UNIVERSIDADE FEDERAL DO MARANHÃO

Curso de Ciência da Computação

José Emanuel Passos Barros

Um Novo Modelo de Rede Neural Interpretável para Predizer o Índice-h de Cientistas

São Luís - MA

2023

José Emanuel Passos Barros

Um Novo Modelo de Rede Neural Interpretável para Predizer o Índice-h de Cientistas

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. Antônio de Abreu Batista Júnior

São Luís - MA

2023

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

Passos Barros, José Emanuel.
Um Novo Modelo de Rede Neural Interpretável para Predizer o Índice-h de Cientistas / José Emanuel Passos Barros. - 2023.
43 p.

Orientador(a): Antônio de Abreu Batista Júnior.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, Universidade Federal do Maranhão, 2023.

1. IA Explicável. 2. Índice-H. 3. LIME. 4. Redes Neurais Artificiais. 5. SHAP. I. Batista Júnior, Antônio de Abreu. II. Título.

José Emanuel Passos Barros

Um Novo Modelo de Rede Neural Interpretável para Predizer o Índice-h de Cientistas

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.

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

Prof. Dr. Carlos de Salles Soares Neto
Examinador Interno
Universidade Federal do Maranhão

Prof. Dr. Diogo Fernando Trevisan
Examinador Externo
Universidade Estadual de Mato Grosso do Sul

São Luís - MA
2023

Dedico à Deus, dono de toda ciência, sabedoria e poder e à memória da minha querida mãe, Jaciléa Silva Passos, por ter me feito quem eu sou.

Agradecimentos

Gostaria de expressar meus agradecimentos a todos que estiveram ao meu lado, suas contribuições, apoio e incentivo foram fundamentais até aqui.

Agradeço primeiramente a Deus, pois toda conquista é concedida por intermédio dEle, até aqui trilhou o caminho ao meu lado e me sustentou com força e coragem em cada adversidade.

Quero agradecer à minha família, a minha avó Ana maria, tios: Adélia, Adriana, Auriléa, Mauro e Clauber, irmãos: Felipe e Mirela, e em memória a minha mãe Jaciléa e meu avô José Ribamar. Por todo o amor, compreensão e encorajamento incondicional que sempre me ofereceram.

À minha noiva, Karla Felícia, minha amada e companheira de todas os momentos, meu agradecimento por estar ao meu lado durante toda a graduação. Foi quem me motivou a seguir confiante no curso. Sua presença torna cada momento muito mais especial e significativo.

Ao meu orientador e coordenador do curso, Prof. Dr. Antônio de Abreu Batista Jr, que me guiou ao longo deste trabalho com sua experiência, obrigado pela atenção, paciência e auxílio prestados em meus momentos de dificuldade.

Aos meus amigos Antoniel, Emival, Daniel, Elton, Felipe Iran, Matheus Boaro, Sousa Renan e demais companheiros de IFMA, que sempre estiveram presentes para me apoiar, minha gratidão pela amizade e pela companhia em momentos importantes.

Aos meus amigos, com os quais trabalhei e trabalho, no Hospital Sarah: Fábio Costa, Fábio Augusto, Marcelo Fontenelle, José Carlos, Flávio Henrique, Laryssa Ribeiro, Paulo Victor, Alexandre Braule e Jadiel Costa. Afonso Lopes, Claudio Aires, Matheus Gomes e todos da UTIC - SEBRAE. Obrigado pelos conhecimentos e experiências passadas que contribuíram no meu crescimento profissional.

A Prof. Lucinete Marques, Prof. Ilma, Prof. Lélia, Andréa Lima e todos os demais professores e assistentes do PPGE - UFMA, pelas experiências e conhecimentos acadêmicos que me proporcionaram.

Agradeço também a todos professores do curso de ciência da computação, que compartilharam seu conhecimento e experiência em sala de aula e amigos que tive a oportunidade de conhecer e estar junto ao longo da graduação em especial da Átletica Lorde. E, por fim, a todos que me ajudaram, de alguma forma, ao longo dessa trajetória, meu sincero reconhecimento. Cada palavra de incentivo, cada gesto de apoio, cada colaboração e conselho contribuíram no meu aprendizado pessoal e profissional.

"A ciência não pode prever o que vai acontecer. Só pode calcular a probabilidade de alguma coisa acontecer".

(César Lattes)

Resumo

Os modelos de aprendizado de máquina, como as redes neurais, tem desempenhado um papel crucial em diversos campos da ciência. Um deles é trazendo previsões altamente precisas sobre os alcances de longo prazo de cientistas. No entanto, apesar do seu amplo sucesso em suas predições, tais modelos sofrem de uma grande fraqueza, a falta de explicabilidade ou transparência de suas decisões e ações autônomas. Atualmente, devido à não compreensão do raciocínio por trás de uma decisão destes modelos, cientistas com menor status podem estar sendo prejudicados. Neste trabalho, propomos uma Rede Neural Artificial (RNA) para predizer o índice-h de pesquisadores e tornamos as suas decisões acessíveis à avaliadores humanos usando os explicadores LIME e SHAP. Este trabalho busca demonstrar que a utilização de um modelo explicável é uma decisão importante e, de fato, fundamental para confiarmos em suas decisões. Nossos experimentos demonstram que as explicativas providas podem auxiliar os tomadores de decisão a tomar decisões mais informadas e a detectar vieses propagados pelo modelo. Futuramente, pretende-se implantar este modelo em uma Aplicação Web para auxiliar pesquisadores, universidades e agências de pesquisa em suas atividades diárias.

Palavras-chave: Redes Neurais Artificiais, Aprendizado Profundo, Redes Acadêmicas, LIME, SHAP, IA Explicável, Índice-H.

Abstract

Machine learning models, such as neural networks, have played a crucial role in many fields of science. One is bringing highly accurate predictions about the long-term achievements of scientists. However, despite their widespread success in their predictions, such models suffer from a major weakness, the lack of explainability or transparency of their autonomous decisions and actions. Currently, due to the lack of understanding of the reasoning behind a decision made by these models, scientists with lower status may be being harmed. In this work, we propose an Artificial Neural Network (ANN) to predict researchers' h-index and make its decisions accessible to human evaluators using LIME and SHAP explainers. This work seeks to demonstrate that the use of an explainable model is an important decision and, in fact, fundamental for us to trust its decisions. Our experiments demonstrate that the provided explanations can help decision makers to make more informed decisions and to detect biases propagated by the model. In the future, it is intended to implement this model in a Web Application to help researchers, universities and research agencies in their daily activities.

Keywords: Artificial Neural Networks, Deep Learning, Academic Networks, LIME, SHAP, Explainable AI, H-index.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

RNA	*Redes Neurais Artificiais*
MLP	*Multilayer Perceptron*
ReLU	*Rectified Linear Unit*
MSE	*Mean Squared Error*
MAE	*Mean Absolute Error*
IA	*Inteligência Artificial*
XAI	*eXplainable Artificial Inteligence*
LIME	*Local Interpretable Model-Agnostic Explanations*
AM	*Aprendizado de Máquina*

Sumário

1 Introdução

A predição do impacto futuro de pesquisadores(CLAUSET; LARREMORE; SINATRA, 2017; HOU et al., 2019; SINATRA et al., 2016) tem atraído um interesse considerável nos últimos anos. Ela interessa à agências de governo, instituições de pesquisa, e aos próprios cientistas que estão sempre orientando-se por ela.

Tradicionalmente o impacto predito do cientista tem sido medido por meio do seu índice-h estimado para anos mais tarde (5, 10 ou 15 anos). O índice-h de Hirsch (HIRSCH, 2005) é uma índice que ajuda a qualificar o impacto dos trabalhos publicados por um pesquisador e se este impacto se concentra em poucos ou vários trabalhos. O índice depende do total de publicações de um cientista e do impacto das suas publicações.

