

Para Haykin (2009), o poder computacional de uma rede neural é medida na estrutura altamente interligada da rede, que permite imitar os processos de sinalização dos neurônios e na sua habilidade de aprendizagem pelo modelo de predição, o que significa que

uma rede treinada pode classificar dados da mesma classe que os dados de aprendizado que nunca viu antes e gerar saídas congruentes. Assim, com essa qualidade de processamento de informação as redes neurais são capazes reconhecer padrões e resolverem problemas incomuns nas áreas de IA, aprendizado de máquina e aprendizado profundo.

Uma Rede Neural profunda, com d camadas escondidas, consiste de d matrizes $A_1, A_2, \cdots, A_d$ e uma função específica $\sigma : \mathbb{R} \mapsto \mathbb{R}$ chamada não-linearidade. A não-linearidade mais utilizada nos dias atuais é a função *rectilinear linear*, $\text{RELU}_b = \max\{0, x - b\}$. Nesta função b é chamado *bias* e é também um parâmetro da rede juntamente com as matrizes $A_1, A_2, \cdots, A_d$. Definindo $y^0 = x^0$, essa rede computa $y^1, y^2, \cdots, y^d$ em que $y^i + 1 = \sigma(A_i y^i)$. A função $\sigma(z)$ denota o vetor obtido aplicando σ a cada coordenada de z. Cada coordenada de um vetor computado y^i representa um nó da rede e cada entrada das matrizes $A_1, A_2, \cdots, A_d$ relaciona-se a uma aresta. A saída da rede é y^d. O tamanho da rede é o número de nós nela. O número de parâmetros é o número de arestas mais o número de nós.

Uma Rede Neural Profunda, portanto, é uma função que mapeia o vetor x^0 para o vetor saída $y^d = f_{A_1, A_2, \cdots, A_d, \vec{b}}(x^0)$.

Portanto, dado um conjunto de dados $D = \{(\mathbf{x}_1, y_1); (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, em que y_i é a saída para a entrada $\mathbf{x}_i = \{x_{i1}; x_{i2}, \cdots, x_{im}\}$, $y_i \in \mathbb{R}$. O método de aprendizado supervisionado tenta aprender a função:

$$f_{A_1, A_2, \cdots, A_d, \vec{b}}(\mathbf{x}), \text{ tal que } f(\mathbf{x}_i) \simeq y_i, \ i = 1, 2, \cdots, m$$

Os parâmetros aprendidos, as matrizes e o vetor de *bias*, não dizem nada sobre a contribuição de cada variável de entrada para uma saída da rede. Por isso a necessidade de métodos de entendimento para desvendar essa relação.

3.5 Predição do índice-h futuro do cientista

O índice-h futuro (ACUNA; ALLESINA; KORDING, 2012) deve ser definido em termos das muitas diferentes características ligadas a carreira do cientista. Uma abordagem de aprendizado de máquina supervisionada é usada para definir o problema.

Sejam $\{x_1, x_2, \cdots, x_n\}$ as entradas para um algoritmo de aprendizado de máquina e $\{y_1, y_2, \cdots, y_n\}$ sejam os alvos, em que $x_i \in \mathbb{R}^d$ representa as d características do cientista i relacionadas a sua carreira, e $y_i \in \mathbb{R}$ o valor do índice-h futuro, calculada Δt anos depois do momento da predição.

O algoritmo tem como *tarefa de aprendizado* (objetivo), aprender uma função aproximada f para estimar y_i dado x_i e Δt (Equação 3.1). A função f deve ser avaliada

