UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Previsão do índice-h de pesquisadores brasileiros de cursos de pós-graduação em computação

Valéria de Carvalho Santos

Monografia - MBA em Inteligência Artificial e Big Data.

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Valéria de Carvalho Santos

Previsão do índice-h de pesquisadores brasileiros de cursos de pós-graduação em computação

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Concentration area: Artificial Intelligence and Big Data

Supervisor: Prof[a]. Dra. Tatiane Nogueira Rios

Versão revisada

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

d278p

de Carvalho Santos, Valéria
Previsão do índice-h de pesquisadores brasileiros de cursos de pós-graduação em computação / Valéria de Carvalho Santos; orientador Tatiane Nogueira Rios. -- São Carlos, 2024.
40 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Predição. 2. Regressão Linear. 3. Pesquisadores. 4. Ciência da Ciência. 5. Pós-Graduação. I. Nogueira Rios, Tatiane, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Valéria de Carvalho Santos

Área de concentração: Inteligência Artificial e Big Data

Prof. Dra. Tatiane Nogueira Rios

São Carlos

2024

Dedico este trabalho aos meus filhos Valentim e Ítalo
que transformaram minha visão de mundo
e me impulsionam a fazer o melhor que posso a cada dia.

AGRADECIMENTOS

Agradeço pela oportunidade de aprofundar meus conhecimentos em Inteligência Artificial através deste curso. Em especial, agradeço à professora Solange Rezende que a cada encontro desperta em mim o desejo de buscar meu crescimento pessoal e profissional.

Agradeço ao meu esposo Jadson pelo companheirismo e compreensão e aos meus filhos Valentim e Ítalo que são minha alegria e força de viver.

Agradeço aos amigos do CSILab-UFOP pela colaboração na condução deste trabalho e por me apoiarem a seguir na pesquisa.

Agradeço à minha orientadora Tatiane Rios pela orientação e parceira na condução deste trabalho. É uma grande honra ser orientada por uma pessoa que admiro tanto.

Por fim, agradeço a todos os professores, tutores, técnicos e colaboradores deste MBA pelo excelente trabalho.

RESUMO

Santos, V C . 2024. 40p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Um problema no meio acadêmico é como mensurar o desempenho de um pesquisador. Uma medida bem aceita internacionalmente pelas principais universidades é o índice-h. A proposta desse índice é quantificar a produtividade e o impacto de cientistas baseando-se nos seus artigos mais citados. O objetivo deste trabalho é prever o índice-h de pesquisadores brasileiros credenciados em programas de pós-graduação da computação daqui a alguns anos. Para isso, foi utilizado o modelo de regressão linear *ElasticNet*. Os resultados são comparados com um modelo de previsão de índice-h proposta anteriormente.

Palavras-chave: Predição. Regressão Linear. Pesquisadores. Ciência da Ciência. Pós-Graduação.

ABSTRACT

Santos, V C . 2024. 40p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

One challenge in the academic field is how to measure a researcher's performance. An internationally well-accepted measure by leading universities is the h-index. The purpose of this index is to quantify the productivity and impact of scientists based on their most cited papers. The goal of this work is to predict the h-index of Brazilian researchers affiliated with computer science graduate programs in the coming years. For this, the ElasticNet linear regression model was used. The results are compared with a previously proposed h-index prediction model.

Keywords: Prediction. Linear Regression. Researchers. Science of Science. Graduate Programs.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A disponibilidade de diferentes conjuntos de dados como OpenAlex (PRIEM; PIWOWAR; ORR, 2022), Web of Science e Scopus impulsionou a literatura na evolução da ciência, com investigações nas relações entre autores, artigos publicados, lugares, instituições e agências financiadoras de pesquisa. Alguns autores usam a expressão "Science of Science" (SciSci) para nomear as pesquisas desse assunto (PRICE, 1963; BOURDIEU, 2004; FORTUNATO *et al.*, 2018). Fortunato *et al.* (2018) acredita que entender os mecanismos básicos de fazer ciência deve direcionar a comunidade científica para otimizar o sucesso de cientistas e ciência como um todo.