O índice-h é frequentemente usado para classificar pesquisadores em processos de seleção, distribuição de recursos de pesquisa, promoções acadêmicas e contratação. A racionalidade por trás do uso desses indicadores como ferramentas de suporte à decisão, nesses contextos, é a força preditiva que supostamente eles carregam (LINDAHL, 2020). Para muitas aplicações, o potencial impacto futuro do avaliado é a preocupação central (PENNER et al., 2013). Os modelos de aprendizado de máquina do índice-h futuro (ACUNA; ALLESINA; KORDING, 2012) inferido por algoritmos de Aprendizado de Máquina, e capturando o potencial impacto futuro do cientista, claramente têm uma vantagem sobre valores correntes destes indicadores tradicionais. Entretanto, vieses de toda natureza (e.g., favorecimento de certos grupos privilegiados) encontrados nesses modelos e a falta de entendimento humano de como eles chegam a uma decisão têm tornado o uso deles inadequados nestes contextos.

Neste trabalho propõe-se um modelo de rede neural do índice-h futuro cujas as suas decisões são acessíveis à humanos. Ao fornecer uma explicação do seus índices-h, os pesquisadores podem entender como seu desempenho é avaliado em relação a seus pares e garantir que a comparação seja feita de forma equitativa. Além disso, os pesquisadores são incentivados a trabalhar para melhorar seu desempenho acadêmico. Eles podem usar as informações para ajustar suas estratégias de publicação, colaboração e divulgação, buscando aumentar sua visibilidade e impacto na comunidade acadêmica.

1.1 Objetivos

A seguir serão apresentados os objetivos do trabalho.

1.1.1 Objetivo Geral

Este trabalho visa desenvolver um modelo de rede neural do índice-h futuro entendível por humanos a ser implantado em uma Aplicação Web para auxiliar administradores da ciência a entenderem o índice-h predito (o sucesso esperado) de pesquisadores.

1.1.2 Objetivos Específicos

Como objetivos específicos: (1) construir um modelo de regressão de rede neural artificial do índice-h futuro do pesquisador usando a biblioteca Keras; (2) Criar explicativas usando técnicas de explicação populares, como LIME (Explicações agnósticas de modelos interpretáveis locais) e SHAP (explanações aditivas SHApley); e (3) Validar o modelo do índice-h futuro usando inferência lógica a partir das explicativas de técnicas de explicação para uma decisão do modelo.

Os resultados bem-sucedidos do projeto fornecerão uma oportunidade de implantar o modelo em uma Plataforma Web para aprimorar os processos de avaliação acadêmica e apoiar os pesquisadores a alcançar todo o seu potencial. Como integrará poderosas técnicas de aprendizado de máquina com uma interface amigável, o sistema preenche a lacuna entre algoritmos complexos e tomada de decisão prática, tornando-se uma ferramenta valiosa para o meio acadêmico e gestão de financiamento de pesquisa. No geral, a aplicação web representa uma contribuição promissora para o avanço acadêmico, promovendo transparência, justiça e tomada de decisão baseada em evidências na avaliação acadêmica e na alocação de financiamento de pesquisa.

2 Fundamentação Teórica

Neste capítulo, será abordado os conceitos teóricos fundamentais utilizados para o experimento realizado, além de uma breve amostra sobre os métodos e ferramentas que foram úteis para conduzir este trabalho. Para o experimento se fez necessário entender regressão e problemas de regressão, além de compreender o Aprendizado de Máquina e quando ela pode ser Aprendizado de Máquina Explicável (XAI). Ao longo do capítulo, mostraremos como a utilização dos frameworks Lime e SHAP podem ser importantes para oferecer explicações do problema abordado.

Em vista disso, juntando os problemas de entendimento humano e decisões éticas sobre modelos de IA por métodos de avalaições preditivas, o experimento traz soluções ao garantir que os modelos de aprendizado de máquina e algoritmos de análise de dados não perpetuem ou ampliem desigualdades existentes na sociedade mas tornem o processo mais transparente e explicável ao entendimento humano, e assim em paralelo com o avanço da IA, os métdoso de aprendizado de máquina são cada vez mais incorporados em decisões críticas, o que torna essencial que seus desenvolvedores assumam a responsabilidade por garantir que seus modelos sejam entendiveis e éticos em suas operações.

2.1 Redes Neurais Artificiais (RNA)

As redes neurais artificiais que existem atualmente são um resultado de um processo de melhoria ao longo dos anos. O seu inicio se deu após os cientistas Warren McCulloch e Walter Pitts produzirem o artigo "A logical calculus of the ideas immanent in nervous activity", publicado em 1943. Eles notaram que o cérebro humano é um importante exemplo de processamento de informações, então é possível comparar o neurônio biológico com a ideia de um neurônio artificial. Uma rede neural é um processador maciçamente e paralelamente distribuído, constituído de unidades de processamento simples, que têm a propensão natural de armazenar conhecimento experimental e torná-lo disponível para o uso. Ela se assemelha ao cérebro em dois aspectos:

1. O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem.

2. Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido (HAYKIN, 2001)

2.1.1 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 $y^d = f_{A_1, A_2, \cdots, A_d, \overrightarrow{b}}(x^0)$.

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:

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

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.

2.1.2 Multilayer Perceptron feedforward

A ideia dos perceptrons de múltiplas camadas (multilayer perceptron - MLP) foi inicialmente formulada por Bryson e Ho em 1969, ao perceberem que havia limitações na estrutura básica nas redes neurais existentes até então (JR.; HO, 1969).

A diferença pode ser tida da seguinte forma: Na rede de camada única, um conjunto de entradas é mapeado diretamente para uma saída usando uma variação generalizada de uma função linear. Essa instanciação simples de uma rede neural também é chamada de perceptron. Nas redes neurais multicamadas, os neurônios são organizados em camadas, em quais as camadas de entrada e saída são separadas por um grupo de camadas ocultas. Este em camadas A arquitetura da rede neural também é chamada de rede feed-forward (NIELSEN, 2015). Dessa forma, temos a maneira que permite o aprendizado de representações mais complexas e abstratas dos dados, o que favorece para uma aprendizado de máquina mais aplicável em inúmeras áreas e problemas.

2.2 Regressão e problemas de regressão

As redes neurais servem para diversos problemas. Ela será abordada aqui para um caso de regressão. Para problemas de regressão, o funcionamento se dar tendo apenas um neurônio de saída e a rede neural deve aprender a fazer previsões numéricas.

A análise de regressão ocupa-se do estudo da dependência de uma variável, a variável dependente, em relação a uma ou mais variáveis, as variáveis explicativas, com o objetivo de estimar e/ou prever a média (da população) ou o valor médio da dependente em termos dos valores conhecidos ou fixos (em amostragem repetida) das explicativas (GUJARATI, 2000).

Portanto, a regressão é uma técnica estatística que mostra a relação mencionada, normalmente representada pela letra x, e uma ou mais variáveis independentes que normalmente são denominadas y. Com o uso da regressão é possível fazer previsões sobre o futuro a partir dos dados já existentes.

2.2.1 Modelo do índice-h futuro

O índice-h de Hirsch (HIRSCH, 2005) é um índice que ajuda a qualificar o impacto dos trabalhos publicados por um pesquisador e se o impacto se concentra em poucos ou vários trabalhos. O índice depende do total de publicações de um cientista e do impacto das suas publicações. Recentemente, Um modelo linear do índice-futuro (ACUNA; ALLESINA; KORDING, 2012) foi proposto. O valor do índice-h futuro do cientista funciona como um substituto (uma medida aproximada) do seu impacto (sucesso) futuro. Os autores usaram aprendizado supervisionado para aprender um modelo linear em termos das características

do cientista. Neste trabalho, uma abordagem similar é usada, mas devido a um melhor desempenho o modelo aprendido é uma rede neural artificial.