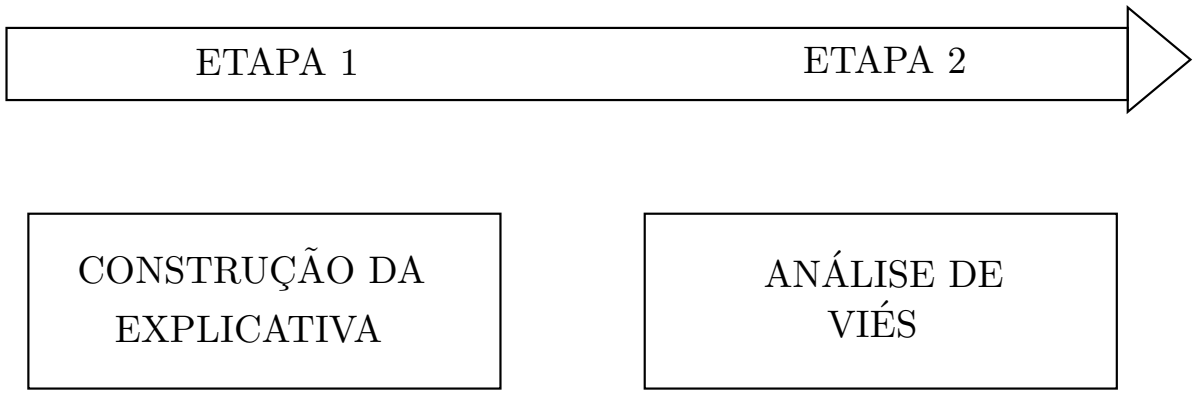
sobre novos exemplos. E, deve ser possível extrair de f a contribuição de cada característica para uma predição.

$$f(y_i|x_i, \Delta t) \approx y_i \tag{3.1}$$

4 Método

Este capítulo apresenta o método. As etapas principais do método estão resumidas no diagrama da Figura 1. Cada etapa é detalhada nas seções seguintes.

Figura 1 – Fluxograma da etapas principais do método.



4.1 Etapa 1

Para estabelecer a relação entre as características do cientista $(x_1, x_2 \cdots, x_d)$ e seu índice-h futuro y, usou-se o método $GradientXInput$/LRP que atribui uma pontuação, algumas vezes também chamada de relevância ou contribuição, para cada variável de entrada. Como indicado na Figura 2, o método produz modelos lineares servindo de explicativas, em que a predição da rede neural y é a variável de resposta do modelo e o coeficiente de cada variável independente representa a sua relevância para a decisão particular.

4.2 Etapa 2

Para visualização de vieses, usou-se boxplots comparativos. As intensidades das contribuições das variáveis de entrada para a decisão da rede neural são comparadas para cada grupo de cientista - eminentes e anônimos. Usa-se uma abordagem visual, por meio de boxplots, para visualização das diferenças.

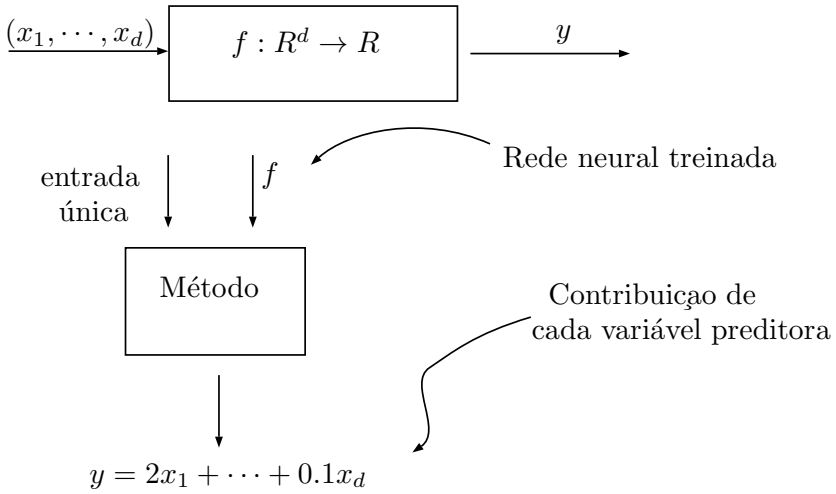


Figura 2 – Método de atribuição produzindo um modelo linear servindo como explicativa do porquê da predição para uma entrada particular.