Os recentes avanços em SciSci abrangem métodos de investigação sobre mobilidade de cientistas entre instituições de perspectivas diferentes, variando de vetor de *embeddings* a redes de ciência (FITZGERALD *et al.*, 2023; WAPMAN *et al.*, 2022; MURRAY *et al.*, 2023; SUGIMOTO *et al.*, 2017).

Outro tópico importante é a predição de produtividade (ACUNA; ALLESINA; KORDING, 2012; DONG; JOHNSON; CHAWLA, 2016; SINATRA *et al.*, 2016), cuja ideia é desvendar os fatores de sucesso de autores e construir modelos de predição. Acuna, Allesina and Kording (2012) propuseram um modelo a partir da regressão linear com regularização *ElasticNet* para predizer o índice-h futuro de autores em um, cinco e dez anos, baseado no número de artigos publicados, idade acadêmica, número de artigos em revistas de alto impacto e outros fatores que influenciam o sucesso científico, no contexto de Neurociência.

Entre as bases de dados bem conhecidas, OpenAlex tem ganhado popularidade por ser uma API livre e de fácil uso. Há milhões de entidades como autores, lugares, instituições, tópicos e conceitos ligados por bilhões de conexões.

No Brasil, programas de pós-graduação (PPGs) recebem notas a cada quatro anos pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) (CAPES, 2024), considerando a produção de seus professores, número de defesas de mestrado e doutorado e outros critérios. As notas variam de 1 a 7, sendo que 1 e 2 não permitem que o programa opere; 3 permite que funcione para cursos de mestrado; 4 a 7 pode ter doutorado, sendo 6 e 7 programas considerados de excelência.

Neste trabalho, uma base de dados de professores de programas de pós-graduação brasileiros em Ciência da Computação foi construída, relacionando dados da Capes e OpenAlex, que permite consultas sobre os registros de publicações desses professores usando a API OpenAlex. A partir desses dados, investiga-se o modelo de regressão linear *ElasticNet* para predizer o índice-h desses professores em 1, 2, 3 e 4 anos. Vale ressaltar

que o modelo *ElasticNet* apresenta propriedades que permitem mitigar o problema de sobreajuste devido à sua regularização, além de incorporar a capacidade de realizar a seleção de atributos relevantes para o modelo, aprimorando a interpretabilidade e a eficácia preditiva (JAMES *et al.*, 2014).

1.1 Hipotesis e Objetivos

A hipótese deste trabalho é que, baseado no modelo proposto por Acuna, Allesina and Kording (2012), predizer o índice-h pode ser efetivamente aplicado ao contexto de pesquisadores de programas brasileiros de pós-graduação em Ciência da Computação.

Assim, o objetivo desta pesquisa é investigar o modelo de regressão linear *ElasticNet* para previsão de índice-h de pesquisadores de programas brasileiros de pós-graduação em Ciência da Computação.

1.2 Organização do texto

Este documento está organizado como se segue. No Capítulo 2 são apresentados os métodos estudados. Os trabalhos relacionados ao tema de pesquisa investigado é apresentado no Capítulo 3. A proposta da pesquisa é apresentada no Capítulo 4. No Capítulo 5 são apresentados os experimentos e resultados. Por fim, o Capítulo 6 apresenta as conclusões.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os métodos de regressão linear e o modelo de predição de Acuna, Allesina and Kording (2012).

2.1 Regressão Linear

Os conceitos apresentados nesta seção foram baseados no livro de (ZHANG *et al.*, 2023).

Problemas de regressão aparecem quando se deseja predizer um valor numérico, por exemplo, predizer preços, tempo de permanência em hospital, entre outros. Suponha que deseja-se estimar os preços de casas baseado na área e na idade da casa. Para desenvolver um modelo para predizer casas, é preciso ter em mãos os dados, incluindo preço de venda, área e idade de cada casa.

De acordo com Zhang *et al.* (2023), regressão linear é a ferramenta mais simples e mais popular para tratar problemas de regressão. Regressão linear surge de simples suposições. Primeiro, assume-se que o relacionamento entre as características x e o alvo y é aproximadamente linear, $E[Y|X = x]$ pode ser expresso como uma soma ponderada das características x. Essa configuração permite que o valor alvo deve ainda desviar do seu valor esperado por conta do ruído de observação. Depois, pode-se supor que qualquer ruído é bem comportado, seguindo uma distribuição Gaussiana. Tipicamente, usa-se n para denotar o número de exemplos no conjunto de dados.