O aprendizado é como segue: em que $\{x_1, x_2, \cdots, x_n\}$ são as entradas para o algoritmo de aprendizado de máquina e $\{y_1, y_2, \cdots, y_n\}$ são 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 objetivo do algoritmo de otimização é aprender uma função aproximada f para estimar y_i dado x_i e Δt (Equação 2.1). A função f deve ser avaliada em dados não vistos antes. E, deve ser possível extrair de f a importância de cada característica individual para uma saída do modelo.

$$f(y_i|x_i, \Delta t) \approx y_i \tag{2.1}$$

2.3 Explicabilidade de modelos de Aprendizado de Máquina (XAI)

As redes neurais também são conhecidas como modelo de caixa preta pois é uma estrutura que pode ser observado em termos de suas entradas e saídas, mas sem revelar seus mecanismos internos. Sua implementação é "opaca"(preta). No contexto de aprendizado de máquina, caixa preta descreve modelos que não podem ser compreendidos apenas olhando para seus parâmetros. No entanto, quando se trata de tomar decisões que afetam a vida das pessoas, como na avaliação de desempenho de um pesquisador , é crucial entender por que estamos tomando aquela decisão. A importância de descrever as saídas da Inteligência Artificial (IA) torna-se evidente nesse ponto.

O problema surge quando muitos algoritmos de aprendizado de máquina aparentemente poderosos em termos de resultados e previsões sofrem de opacidade, tornando impossível obter insights sobre seus mecanismos internos. Isso é um problema crítico, pois confiar em decisões-chave de um sistema que não consegue se explicar traz riscos evidentes.

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, neste trabalho usamos os métodos explicativos SHAP e LIME nas predições do nosso modelo com o intuito de modelar a transparência, fornecendo explicações interpretáveis que pesquisadores e demais usuários podem entender e confiar.

2.3.1 Método SHAP

Esta é uma explicação de modelos de aprendizado de máquina com valores de Shapley, este algoritmo foi publicado pela primeira vez em 2017 por Lundberg e Lee (LUNDBERG; LEE, 2017). Os valores de Shapley são uma abordagem amplamente utilizada da teoria dos jogos cooperativos. A ideia é que você considere cada característica como um jogador e o conjunto de dados como uma equipe. Cada jogador dá a sua contribuição para o resultado da equipe. A soma dessas contribuições nos dá o valor da variável de destino dados alguns valores das característica. A atribuição de importância da fórmula do SHAP ajuda a entender como cada recurso contribui para a decisão final do modelo.

$$\phi_i = \frac{1}{N} \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} \left[f(S \cup \{i\}) - f(S) \right] \tag{2.2}$$

Resumidamente, a fórmula do SHAP calcula a diferença nas previsões do modelo ao adicionar o recurso i a diferentes subconjuntos S de outros recursos. Em seguida, essas diferenças são ponderadas pelo coeficiente binomial e somadas para obter o valor SHAP (ϕ_i) para o recurso i.

Pode-se citar como benefícios da abordagem do uso de valores SHAP:

1. Interpretabilidade global: os valores SHAP não apenas mostram a importância da característica, mas também mostram se a característica tem um impacto positivo ou negativo na média das previsões do modelo.

2. Os valores SHAP podem ser usados para explicar uma grande variedade de modelos, incluindo modelos lineares, modelos baseados em árvore e redes neurais , enquanto outras técnicas só podem ser usadas para explicar tipos limitados de modelos.

Em poucas palavras, os valores SHAP são usados sempre que você tem um modelo complexo (pode ser um aumento de gradiente, uma rede neural, ou qualquer coisa que

tome alguns recursos como entrada e produz algumas previsões como saída) e você deseja entender quais decisões o modelo está tomando.

Existem duas principais abordagens no framework para o calculo dos valores SHAP, KernelSHAP e PermutationSHAP, ambos visam alcançar os valores de Shapley, a diferença entre eles reside em seus métodos de computação subjacentes. O KernelSHAP usa uma média ponderada das previsões do modelo com base em uma função do kernel, enquanto o PermutationSHAP observa diretamente o impacto das permutações de recursos nas previsões do modelo. O KernelSHAP tende a ser mais preciso, mas computacionalmente caro, enquanto o PermutationSHAP é computacionalmente eficiente, mas pode fornecer aproximações um pouco menos precisas, especialmente para modelos complexos e não lineares. Aqui por enquanto abordaremos o PermutationSHAP pela complexidade do modelo não ser tão extensa ainda e pela limitação dos recursos computacionais disponíveis.

2.3.2 Método LIME - Local Interpretable Model-Agnostic Explanations

Dentro do Aprendizado de Máquina Explicável, encontra-se o arcabouço LIME (RIBEIRO; SINGH; GUESTRIN, 2016), um método que fornece as explicações para as decisões do modelo de AM humanamente entendíveis.

LIME é um algoritmo que pode explicar as previsões de qualquer classificador ou regressor de forma fiel, por aproximação localmente com um modelo interpretável. O LIME aborda a explicação de um modelo de aprendizado de máquina ao redor de uma instância específica de entrada x. A ideia central é criar uma explicação local para a previsão do modelo $f(x)$ criando um modelo interpretável local $g(z)$, onde z é um vetor de características que aproxima x na vizinhança local. O modelo $g(z)$ é construído usando pesos otimizados para fornecer uma explicação que seja interpretável pelo ser humano. As etapas do fornecimento das explicações pelo framework ocorre da seguinte forma:

Amostragem dos dados próximos: Amostra-se N instâncias z próximas à instância de interesse x a partir de uma distribuição específica. Geralmente, usa-se o método de amostragem por distância, como amostragem com pesos ou amostragem por kernel.

Cálculo das predições de f para z: Realiza-se a predição do modelo f para todas as instâncias z geradas na etapa anterior.

Ponderação dos dados de treinamento: Calcula-se os pesos para as amostras z com base em sua proximidade com a instância x usando uma função de distância, como a distância euclidiana ou a distância de similaridade.

- Treinamento do modelo interpretável: Utiliza-se os dados ponderados gerados para treinar um modelo interpretável, geralmente um modelo linear, como regressão

logística ou regressão linear.

- Obtenção das explicações: As características importantes são extraídas do modelo interpretável como as explicações locais, que fornecem informações sobre como as características influenciam a previsão do modelo f para a instância x.

3 Arcabouço de explicação do índice-h futuro

Figura 1 mostra as duas etapas principais deste trabalho ao todo: (1) A construção do modelo do índice-h futuro e (2) A obtenção das explicativas para as predições do modelo. A seguir é dado mais detalhes de cada uma.