5 Resultados e Discussão

5.1 Configuração Experimental

5.1.1 Seleção da Amostra e Arquitetura da Rede

Grupos de diferentes status foram selecionados aleatoriamente. Cientistas dos dois grupos publicaram em muitos periódicos diferentes, mas os do grupo 1 tem uma rede de colaboradores tímida enquanto os do grupo 2, uma rede densa.

A Figura 3 apresenta a arquitetura da rede neural. A rede possui 4 camadas, usando a função de ativação ReLU (do ingês, Rectified Linear Activation Function), escolhida por ser um modelo fácil de treinar e de bom desempenho. Foi configurada com um batch size igual a 50. O algoritmo pega as primeiras 50 amostras do conjunto de dados de treinamento e treina a rede. Em seguida, ele pega uma outra parte de 50 e treina a rede novamente. Adotou-se uma abordagem hold-out, 80 por cento da base foi utilizada para treino e 20 para os testes. Após testes, foi verificado que 100 épocas já era necessário para mostrar a evolução do problema.

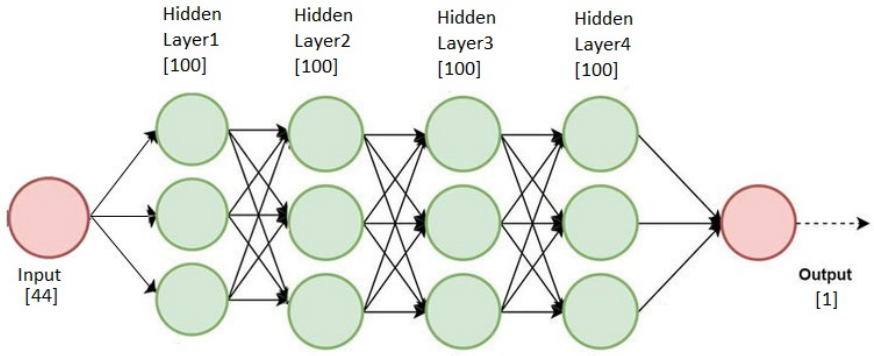


Figura 3 – Arquitetura da rede.

5.1.2 Conjunto de Dados

Neste trabalho, nosso método será testado em um conjunto de dados de quase 4 milhões de artigos científicos, escritos por mais de 700 mil autores, a maioria deles da Ciência da Computação, e publicados em quase 7 mil conferências e periódicos, cobrindo tópicos, sobretudo, da área da Ciência da Computação.

Weihs e Etzioni (2017) têm tornado público em um repositório[1], uma versão pré

[1] https://github.com/Lucaweihs/impact-prediction, acessado em 22 de julho de 2022.

Tabela 1 – Descrição das Características do Cientista usadas para predizer seu índice-h futuro.

Nome da Característica	**Descrição**
author_hindice	índice-h
author hindex delta	Mudança no índice h nos últimos dois anos
author citation count	Contagem cumulativa de citações
author key citation count	Contagem cumulativa de citações de chave
author citations delta 0,1	Citações este ano e um ano atrás
author key citations delta 0,1	Principais citações este ano e um ano atrás
author mean citations per paper	Número médio de citações por artigo
author mean citation per paper delta	Mudança na média de citações por artigo nos últimos dois anos
author mean citations per year	Número médio de citações por ano
author papers	Número de artigos publicados
author papers delta	Número de artigos publicados nos últimos dois anos
author mean citation rank	Classificação do autor (entre 0 e 1) entre todos os outros autores
author unweighted pagerank	PageRank do autor na rede de coautoria não ponderada
author weighted pagerank	PageRank do autor na rede de coautoria ponderada
author age	Duração da carreira (anos desde o primeiro artigo publicado)
author recent num coauthors	Número total de coautores nos últimos dois anos
author max single paper citations	Número máximo de citações para qualquer artigo do autor
venue hindex mean, min,max	Índices H de locais que o autor publicou
venue hindex delta mean, min,max	Alteração do índice h de 2 anos para locais que o autor publicou
venue citations mean, min,max	Média de citações por artigo de locais que o autor publicou
venue citations delta mean, min,max	Mudança na média de citações por artigo nos últimos dois anos para locais que o autor publicou
venue papers mean, min, max	Número de artigos em locais em que o autor publicou
venue papers delta mean, min, max	Mudança no número de trabalhos em locais em que o autor publicou nos últimos dois anos
venue rank mean, min, max	Ranks de locais (entre 0-1) em que o autor publicouestabelecido determinado pelo número médio de citações por artigo
venue max single paper citations mean, min, max	Número máximo de citações de qualquer artigo publicado em um local recebeu para cada local que o autor publicou
total num venues	Número total de locais publicados