A suposição de linearidade significa que o valor esperado do alvo pode ser expresso como a soma das características:

$$price = w_{area}.area + w_{age}.age + b \tag{2.1}$$

Aqui, w_{area} e w_{age} são chamados pesos e b é chamado bias. Os pesos determinam a influência de cada feature na predição. O bias determina o valor da estimativa quando todas as features são zero.

Pode-se dizer que é uma transformação afim de características de entrada, que é caracterizada por uma transformação linear de características via uma soma ponderada, combinada com uma transladação via o bias adicionado. Dado um conjunto de dados, o objetivo é escolher os pesos w e o bias b que, na média, torna as predições de modelo justas aos preços verdadeiros observados nos dados o mais próximo possível.

Em aprendizado de máquina, normalmente trabalha-se com conjunto de altas dimensões, onde é mais conveniente empregar notação de álgebra linear compacta. Quando

as saídas consistem de d características, cada índice é assinado (entre 1 e d) e a predição $\hat{y}$ é expressa como:

$$\hat{y} = w_1 x_1 + ... + w_d x_d + b. \tag{2.2}$$

Coletando todas as características em um vetor $x \in R^d$ e todos os pesos em um vetor $w \in R^d$, pode-se expressar os modelos compactamente pelo produto de w e x:

$$\hat{y} = w^T x + b. \tag{2.3}$$

Em 2.3, o vetor x corresponde às características de um único exemplo. Frequentemente, convém se referir às características do conjunto de dados inteiro de n exemplos via a matriz $X \in R^{n_x d}$. Aqui, X contém uma linha para cada amostra e uma coluna para cada característica. Para uma coleção de características X, as predições $\hat{y} \in R^n$ podem ser expressas pelo produto matriz-vetor:

$$\hat{y} = Xw + b, \tag{2.4}$$

Dadas as características de um conjunto de treinamento X e os rótulos correspondentes y, o objetivo da regressão linear é encontrar o vetor de pesos w e o termo bias b, tal que, dadas as características de um novo exemplo amostrado da mesma distribuição como X, o rótulo do novo exemplo será predito com o menor erro.

Mesmo acreditando que o melhor modelo para predizer y dado x é linear, deve-se esperar encontrar um conjunto de dados real de n exemplos, onde $y^{(i)}$ é exatamente igual a $w^T x^{(i)} + b$ para todo $1 \leq i \leq n$. Por exemplo, quaisquer que sejam os instrumentos usados para observar as características X e os rótulos y, pode haver uma pequena quantidade de erro de medição. Assim, um termo de ruído é incorporado para contar com esses erros.

Antes de buscar pelos melhores parâmetros w e b, é preciso: (i) uma medida de qualidade de algum modelo dado; e (ii) um procedimento para atualizar o modelo para melhorar sua qualidade.

2.1.1 Loss Function

Para ajustar o modelo para os dados é preciso estar de acordo com alguma medida de qualidade. *Loss functions* quantifica a distância entre os valores reais e os valores preditos do alvo. A *loss* vai normalmente ser um número não-negativo onde valores menores são melhores e predições perfeitas incorrem em *loss* de 0.

Para problemas de regressão, a função de *loss* mais comum é o erro quadrático médio ou *Root Mean Square Error* (RMSE). Quando a predição para um exemplo i é $\hat{y}^{(i)}$ e o rótulo verdadeiro correspondente é $y^{(i)}$, o erro quadrático é dado por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.5}$$

Um aspecto importante do RMSE é que ele é sensível a *outliers* devido ao fato de que os erros são elevados ao quadrado. Portanto, grandes erros têm um impacto desproporcional na métrica.

2.1.2 Modelo ElasticNet

O modelo de regressão ElasticNet, proposto por Zou and Hastie (2005a), é uma regressão linear que combina as penalidades de regressão Lasso e Ridge. A função objetivo minimizada durante a otimização é dada por:

$$J(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + \frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2\right) \tag{2.6}$$

onde:

- y_i são os valores reais,
- $\hat{y}_i$ são os valores preditos,
- β_j são os coeficientes do modelo,
- λ é o parâmetro de regularização,
- α controla a mistura entre Lasso (L1) e Ridge (L2).