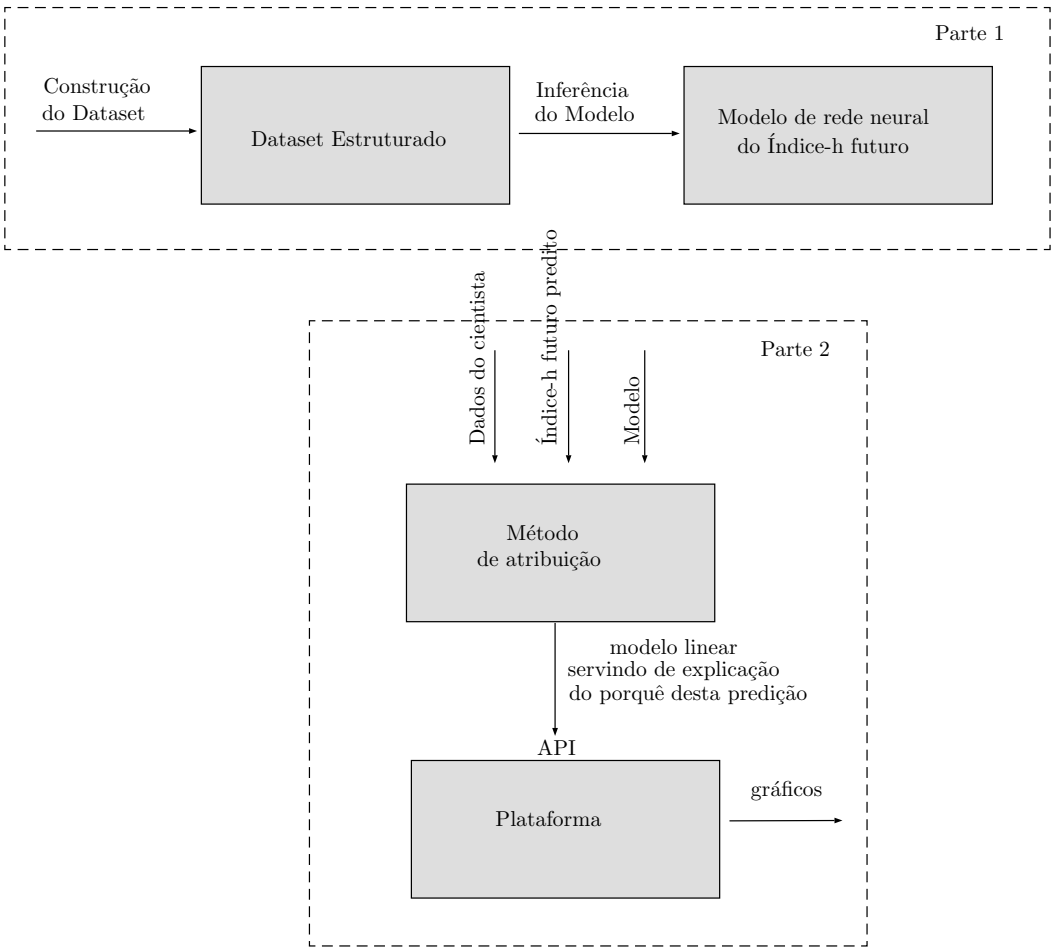

Figura 1 – Operação da Plataforma integrada com os métodos de entendimento (atribuição).

O projeto como um todo envolve as diversas etapas descritas na estrutura acima, desde a criação do modelo de previsão do índice-h até sua implantação na aplicação web. No entanto, o escopo desta parte específica do trabalho se concentra exclusivamente na criação do modelo de previsão do índice-h de pesquisadores. Contudo, é importante destacar que o trabalho como um todo abrange também a implementação da aplicação web, onde o modelo desenvolvido será integrado para fornecer previsões e explicações aos usuários finais. Essa etapa será realizada posteriormente, visando disponibilizar uma ferramenta acessível, para os cientistas que desejam avaliar e compreender o impacto acadêmico de suas pesquisas.

4 Conjunto de Dados

Neste trabalho, o modelo a ser avaliado foi treinado apartir de um conjunto de dados subjacente ao serviço de pesquisa Semantic Scholar criado pelo Allen Institute for A.I.

Tabela 1 – Descrição das Características dos Cientistas usadas no conjunto de dados.

Nome da Característica	**Descrição**
author_hindex	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

A base contém 836024 valores com 44 caracteristicas para um determinado autor, sendo o indice-h de cada autor a caracteristica alvo utilizada para construírmos um modelo

de regressão para prever o Indice-H do autor com base nas outras 43 caracteristicas e recursos da base.

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

4.1 Análise dos dados

Um levantamento inicial, utilizando a base de dados já pré-processada, em cima das estatísticas nos mostrou um indicativo de desbalanceamento dos valores do conjunto, a Média (AVG) e Desvio padrão (STD) dos valores do indice-h do conjunto de dados estão muito mais próximos do valor mínimo da medida do que do valor máximo, como mostra o gráfico abaixo:

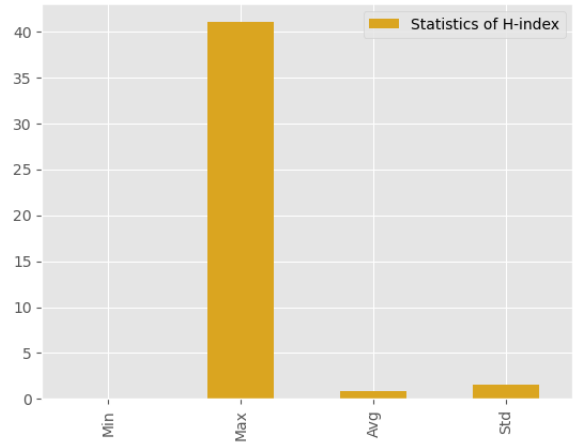

Figura 2 – Estatísticas do valor alvo (índice-h)

Quando a maioria dos valores do alvo está concentrada em uma faixa estreita, e há poucos valores altos, isso pode afetar significativamente o desempenho do modelo de aprendizado de máquina. O desbalanceamento pode causar problemas na capacidade do modelo de aprender padrões para os valores altos, uma vez que a proporção de exemplos com esses valores é limitada. Isso pode levar a um viés do modelo em direção a predições mais baixas, uma vez que a média da base de dados está próxima de zero. Para confirmar tais evidências usaremos, mais adiante, os métodos de explicação das bibliotecas SHAP e LIME em amostras de preditas pelo modelo treinado.

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

5 Configuração Experimental

5.1 Configuração

5.1.1 Normalização dos dados

A normalização de dados foi uma etapa crucial na preparação de dados para treinar o modelo de aprendizado de máquina. A normalização é particularmente importante ao usar algoritmos sensíveis à escala dos dados de entrada, como algoritmos de otimização baseados em gradiente usados em redes neurais.

Um método popular de normalização é a normalização de pontuação Z (também conhecida como padronização). A normalização do Z-score dimensiona os dados para ter uma média de 0 e um desvio padrão de 1. A fórmula em Python para calcular o Z-score para os ponto de dados foi dada por:

```
data_normalized = ((dataset-dataset.mean())/dataset.std())
```

Importância da normalização:

• Velocidade de convergência: Ajudou os algoritmos de otimização a convergirem mais rapidamente padronizando a escala dos recursos, tornando-os comparáveis e tentando evitar que os recursos de maiores magnitudes dominem o processo de aprendizagem.

• Desempenho aprimorado: a normalização aprimorou o desempenho do modelo, evitando gradientes de desaparecimento ou explosão que podem ocorrer ao trabalhar com dados não normalizados em modelos de aprendizado profundo.

• Interpretabilidade aprimorada: Quando os recursos estão na mesma escala, fica mais fácil interpretar os coeficientes do modelo ou as importâncias dos recursos, pois são diretamente comparáveis.

Divisão dos dados de treinamento e teste:

Em seguida realizamos o processo de divisão, dimensionamento e transformação do conjunto de dados para treinamento e teste usando o método train_test_split da biblioteca scikit-learn:

```
X_train, X_test, y_train, y_test = train_test_split(features, target,
                                    test_size=0.20, random_state=42)])
```

Neste modelo, o test_size parâmetro é definido como 0,2, o que significa que 20%

dos dados serão usados para teste e o restante (80%) será usado para treinamento. O random_state parâmetro garante a reprodutibilidade, pois configurá-lo para o valor 42 garante que essa mesma divisão seja gerada toda vez que você executar o código.

Assim, O número de registros no conjunto de treinamento é de 668.819 registros e o conjunto de teste é de 167.205 registros de diferentes, Separados em variáveis de feição (features) e variável de destino (target). As 43 características são as variáveis independentes que foram usadas para fazer previsões e o indice-h é a variável dependente prevista.

5.1.2 Configuração do Modelo de RNA

O modelo de rede neural Keras fornecido é uma rede neural feedforward projetada usando a API sequencial em Keras. Consiste em múltiplas camadas densas (totalmente conectadas) com a função de ativação da unidade linear retificada (ReLU) para camadas intermediárias e uma função de ativação linear para a camada de saída. O modelo é compilado com a função de perda Mean Squared Error (MSE), o otimizador Adam e métricas adicionais para monitorar durante o treinamento.

O modelo de rede neural é definido como um Sequential model em Keras. Um Sequential model é apropriado para redes neurais feedforward, onde cada camada segue a anterior sequencialmente.

Detalhes da configuração e Arquitetura do modelo:

```
def build_model():
  model = keras.Sequential([
    layers.Dense(100, activation='relu', input_shape=(43)),
    layers.Dense(100, activation='relu'),
    layers.Dense(100, activation='relu'),
    layers.Dense(50, activation='relu'),
    layers.Dense(1)
  ])
```

A Arquitetura de camadas Do modelo consiste em várias camadas densas (totalmente conectadas). Cada camada densa possui um número especificado de neurônios (unidades) e uma função de ativação.