processada deste conjunto de dados, que utilizaram para predizer medidas de impacto baseadas em citações. Neste trabalho, inicialmente, esta versão pré-processada será adotada. Futuramente, novas conjuntos de dados serão incluídos. Como eles fizeram, as 44 características referente ao cientista usadas para predizer suas medidas de impacto serão usadas também aqui. A Tabela 1 lista elas.

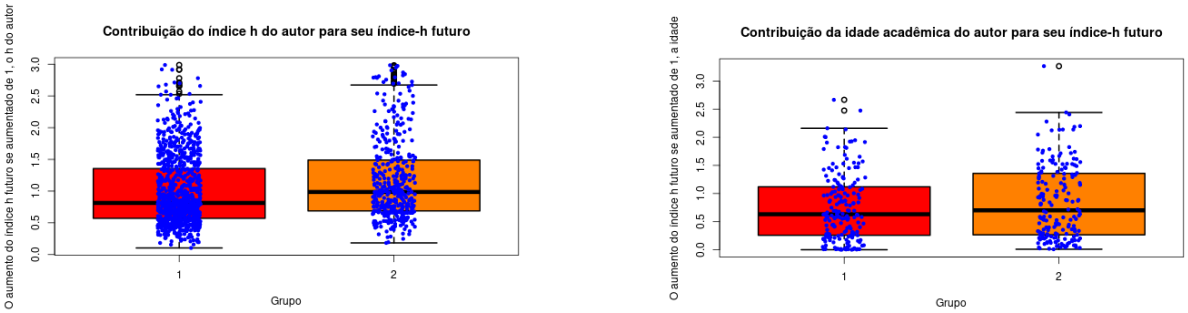
5.2 Resultados

Para avaliar o desempenho do modelo nos dois grupos, a raiz quadrada do erro quadrático médio (RMSE do inglês Root Mean Squared Error) foi usada. A Tabela 2 mostra o desempenho nos diferentes conjuntos. Um erro muito acentuado já evidencia algum prejuízo para o grupo 1.

Tabela 2 – Desempenho do Modelo.

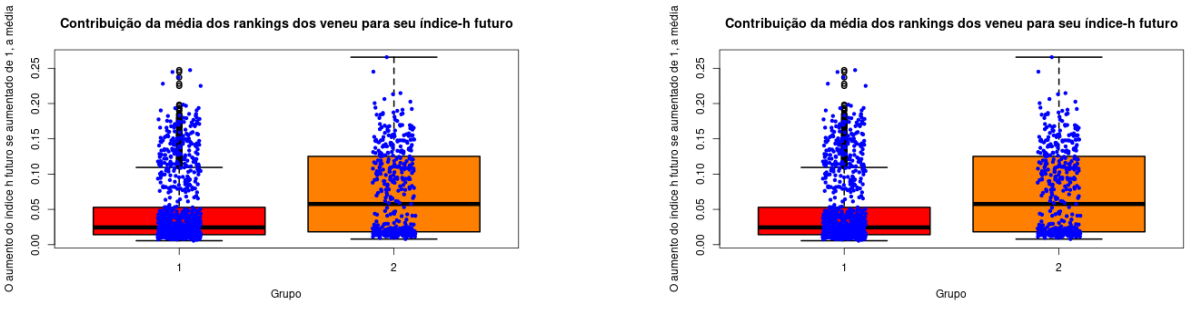
Conjunto	RMSE
de Teste	1.35
de Treino	1.33
Grupo1	4.11
Grupo2	2.6

Os boxplots comparativos nas Figuras 4 e 5, claramente mostram pontuações de valores superestimados para cientistas do grupo 2. O método *Gradient x input* atribui pontuações desproporcionais para cientistas desse grupo, caracterizando favorecimento.