Para o problema de predição de índice-h investigado neste trabalho,

- y_i corresponde a $h_t^{(i)}$ é o índice-h observado para cada autor i no tempo t,
- $\hat{y}_i$ corresponde a $\hat{h}_{t+\tau}^{(i)} = \beta_0 + \beta_1\sqrt{n_t^{(i)}} + \beta_2 h_t^{(i)} + \beta_3 y_t^{(i)} + \beta_4 j_t^{(i)} + \beta_5 q_t^{(i)}$, o modelo de predição para o instante de tempo $t + \tau$,

2.2 Modelo de Predição de Sucesso Acadêmico

Acuna, Allesina and Kording (2012) propuseram um modelo para predizer o índice-h futuro de pesquisadores baseado nas seguintes características: número de artigos escritos, índice-h atual, quantidade de anos desde a primeira publicação, número de artigos distintos publicados e número de artigos publicados nas revistas *Nature*, *Science*, *Nature Neuroscience*, *Proceedings of the National Academy of Sciences* and *Neuron*.

Começando com neuroscientistas, os autores tentaram predizer o índice-h de cada cientista em um, cinco e dez anos no futuro. Os valores dos coeficientes de cada

característica foram otimizados usando regressão linear com regularização *ElasticNet* e os modelos são apresentados, respectivamente, nas equações 2.7, 2.8 e 2.9. De acordo com os autores, as equações aproximadas são razoavelmente precisas para ciências da vida e menos significativas para outras áreas de pesquisa.

$$h_{+1} = 0.76 + 0.37 + \sqrt{n} + 0.97h - 0.07y + 0.02j + 0.03q \tag{2.7}$$

$$h_{+5} = 4 + 1.58\sqrt{n} + 0.86h - 0.35y + 0.06j + 0.2q \tag{2.8}$$

$$h_{+10} = 8.73 + 1.33\sqrt{n} + 0.48h - 0.41y + 0.52j + 0.82q \tag{2.9}$$

Neste trabalho, baseado no modelo proposto por Acuna, Allesina and Kording (2012), a regressão linear com regularização *ElasticNet* é utilizada para a base de dados de professores de programas de pós-graduação brasileiros em Ciência da Computação. Esta proposta se destaca pelo possibilidade de entender a influência de cada característica na predição do índice-h.

3 TRABALHOS RELACIONADOS

Este capítulo apresenta alguns trabalhos da literatura que abordam o problema de predição de sucesso acadêmico.

Dong, Johnson and Chawla (2016) abordaram duas perguntas feitas por muitos pesquisadores científicos: “Como meu índice h evoluirá ao longo do tempo e quais dos meus artigos previamente ou recentemente publicados contribuirão para isso?”. Eles realizaram duas tarefas relacionadas. Primeiro, desenvolveram um modelo para prever os futuros índices-h dos autores com base em seu impacto científico atual. Segundo, examinaram os fatores que impulsionam os artigos a aumentarem os futuros índices-h previstos dos autores. Aproveitando os fatores relevantes, pode-se prever o índice-h de um autor em cinco anos com um valor R^2 de 0,92 e se um artigo previamente recentemente publicado contribuirá para esse futuro índice-h com uma pontuação F_1 de 0,99. Os autores descobriram que a autoridade no tópico e o local de publicação são cruciais para essas previsões eficazes, enquanto a popularidade do tópico é irrelevante. Além disso, desenvolveram uma ferramenta online que permite que os usuários gerem previsões informadas do índice-h. O trabalho demonstra a previsibilidade do impacto científico e pode ajudar os pesquisadores a aproveitar sua posição acadêmica.

Sinatra *et al.* (2016) quantificaram as mudanças no impacto e na produtividade ao longo de uma carreira científica, descobrindo que o impacto medido por publicações influentes é distribuído aleatoriamente dentro da sequência de publicações de um cientista. O impacto aleatório permitiu que os autores formulassem um modelo estocástico que desvincula os efeitos da produtividade, habilidade individual e sorte, revelando a existência de padrões universais que regem o surgimento do sucesso científico. O modelo atribui um parâmetro individual único, Q, a cada cientista, que se mantém estável durante a carreira e prevê com precisão a evolução do impacto de um cientista, desde o índice-h até as citações cumulativas e reconhecimentos independentes, como prêmios.