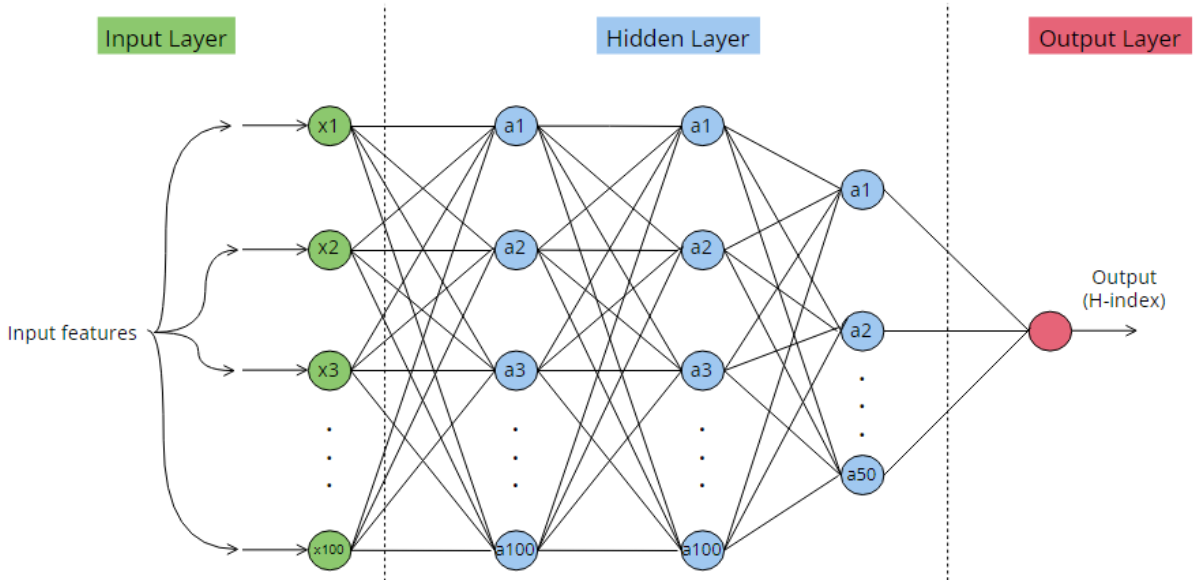

Figura 3 – Diagrama do modelo de Rede Neural

1. Camada de entrada: A primeira camada é a camada de entrada com 100 neurônios (unidades) e usa a função de ativação ReLU. O input_shape=(43)parâmetro define a forma dos dados de entrada, indicando que cada amostra de entrada possui 43 características.

2. Camadas Intermediárias: O modelo contém três camadas densas intermediárias, duas com 100 neurônios outra com 50 e a função de ativação ReLU. Essas camadas destinam-se a capturar padrões e relacionamentos complexos nos dados.

3. Camada de Saída: A última camada é a camada de saída com 1 neurônio. Como não há função de ativação especificada para esta camada, ela é considerada uma ativação linear por padrão. Isso é adequado para tarefas de regressão em que queremos que o modelo gere diretamente um valor contínuo.

Depois de definir a arquitetura, o modelo é compilado com a função de perda especificada, otimizador e métricas adicionais para monitorar durante o treinamento.

```
model.compile(loss='mse',
              optimizer= 'adam',
              metrics=['mae', 'mse', 'accuracy'])
```

1. loss = 'mse': a perda do erro quadrático médio (MSE) é usada para tarefas de regressão, em que o objetivo é minimizar a diferença quadrada entre os valores previstos e os valores de destino reais.

2. optimizer = 'adam': O otimizador Adam é um algoritmo de otimização de taxa de aprendizado adaptativo comumente usado no treinamento de redes neurais. Ele adapta a taxa de aprendizado durante o treinamento para melhorar a velocidade de convergência e a estabilidade.

3. metrics = ['mae', 'mse', 'accuracy']: durante o treinamento, o modelo monitorará três métricas: erro absoluto médio (MAE), erro quadrático médio (MSE) e precisão.

5.1.3 Avaliação do modelo

O processo de treinamento do modelo envolve o uso dos dados de treinamento fornecidos (X_train e y_train) para otimizar os parâmetros do modelo (pesos e vieses) com base na função de perda especificada (MSE) e otimizador (Adam). O processo de treinamento atualiza iterativamente os parâmetros do modelo por 100 épocas e atualizações em 64 lotes, com o objetivo de minimizar a perda nos dados de treinamento e melhorar a capacidade do modelo de generalizar para dados não vistos. A validação durante o treinamento permite monitorar o desempenho do modelo em um conjunto de dados separado (X_test, y_test) ao final de cada época. Isso ajuda a avaliar a capacidade de generalização do modelo.

```
history = model.fit(
  X_train, y_train, epochs=100, batch_size=64,
  validation_data=(X_test, y_test), verbose=0)
```

Após a execução do método model.fit(), o history objeto conterá o histórico de treinamento, incluindo perdas e valores de métricas para cada época.

Assim, usamos essas informações para analisar o desempenho do modelo e visualizar o resultado do treinamento. As implicações trazem um bom indicador de que o modelo está funcionando razoavelmente bem com um baixo valor da métrica de perda (MSE) que representa a média quadrada entre os valores previstos do modelo e os valores de destino reais. No entanto, o modelo atinge uma perda baixa de 0,021871 no seu melhor caso e tem um desempenho com uma perda de 0,228763 no seu pior caso. Tal diferença entre o melhor e o pior desempenho pode ser uma indicação de sobreajuste do modelo.

5.2 Configuração dos Explicadores

As previsões e explicações feitas por um modelo bem-sucedido é influenciada por diferentes níveis de importância de vários recursos comportamentais de entrada. No entanto, devido à natureza complexa das redes neurais profundas, não podemos avaliar

diretamente e de forma aprofundada a importância das características e o desempenho do modelo no geral ou para casos específicos, pois o modelo se comporta como uma "caixa preta". Para superar essa limitação, podemos usar métodos de explicabilidade que fornecem aproximações de como cada recurso contribui para a previsão de um determinado pesquisador e como eles se aproximam dos valores reais.

Para investigar esse aspecto, nos concentramos em dois métodos populares de explicabilidade baseados em instâncias encontrados na literatura. Esses métodos vêm de diferentes famílias e oferecem diferentes abordagens para interpretar o comportamento do modelo. Os dois métodos de explicabilidade considerados são descritos a seguir.

5.2.1 SHAP

Em um primeiro experimento, configuramos o explicador SHAP usando o modelo e os dados de treinamento e, em seguida, calculamos os valores SHAP para os dados de teste.

```
explainer = shap.Explainer(model, X_train)

# Calcula os valores SHAP para as previsoes do conjunto de teste
shap_values = explainer(X_test)
```

Primeiro de tudo, criamos um objeto chamado explainer. É o objeto que recebe, como entrada, o método de previsão do nosso modelo e o conjunto de dados de treinamento. Para tornar o modelo SHAP agnóstico, ele realiza uma perturbação em torno dos pontos do conjunto de dados de treinamento e calcula o impacto dessa perturbação no modelo. É um tipo de técnica de reamostragem, cujo número de amostras é definido posteriormente. O resultado é uma estimativa estatística dos valores SHAP.

- Shap_values é uma matriz 2D em que Cada linha pertence a uma única previsão feita pelo modelo. Cada coluna representa uma característica usado no modelo. Cada valor SHAP representa o quanto essa característica contribui para a saída da previsão dessa linha.

Para gerar a explicação global com o SHAP, usamos o método summary_plot():

```
shap.summary_plot(shap_values, X_test, feature_names=feature_names)
```

Esse método cria um gráfico de resumo que exibe a importância do recurso para todos os pontos de dados no conjunto de dados de teste. Os resultados serão discutidos no capítulo seguinte.

5.2.2 LIME

Na segunda abordagem explicativa, a técnica LIME é usada para fornecer explicações locais para as previsões do modelos, ele cria um modelo mais simples e interpretável que atue como uma aproximação local do nosso modelo de caixa preta. Em vez de tentar entender todo o modelo complexo, ele aproxima as previsões do nosso modelo nas proximidades de um ponto de dado específico usando o modelo mais simples, facilitando a compreensão do comportamento do modelo real para essa instância específica.