(a) Índice-h (b) Idade acadêmica

Figura 4 – Grupo dois recebe recompensa maior. Ao contrário do grupo 1, este grupo tem uma rede de colaboradores densa.



(a) Venue (b) Idade acadêmica

Figura 5 – Grupo dois recebe recompensa maior. Ao contrário do grupo 1, este grupo tem uma rede de colaboradores ampla.

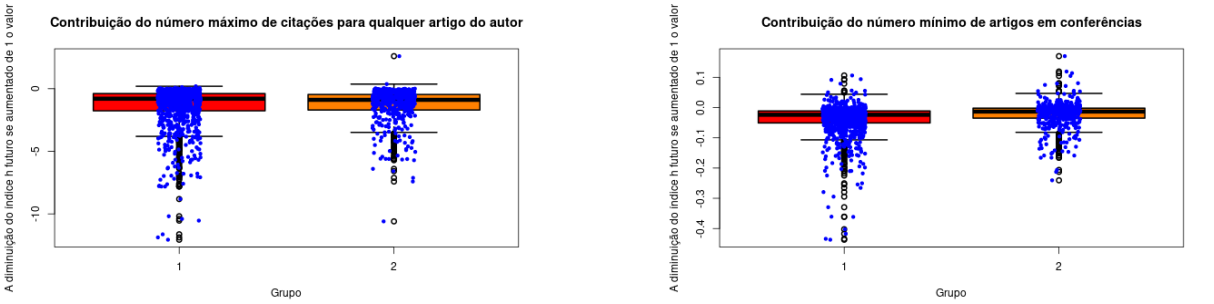
A Tabela 3 compara a importância das características provida pelo método *Gradient x Input* para um cientista de cada grupo. Os dados da Tabela confirmam valores muito

elevados, e sem uma clara justificativa, para o cientista que já tem muito (cientista do grupo 2).

Tabela 3 – Comparativo de dois cientistas.

	Contribuição para o índice-h futuro	
Métricas	**Cientista do grupo 1**	**Cientista do grupo 2**
author_hindex	0.15623327	1.3487122
author_hindex_delta	0	0.24888946
author_citation_count	0.08490858	0.55009615
author_key_citation_count	0	0.051654607
author_citations_delta_1	0	0.43979183
author_citations_delta_0	0.09972202	2.1232595
author_key_citations_delta_1	0	0
author_key_citations_delta_0	0	0.026112486
author_mean_citations_per_paper	-0.0036064207	-0.047694266
author_mean_citation_per_paper_delta	0.0043373895	-0.033722505
author_mean_citations_per_year	0.010864068	0.77740014
author_papers	0.6169457	-0.4578309
author_papers_delta	3.550577	0.8978659
author_mean_citation_rank	3.9193685e-05	0.040690333
author_unweighted_pagerank	0.05341546	0.07003153
author_weighted_pagerank	0.051683865	0.074496046
author_age	-0.6329136	0.18181466
author_recent_num_coauthors	0.92988664	0.10289526
author_max_single_paper_citations	-0.009810526	-0.29278713
venue_hindex_max	-0.060999997	0.14888383
venue_hindex_min	-0.00016569924	0.022596203
venue_hindex_mean	0.09722936	0.5113547
venue_hindex_delta_max	0.061802566	0.14412118
venue_hindex_delta_min	0	0.044232685
venue_hindex_delta_mean	0.05444467	0.19394426
venue_citations_max	0.043207914	0.016315954
venue_citations_min	0.00042802593	0.0006928835
venue_citations_mean	0.01865556	0.21309294
venue_citations_delta_max"	0.016885744	0.040419284
venue_citations_delta_min	-0.0072180727	7.674358e-05
venue_citations_delta_mean	0.0014046615	0.041313987
venue_papers_max	-0.66313314	0.154364
venue_papers_min	-0.045256943	-0.009888934
venue_papers_mean	2.4030704	0.8855413
venue_papers_delta_max	-1.1745499	-0.078069955
venue_papers_delta_min	0	0
venue_papers_delta_mean	0.25882375	-0.038642164
venue_rank_max	0.0099225305	0.12387629
venue_rank_min	0.000459027	0.0005179575
venue_rank_mean	0.006828638	0.092050016
venue_max_single_paper_citations_max	-0.18594588	1.6100341
venue_max_single_paper_citations_min	-0.002636009	-0.0007255984
venue_max_single_paper_citations_mean	0.17989683	1.352639
totalNumVenues	-0.2522878	0.593299