Um estudo empírico da força preditiva do índice-h comparado com outros indicadores foi reportado por Hirsch (2007). Os resultados indicam que o índice-h é melhor que outros indicadores considerados - quantidade total de citações, citações por artigo, quantidade total de artigos - em predizer sucesso científico futuro. O estudo foi feito pela seleção de grupos de físicos e a observação do comportamento (medidas estatísticas) dos indicadores em períodos de tempo determinado.

Wen, Wu and Chai (2020) propuseram um modelo de predição baseado em rede neural recorrente GRU (GRU-CPM - citation number prediction model based on the recurrent neural network method with gated recurrent unit). Os autores usaram o método

PCA (*Principal Component Analysis*) para extrair as características de conjuntos de dados reais que são úteis para predizer o número de citações em artigos. Essas características foram usadas como entrada para o GRU-CPM. Os resultados da predição foram comparados com outros modelos de regressão (LSTM, Regressão Linear, SVR e Random Forest) e demonstraram que o GRU-CPM apresenta maior acurácia e convergência mais rápida.

Um estudo feito por Abbasi, Hossain and Owen (2012) investiga se a colaboração interna e externa (a níveis institucional e nacional) entre cientistas associados com impacto na pesquisa está relacionado ao desempenho da pesquisa. Os autores extraíram dados da Scopus, pesquisando por publicações em revistas com a frase 'information science' no título ou palavras-chave ou resumos. O teste de classificação de correlação de Spearman é usado para examinar as hipóteses. As análises dos dados coletados mostram que o impacto das publicações está fortemente associado com indicadores de colaboração baseado nas afiliações dos autores. Entretanto, as colaborações internas, tanto em nível institucional quanto nacional, mostram associações mais fortes com impacto nas publicações do que as colaborações externas.

Um modelo de previsão de impacto científico a longo prazo (10 anos) baseado em extração de características de múltiplos campos foi proposto por Wu *et al.* (2019). O fluxo de trabalho do modelo consiste em engenharia de características e conjunto de modelos. Na engenharia de características, são extraídas características de atributos, de séries temporais e de redes heterogêneas com base em três campos diferentes. Além disso, ao extrair a característica de rede heterogênea, foi proposto um método de avaliação de impacto científico baseado em rede acadêmica heterogênea, que considera tanto o tempo de publicação quanto os fatores de ordem dos autores. No conjunto de modelos, foram ajustados o modelo básico e o modelo de ruído para um conjunto de treinamento diferente para aproveitar ao máximo as informações do conjunto de dados original. Segundo os autores, os resultados demonstram que o modelo proposto pode melhorar de forma estável a precisão da previsão do impacto científico dos pesquisadores e também oferece um padrão de previsão para o problema de previsão a longo prazo.

4 PREVISÃO DE ÍNDICE-H DE PESQUISADORES BRASILEIROS DE PROGRAMAS DE PÓS-GRADUAÇÃO

Neste capítulo, a metodologia de desenvolvimento do trabalho é apresentada. Na seção 4.1, a base de dados é descrita. Nas seções seguintes, os métodos utilizados são apresentados.

4.1 Base de Dados

Os dados de 1511 professores de programas pós-graduação em ciência da computação foram coletados usando a API OpenAlex (PRIEM; PIWOWAR; ORR, 2022). Desses dados, cinco características foram consideradas, inspirado pelo trabalho de Acuna, Allesina and Kording (2012):

- n_t: número de artigos escritos pelo professor até o tempo t.
- h_t: índice-h atual no tempo t.
- y_t: número de anos desde que o professor publicou seu primeiro artigo até o tempo t.
- j_t: número de revistas distintas em que o professor publicou até o tempo t.
- k_t: número de artigos publicados em revistas alto impacto até o tempo t.