Primeiramente inicializamos um objeto de explicação LIME para dados tabulares com modo de regressão:

```
explainer_lime = lime_tabular.LimeTabularExplainer(X_train, mode="
                                    regression", feature_names=
                                    feature_names)
```

Depois de criar o objeto explainer_lime, podemos usá-lo para gerar as explicações locais. Por ele o LIME gerará o modelo local interpretável em torno da instância de interesse e o usará para aproximar o comportamento do nosso modelo. Aqui executamos a explicação local para um ponto de dado de interesse (X_test[instance_interest]) do conjunto de dados de teste:

```
explanation = explainer_lime.explain_instance(X_test[instance_interest],
                                     model.predict)
```

O objeto de explicação resultante (explanation) fornece informações sobre as contribuições de recursos individuais para a previsão do modelo para aquela instância, ajudando a interpretar o processo de tomada de decisão do modelo nas proximidades da instância de interesse. Em seguida, ele perturba ou altera ligeiramente esse ponto de dado para criar um novo conjunto de dados sintético. O conjunto de dados sintético inclui o ponto de dados original e alguns outros pontos de dados semelhantes a ele. Esses pontos formam uma "vizinhança local"em torno do ponto de dados específico.

5.3 Resultados preliminares e Discussão

A seguir discutiremos os resultados das explicações geradas pelos métodos aplicados no capítulo anterior.

5.3.1 Desempenho do Modelo

A tabela 2 mostra o desempenho do modelo proposto. O modelo apresenta uma acurácia alta. Isto está em concordância com outros achados.

Tabela 2 – Histórico das medidas de avaliação do modelo.

	loss	mse	accuracy	val_loss	val_mse	val_accuracy
mean	0.038909	0.038909	0.860545	0.043400	0.043400	0.860096
std	0.025901	0.025901	0.003963	0.015864	0.015864	0.001378
min	0.021871	0.021871	0.822500	0.032093	0.032093	0.849653
max	0.228763	0.228763	0.861634	0.119561	0.119561	0.860792

5.3.2 Discussão das explicações globais obtidas com SHAP

Para as explicações globais, o SHAP forneceu um resumo da importância dos recursos em todo o conjunto de dados, medimos a contribuição das características de entrada para previsões globais (média de previsões de todas as instâncias do modelo), mostrando a importância das características e se as mesmas potencialmente demonstram existir vieses no conjunto de dados.

A Observação principal é que o impacto de uma característica não depende apenas de uma única característica, mas de todo o conjunto de características no conjunto de dados. Assim, calculamos o impacto de cada recurso na variável de destino (chamada valor de SHAP) usando cálculo combinatório e treinando novamente o modelo sobre toda a combinação de características. Assim, o valor absoluto médio do impacto de uma característica contra a variável de destino é usado como uma medida de sua importância. Em seguida, cada recurso, com seus valores de SHAP, contribui para empurrar a saída do modelo, desse valor base para a esquerda e para a direita. Com o método summary_plot , podemos plotar gráficos de pontos para visualizar o impacto da direcionalidade dos recursos.

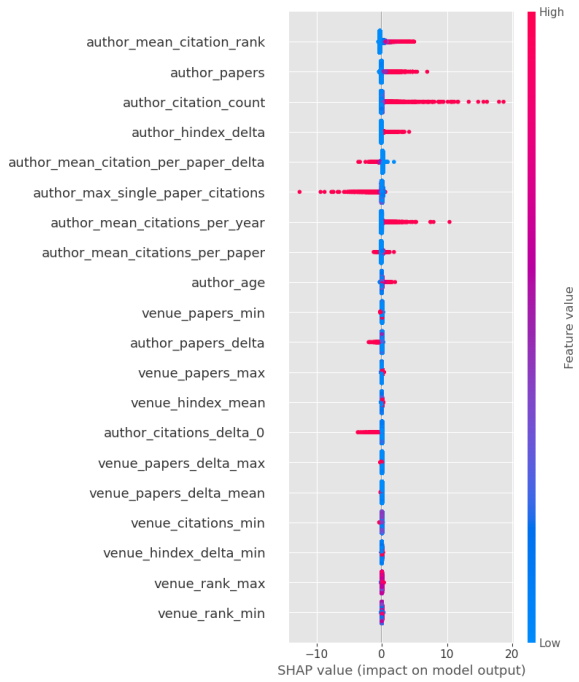

Figura 4 – Pontos de dados

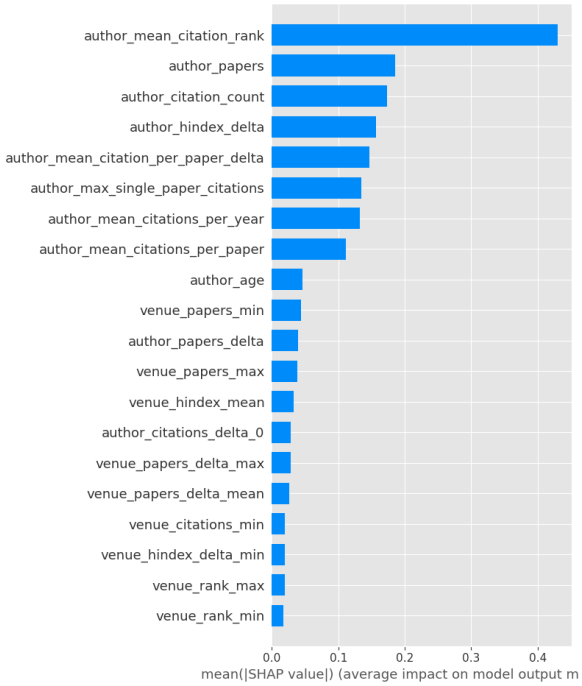

Figura 5 – Impacto dos recursos

O valor SHAP positivo significa uma característica com impacto positivo na previsão e SHAP negativo uma característica que impactou negativamente na previsão, ambas tem a mesma escala de impacto, não significando que o SHAP negativo impactou menos, mas sim que pode ter impactado muito ou pouco de forma negativa ao tentar se aproximar da real previsão. A seguir, alguns insights que podemos concluir no gráfico acima:

1. A orientação vertical mostra qual característica do pesquisador está representando.
2. As cores mostram quanto os valores das característica foram altos ou baixos para a linha do conjunto de dados.
3. A orientação horizontal mostra se o efeito desse valor causou uma previsão maior ou menor.
4. Valores mais altos para as características author_mean_citation_rank, author_papers e author_citation_count, author_hindex_delta e author_mean_citations_per_year tem maior impacto positivo na predição do índice-h.
5. Valores mais altos para as características author_mean_citation_per_paper_delta, author_max_single_paper_citations, author_papers_delta, author_citations_delta_0 tem maior impacto negativo na predição índice-h.

Este gráfico de resumo oferece uma visão global do comportamento do modelo e a importância relativa das diferentes características dos autores ao fazer previsões em

todo o conjunto de dados. Ele ajuda a identificar quais recursos têm o impacto mais significativo nas previsões do modelo em média e como eles afetam a saída. No contexto de uso do nosso trabalho, ele também nos ajuda a identificar vieses no modelo, principalmente ao observar que as característica: author_citation_count, que a armazena o número de contagem cumulativa do autor e o número de publicações do autor (author_papers) que são as métricas de maior relevância no índice de Hirsch, não se caracterizam como os recursos de maior importância global no nosso modelo, no qual tem como principal recurso o rank do autor entre todos os outros autores em termos de citações médias por ano (author_mean_citation_rank). Esses padrões consistentes, indicam que o modelo pode acabar sendo sistematicamente tendencioso, produzindo estimativas a favor ou contra certos valores de recursos em suas previsões, que sistematicamente se desviam da verdadeira relação entre as que seriam as principais variáveis de interesse do índice-h.

Além disso, com relação ao uso do explicador SHAP através da aplicação, para os usuários finais do nosso modelo, estamos ponderando como tais métodos podem ser usados para uma explicação global dentro de um determinado grupo de cientistas, como no domínio de universidades e instituições de pesquisa onde o explicador traria uma visão global limitada aquele grupo e seus pesquisadores.