O resultado da Tabela 6 mostra que as discriminações podem ser ainda mais perversas, quando a característica tem o efeito de diminuir o valor da predição, a diminuição do índice-h futuro é bem menor para cientistas do grupo 2.



(a) Número máximo de citações do seu melhor artigo.

(b) Número mínimo de artigos que já publicou.

Figura 6 – Grupo dois é ainda favorecido, ainda que a contribuição tenha um efeito de diminuição do valor do índice-h futuro do cientista.

Explicações para predições de redes neurais estão cada vez mais se tornando um fator vital para muitas aplicações. Este trabalho, apresenta uma abordagem nova para identificar enviesamentos em explicações providas pelo método de explicação do Gradient x input. Nossa abordagem, apesar de simples, representa um avanço sobre o nosso conhecimento do assunto. Pela primeira vez, provemos evidências de favorecimentos nessas explicações.

Como indicado por Bornmann et al. (2020), em um trabalho anterior, a evidência que nós encontramos aponta para uma atribuição de pontuações desproporcionais, dada pelo método do Gradient x input, para cientistas em posições privilegiadas, confirmando nossa suspeita (de favorecimento).

Inevitavelmente, existem alguns problemas devido a seleção da amostra. Nossa seleção foi randômica e aplicou-se uma visão estreita. Em trabalho futuro, pretende-se estender nossas análises para mais grupos.

6 Conclusão

Neste trabalho, é apresentado um novo método para identificar as pontuações desproporcionais atribuídas a cientistas em posições privilegiadas pelo método de explicação *gradient x input*. Tem-se obtido resultados preliminares demonstrando que as explicações providas têm favorecido cientistas com status elevados, e consequentemente, tem ampliado as diferenças de impacto e produtividade já existentes entre cientistas com status de estrelas e aqueles em minoria nos espaços da ciência.

Este estudo é o primeiro passo para melhorar o nosso entendimento sobre o assunto e aliviar essa questão. No entanto, mais trabalho focado e aprofundado precisa ser feito.

Devido ao pioneirismo deste trabalho, espera-se contribuir para o início de novas abordagens do tema. Além do que foi realizado, espera-se prosseguir nas seguintes melhorias:

- Adquirir novas bases de dados para reforçar nossas descobertas.
- Demostrar outros métodos que discriminam entre cientistas respeitados e não respeitados.
- Aprofundar/testar com outros modelos (arquiteturas de redes neurais) utilizando outros métodos de atribuição baseado em gradiente.
- Atingir melhorias na prova do método proposto.

Referências

ACUNA, D. E.; ALLESINA, S.; KORDING, K. P. Predicting scientific success. *Nature*, v. 489, n. 201, p. 201–202, 2012. Citado 2 vezes nas páginas 15 e 20.

BIANCHINI, S.; MüLLER, M.; PELLETIER, P. Artificial intelligence in science: An emerging general method of invention. *Research Policy*, v. 51, n. 10, p. 104604, 2022. Citado na página 13.