4.2 Regressão *ElasticNet*

Regressão *ElasticNet*, como proposta por Zou and Hastie (2005b), é uma regressão linear que combina as penalidades de regressão de Lasso e Ridge. A função objetivo minimizada durante a otimização é dada por:

$$\min_{\beta} \left(\frac{1}{2n} \sum_{i=1}^{n} (h_{t+\tau}^{(i)} - \hat{h}_{t+\tau}^{(i)})^2 + \alpha\lambda \sum_{i=1}^{p} \|\beta_i\|_1 + \frac{\alpha}{2}(1-\lambda) \sum_{i=1}^{p} \|\beta_i\|_2^2 \right), \tag{4.1}$$

onde:

- p é o número de coeficientes,
- n é o número de exemplos,
- $h_{t+\tau}^{(i)}$ é o índice-h observado para o autor i no tempo $t+\tau$,
- $\hat{h}_{t+\tau}^{(i)} = \beta_0 + \beta_1 \sqrt{n_t^{(i)}} + \beta_2 h_t^{(i)} + \beta_3 y_t^{(i)} + \beta_4 j_t^{(i)} + \beta_5 k_t^{(i)}$ é o modelo de predição para o instante de tempo $t+\tau$,
- β_i são os coeficientes do modelo,

- α é a constante que multiplica os termos de penalidade e λ é o parâmetro combinado entre as penalidades $L1$ e $L2$.

5 EXPERIMENTOS E RESULTADOS

A configuração experimental consiste em separar os dados em conjuntos de treino e teste, treinando cada modelo no conjunto de treinamento e avaliando sua performance no conjunto de teste

A regressão *ElasticNet* utiliza uma abordagem de dados estáticos para previsão. Esse modelo usa os dados disponíveis até um ponto fixo para prever o índice-h nos horizontes de $\tau = 1$, $\tau = 2$, $\tau = 3$ e $\tau = 4$ anos.

O modelo *ElasticNet* é treinado utilizando uma validação cruzada estratificada por autor com 10 *folds*. Dentro de cada *fold*, 80% dos dados são usados para treinamento e 20% para teste. A mistura de penalidades L_1 e L_2 é fixada com $\lambda = 0.5$ e $\alpha = 1$. Também são apresentados resultados para dados estratificados sobre notas de programas de pós-graduação e sobre idade acadêmica para verificar a importância dos coeficientes de regressão para diferentes grupos.

O desempenho do modelo *ElasticNet* investigado neste estudo foi avaliado usando RMSE, em horizontes de previsão de 1, 2, 3 e 4 anos, segundo duas divisões. A primeira divisão considera a estratificação por notas dos programas (3, 4, 5, 6, 7), apresentada na Tabela 1. A segunda considera a idade acadêmica, em que foram considerados pesquisadores juniors os que tinham menos de 15 anos de conclusão do doutorado e seniors os que tinham mais de 15. A avaliação do desempenho por idade acadêmica é apresentada na Tabela 2 .

Tabela 1 – Desempenho em termos de RMSE considerando as notas dos programas.

	RMSE (1 ano)	**RMSE (2 anos)**	**RMSE (3 anos)**	**RMSE (4 anos)**
Todas	0.0167 (± 0.0063)	0.0268 (± 0.0116)	0.0357 (± 0.0167)	0.0439 (± 0.0217)
Nota 3	0.0243 (± 0.0095)	0.0384 (± 0.0170)	0.0506 (± 0.0246)	0.0617 (± 0.0317)
Nota 4	0.0150 (± 0.0056)	0.0237 (± 0.0103)	0.0313 (± 0.0148)	0.0382 (± 0.0197)
Nota 5	0.0208 (± 0.0070)	0.0328 (± 0.0131)	0.0434 (± 0.0194)	0.0528 (± 0.0250)
Nota 6	0.0216 (± 0.0073)	0.0346 (± 0.0132)	0.0461 (± 0.0187)	0.0563 (± 0.0241)
Nota 7	0.0201 (± 0.0066)	0.0326 (± 0.0121)	0.0437 (± 0.0179)	0.0536 (± 0.0235)

Tabela 2 – Desempenho em termos de RMSE considerando a idade acadêmica.