5.4 Discussão das explicações Locais com LIME

Com o LIME, selecionamos alguns pontos de dados individuais para os quais desejamos uma explicação, separamos duas instâncias de quatro grupos para previsões de índice-h de pesquisadores, os grupos remetem à amostras de pesquisadores com niveis de indice-h baixo(Figuras 6 e 7), médio((Figuras 8 e 9) e aproximadamente alto((Figuras 10, 11, 12 e 13) em relação ao conjunto de teste. Essas explicações locais são valiosas para entender o processo de tomada de decisão do modelo para um autor específico.

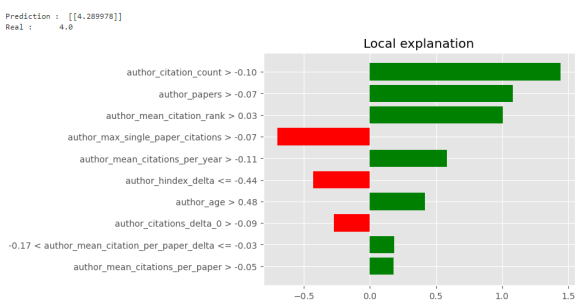

Figura 6 – 1ª Amostra H-index = 4

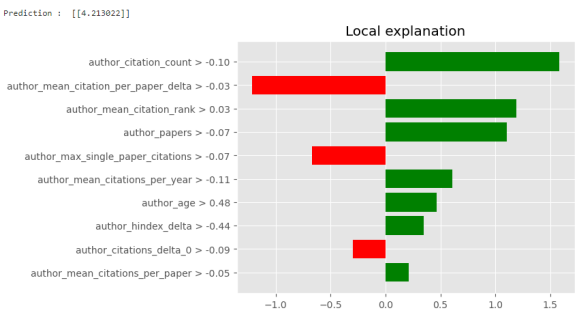

Figura 7 – 2ª Amostra H-index = 4

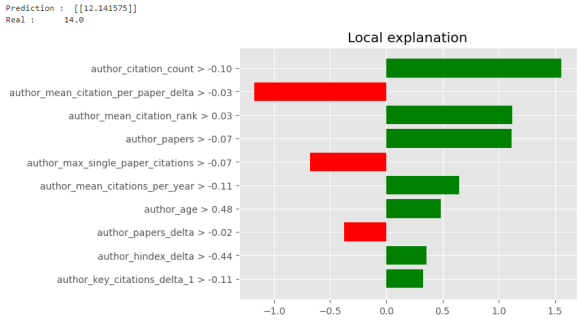

Figura 8 – 1ª Amostra H-index = 14

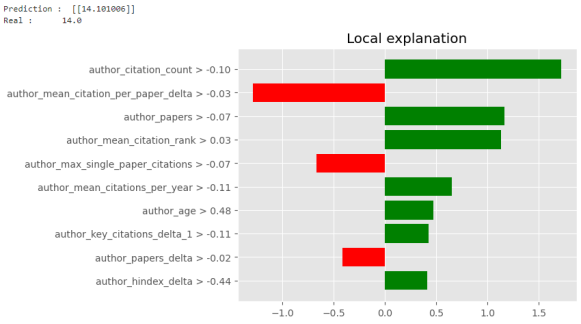

Figura 9 – 2ª Amostra H-index = 14

Nos dois primeiros grupos, a explicação do modelo LIME aproximado da nossa rede demonstra como o modelo prediz com ótima precisão para valores mais baixos como os de índice-h 4. Porém nas amostras de ìndice-h com valor real = 14, podemos ter casos de predições bem precisas mas de outras um pouco distante do valor real.

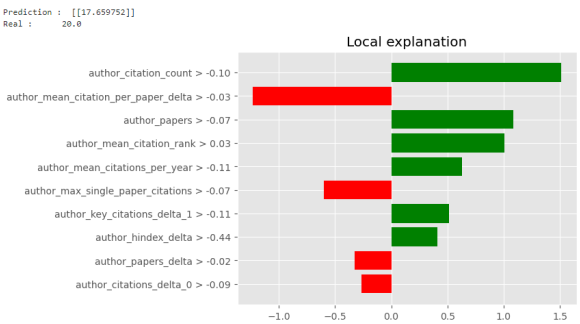

Figura 10 – 1ª Amostra H-index = 20

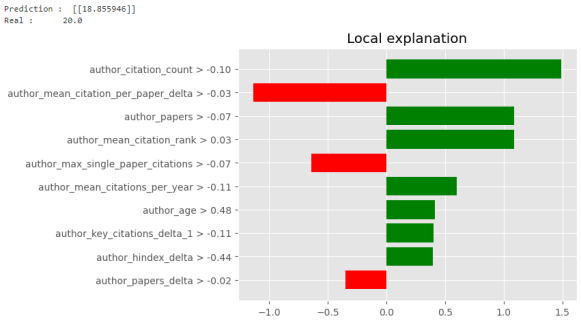

Figura 11 – 2ª Amostra H-index = 20

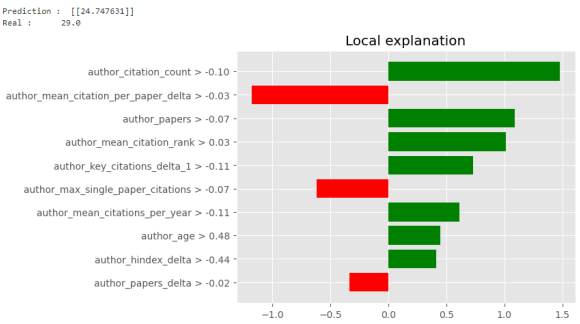

Figura 12 – 1ª Amostra H-index = 29

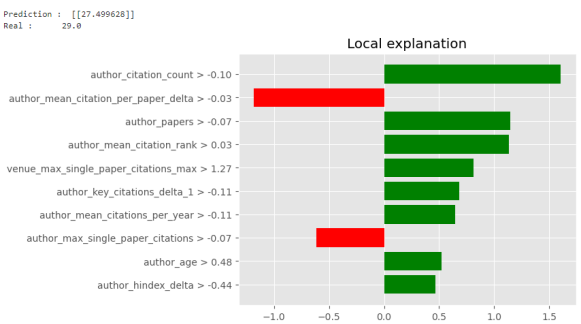

Figura 13 – 2ª Amostra H-index = 29

Posteriormente aumentando o valor de predição das amostras, notamos que o modelo começa a se distanciar no valor predito para o valor real a medida que testamos autores com indicadores mais elevados. Tal indício, evidencia o problema de desbalanceamento no conjunto de dados, que consequentemente dificulta que o modelo generalize para os dados probabilísticos minoritários durante o treinamento, pois há poucos valores altos. Assim, A qualidade de generalização do modelo é afetada pois a ótima precisão do modelo se concentra numa faixa estreita do valor alvo.

Outrossim, o explicador possibilitou enxergar e ter uma compreensão individual para cada cientista. Os cientistas podem ter trajetórias de carreira e focos de pesquisa únicos, ao usar os explicadores, podemos oferecer explicações personalizadas para cada cientista, mostrando a eles como suas características individuais e marcos de carreira contribuem para suas previsões de índice h. Essa abordagem personalizada pode levar a um envolvimento mais profundo com a aplicação a ser desenvolvida e aprimorando a experiência do usuário.