BOL, T.; VAAN, M. de; RIJT, A. van de. The matthew effect in science funding. *Proceedings of the National Academy of Sciences*, v. 115, n. 19, p. 4887–4890, 2018. Citado na página 13.

BORNMANN, L.; GANSER, C.; TEKLES, A.; LEYDESDORFF, L. Does the halpha-index reinforce the matthew effect in science? the introduction of agent-based simulations into scientometrics. *Quantitative Science Studies*, v. 1, n. 1, p. 331–346, 2020. Citado 2 vezes nas páginas 15 e 28.

DHURANDHAR, A.; SHANMUGAM, K.; LUSS, R.; OLSEN, P. A. Improving simple models with confidence profiles. *Advances in Neural Information Processing Systems*, v. 31, 2018. Citado na página 18.

GIERYN, T. F. *Science and Social Structure: A Festschrift for Robert K. Merton*. [S.l.]: New York Academy of Sciences, 1980. Citado na página 16.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. Citado na página 19.

HAYKIN, S. S. *Neural networks and learning machines*. Third. Upper Saddle River, NJ: Pearson Education, 2009. Citado na página 19.

KÖCHLING, A.; WEHNER, M. C. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development. *Business Research*, v. 13, n. 3, p. 795–848, Nov 2020. Citado na página 13.

LUNDBERG, S. M.; NAIR, B.; VAVILALA, M. S.; HORIBE, M.; EISSES, M. J.; ADAMS, T.; LISTON, D. E.; LOW, D. K.-W.; NEWMAN, S.-F.; KIM, J. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, Nature Publishing Group, v. 2, n. 10, p. 749–760, 2018. Citado na página 18.

MERTON, R. K. The matthew effect in science. *Science*, American Association for the Advancement of Science, v. 159, n. 3810, p. 56–63, 1968. Citado 2 vezes nas páginas 13 e 16.

MITCHELL, T. M. *Machine Learning*. 1. ed. USA: McGraw-Hill, Inc., 1997. ISBN 0070428077. Citado na página 17.

O'NEILL, O. Linking trust to trustworthiness. *International Journal of Philosophical Studies*, Taylor & Francis, v. 26, n. 2, p. 293–300, 2018. Citado na página 18.

ROSCHER, R.; BOHN, B.; DUARTE, M. F.; GARCKE, J. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, IEEE, v. 8, p. 42200–42216, 2020. Citado na página 17.

SAMEK, W.; MONTAVON, G.; VEDALDI, A.; HANSEN, L. K.; MÜLLER, K. (Ed.). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. [S.l.]: Springer, 2019. v. 11700. (Lecture Notes in Computer Science, v. 11700). Citado na página 18.

SELBST, A. D.; BAROCAS, S. The intuitive appeal of explainable machines. *Fordham L. Rev.*, HeinOnline, v. 87, p. 1085, 2018. Citado na página 18.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning important features through propagating activation differences. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. [S.l.]: JMLR.org, 2017. (ICML'17), p. 3145–3153. Citado 2 vezes nas páginas 13 e 19.

SILVA, J. A. Teixeira da. The matthew effect impacts science and academic publishing by preferentially amplifying citations, metrics and status. *Scientometrics*, Springer-Verlag, Berlin, Heidelberg, v. 126, n. 6, p. 5373–5377, jun 2021. Citado na página 15.

WANG, D.; BARABáSI, A.-L. The matthew effect. In: _____. *The Science of Science*. [S.l.]: Cambridge University Press, 2021. p. 28–38. Citado na página 16.

WEIHS, L.; ETZIONI, O. Learning to predict citation-based impact measures. In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. [S.l.: s.n.], 2017. p. 1–10. Citado na página 24.

ZENG, A.; FAN, Y.; DI, Z.; WANG, Y.; HAVLIN, S. Impactful scientists have higher tendency to involve collaborators in new topics. *Proceedings of the National Academy of Sciences*, v. 119, n. 33, p. e2207436119, 2022. Citado na página 15.