6 Conclusão

A interpretabilidade é essencial para garantir a qualidade e confiabilidade das previsões. Seus métodos de explicabilidade tornam-se importantes não apenas para especialistas em aprendizado de máquina, mas também para usuários em todas as áreas. Desse modo, com o intuito de tornar tais ferramentas acessíveis, neste trabalho buscamos propor um modelo de RNA interpretável para a predição do índice-h de pesquisadores, justificando a importância da aplicação de métodos, além dos convencionais, para compreender o comportamento do modelo e do problema ao qual foi aplicado em detalhes. Os resultados do treinamento do modelo foram promissores, alcançando um (MSE) de 0.0389 no conjunto de teste. No entanto, também podemos uma diferença significativa entre o melhor e o pior desempenho do modelo, sugerindo a possibilidade de sobreajuste e a necessidade de lidar com o desbalanceamento dos dados. Por meio das explicações locais fornecidas pelo método LIME, foi possível identificar padrões de predição mais precisos de forma individual para cada autor, como autores com valores mais baixos de índice-h, que tem suas previsões com maior exatidão, enquanto para autores com medidas do índices-h mais elevadas, o desempenho do modelo começava a se distanciar dos valores reais. Essa observação apontou para o impacto do desbalanceamento nos dados de treinamento, que afetou a capacidade de generalização do modelo, especialmente para os valores minoritários menos representativos. Essa compreensão é essencial para melhorar o desempenho do modelo em diferentes faixas de valores de índice-h e garantir que ele seja justo e confiável em todas as situações.

Ao analisar os resultados das explicações para a RNA, as métricas globais fornecidas pelo SHAP mostraram ser benéficas, pois forneceu medidas numéricas simples para avaliar a importância dos recursos do modelo. Isso permitiu comparações fáceis entre os recursos, tornando os resultados visualmente apresentáveis (4, 5) até mesmo para um público não técnico. Identificamos que características como o rank do autor em termos de citações médias por ano mostraram-se mais influentes nas previsões do modelo em comparação com métricas tradicionais, como o número total de citações do autor e o número de publicações. Essa observação sugeriu a possibilidade de vieses no modelo, indicando que as métricas de maior relevância no índice-h não se destacaram como o recurso de maior importância global no nosso modelo interpretável. Esse achado é crucial para promover a transparência e equidade nas avaliações acadêmicas, permitindo aos comitês de seleção e pesquisadores compreender como as características específicas são valorizadas por outras perspectivas.

Em suma, o trabalho apresentou resultados promissores do modelo junto a técnicas de explicabilidade. A combinação das explicações foram fundamentais para compreender o comportamento do modelo em diferentes contextos e contribuir para a melhoria do

processo de tomada de decisão, também proporcionou, do ponto de vista científico, uma visão mais abrangente e detalhada de previsão do índice-h. Com essa interpretabilidade promovemos uma experiência mais envolvente para os usuários finais, permitindo que eles valorizem a aplicação como uma ferramenta valiosa em suas carreiras acadêmicas. Em vista disso, o trablho abre novas perspectivas para a aplicação de IA Explicável no meio acadêmico, com ênfase na ética, equidade e responsabilidade, para promover o avanço científico e o desenvolvimento da comunidade como um todo.

6.1 Trabalhos futuros

Há uma série de caminhos para trabalhos futuros que podem ser mais bem explorados. Neste contexto, os trabalhos futuros envolvem a criação de uma plataforma interativa que permita aos usuários acessar e interagir com o modelo, para ampliar o alcance e utilidade do trabalho desenvolvido. No geral, obtemos um modelo decente, que será melhorado a longo prazo. Ao implantá-lo na aplicação Web testaremos e aplicaremos melhoras no seu aprendizado, balanceando o modelo com novos valores, reajustando a base de dados e seus valores principalmente focando em técnicas, para uma melhorar a generalização do modelo.

Algumas etapas do trabalho futuro a serem analisadas e desenvolvidas:

1. Interface Amigável e Explicativa: O primeiro passo na implementação seria projetar uma interface amigável e explicativa para a aplicação web. Essa interface deve permitir que os usuários entendam facilmente o propósito do modelo, suas funcionalidades e como interpretar as previsões fornecidas.

2. Incorporação dos Explicadores: A aplicação web deve incorporar os explicadores SHAP e LIME, permitindo aos usuários visualizar e compreender as explicações para suas previsões individuais. Isso garantirá que os cientistas em início de carreira ou usuários não técnicos possam compreender como suas características individuais influenciam suas previsões de índice-h.

3. Visualização Gráfica dos Resultados: A apresentação visual dos resultados é essencial para facilitar a interpretação dos usuários. Gráficos, tabelas e visualizações interativas podem ser utilizados para mostrar as contribuições de cada recurso e como eles afetam a previsão do índice-h.

4. Melhoria da Usabilidade: A aplicação web deve ser projetada para ser intuitiva e fácil de usar. Recursos como filtros, opções de seleção e ferramentas de busca podem ser adicionados para melhorar a experiência do usuário e facilitar a exploração dos dados.

5. Análise de Erro em Tempo Real: A aplicação pode ser aprimorada com a capacidade de analisar erros em tempo real, destacando casos em que o modelo pode fazer previsões imprecisas ou inconsistentes. Isso permitirá que os usuários compreendam melhor as limitações do modelo e ofereçam insights valiosos para refinamentos futuros.

6. Incentivo à Participação e Feedback dos Usuários: A implementação da aplicação web pode incluir mecanismos para incentivar a participação dos usuários e coletar feedback valioso. Pesquisadores podem fornecer comentários sobre as previsões recebidas e a experiência geral, o que permitirá aprimorar continuamente o modelo e a aplicação.

7. Escalabilidade e Segurança: À medida que a aplicação web ganhar usuários, a escalabilidade e segurança se tornaram questões essenciais. Portanto, abordagens de otimização da aplicação para lidar com as solicitações de usuários e garantir a segurança dos dados pessoais dos pesquisadores devem ser garantidas.

Assim realizando uma série de estudos e planejamentos, para garantir que a implantação futura do modelo em uma aplicação web seja bem-sucedida, promoveremos a acessibilidade das explicações e insights para pesquisadores, universidades e agências de pesquisa, tornando-se uma valiosa ferramenta para aprimorar os processos de avaliação acadêmica e a alocação de financiamento de pesquisa de forma eficiente, segura e que atenda às necessidades dos usuários. .

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 14 e 18.

CLAUSET, A.; LARREMORE, D. B.; SINATRA, R. Data-driven predictions in the science of science. *Science*, v. 355, n. 6324, p. 477–480, 2017. Citado na página 14.

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 20.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. Citado na página 17.

GUJARATI, D. N. *Basic Econometrics*. [S.l.]: McGraw-Hill Education, 2000. Citado na página 18.

HAYKIN, S. S. *Redes neurais: princípios e prática*. 2ª ed.. ed. [S.l.]: Bookman, 2001. Citado na página 16.

HAYKIN, S. S. *Neural networks and learning machines*. Third. Upper Saddle River, NJ: Pearson Education, 2009. Citado na página 17.

HIRSCH, J. E. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 102, n. 46, p. 16569–16572, 2005. Citado na página 14.

HOU, J.; PAN, H.; GUO, T.; LEE, I.; KONG, X.; XIA, F. Prediction methods and applications in the science of science: A survey. *Computer Science Review*, v. 34, p. 100197, 2019. Citado na página 14.

JR., A. C. B.; HO, Y.-C. Applied optimal control: Optimization, estimation, and control. *Blaisdell Publishing Company*, 1969. Citado na página 18.

LINDAHL, J. *In search of future excellence : bibliometric indicators, gender differences, and predicting research performance in the early career*. 60 p. Tese (Doutorado) — Umeå University, Department of Sociology, 2020. Citado na página 14.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, v. 30, 2017. Citado na página 21.

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 20.

NIELSEN, M. A. *Neural Networks and Deep Learning: A Textbook*. [S.l.]: Deterministic Books, 2015. Citado na página 18.

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 20.

PENNER, O.; PAN, R. K.; PETERSEN, A. M.; KASKI, K.; FORTUNATO, S. On the predictability of future impact in science. *Scientific Reports*, v. 3, n. 3052, 2013. Citado na página 14.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1135–1144. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939778>. Citado na página 22.

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 19.

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 20.

SELBST, A. D.; BAROCAS, S. The intuitive appeal of explainable machines. *Fordham L. Rev.*, HeinOnline, v. 87, p. 1085, 2018. Citado na página 20.

SINATRA, R.; WANG, D.; DEVILLE, P.; SONG, C.; BARABÁSI, A.-L. Quantifying the evolution of individual scientific impact. *Science*, v. 354, n. 6312, 2016. Citado na página 14.

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 26.