	RMSE (1 ano)	**RMSE (2 anos)**	**RMSE (3 anos)**	**RMSE (4 anos)**
Juniors	0.0162 (± 0063)	0.0262 (± 0.0116)	0.0352 (± 0.0169)	0.0434 (± 0.0218)
Seniors	0.0179 (± 0.0065)	0.0283 (± 0.0119)	0.0374 (± 0.0173)	0.0456 (± 0.0225)

No geral, o modelo apresenta um RMSE baixo e consistente, demonstrando sua robustez para fazer as previsões. Em cada divisão considerada, o RMSE apresenta valor mais baixo para um ano e vai aumentando progressivamente a cada ano. Esse comportamento é esperado visto que, a partir do ano atual, o horizonte de tempo vai aumentando, o que pode dificultar a predição dos modelos.

Quando o modelo é avaliado considerando todos os exemplos, o RMSE é menor do que cada conjunto de exemplos por nota, exceto para a nota 4. Esse resultado mostra a robustez do modelo para fazer previsão para todas as notas. Esse resultado também representa o bom desempenho do modelo para fazer a previsão do índice-h dos pesquisadores de todas as idades acadêmicas.

Observando o RMSE por nota, percebe-se que os menores valores correspondem à nota 4. Isso pode ter acontecido porque os programas de nota 4 são o que apresentam maior frequência no conjunto de dados, faciltando o aprendizado do modelo. A Tabela 3 apresenta as frequências de cada nota. Por outro lado, as notas 3 e 6, que são menos frequentes, apresentam os maiores valores de RMSE, apesar de serem muito próximos.

Tabela 3 – Frequência das notas dos programas.

Nota	**Frequência**
4	32,03%
7	24,62%
5	18,66%
3	15,95%
6	8,73%

A Tabela 4 apresenta os valores dos coeficientes de regressão *ElasticNet* para $\hat{h}_{t+\tau}^{(i)} = \beta_0 + \beta_1\sqrt{n_t^{(i)}} + \beta_2 h_t^{(i)} + \beta_3 y_t^{(i)} + \beta_4 j_t^{(i)} + \beta_5 k_t^{(i)}$, cujos β's correspondem respectivamente a:

- β_0: coeficiente de inclinação.
- n_t: número de artigos escritos pelo professor até o tempo t.
- h_t: índice-h atual no tempo t.
- y_t: número de anos desde que o professor publicou seu primeiro artigo até o tempo t.
- j_t: número de revistas distintas em que o professor publicou até o tempo t.
- k_t: número de artigos publicados em revistas alto impacto até o tempo t,

Foram considerados os horizontes de predição $\tau = 1$, $\tau = 2$, $\tau = 3$ e $\tau = 4$ anos, e seus valores-p correspondentes para todo o conjunto de dados (primeira linha da tabela) e para cada nota de programa, de 3 a 7 (linhas seguintes). Na estatística, o valor-p é uma medida que ajuda a determinar a significância dos resultados de um teste estatístico. É comum considerar que se o valor-p $< 0{,}05$, o resultado tem diferença significativa e, valor-p $> 0{,}05$ não tem diferença estatisticamente significativa. Os valores-p maiores que 0,05 estão em negrito na tabela indicando que os respectivos valores dos coeficientes não são estatisticamente significativos e não serão considerados na análise. Observa-se que esses

valores altos de valores-p ocorrem para os mesmos coeficientes em diferentes horizontes de predição, como é o caso do β_1 para todas as notas.

Tabela 4 – Coeficientes de regressão *ElasticNet* para $\hat{h}_{t+\tau}^{(i)} = \beta_0 + \beta_1\sqrt{n_t^{(i)}} + \beta_2 h_t^{(i)} + \beta_3 y_t^{(i)} + \beta_4 j_t^{(i)} + \beta_5 k_t^{(i)}$, considerando os horizontes de predição $\tau = 1$, $\tau = 2$, $\tau = 3$ e $\tau = 4$ anos, e seus valores-p correspondentes. Valores-p maiores que 0,05 estão em negrito.

Modelo	β	β para $\tau = 1$	p-values $\tau = 1$	β para $\tau = 2$	p-values $\tau = 2$	β para $\tau = 3$	p-values $\tau = 3$	β para $\tau = 4$	p-values $\tau = 4$
Todas	β_0	0.0101	0.0000	0.0207	0.0000	0.0318	0.0000	0.0435	0.0000
	β_1	0.0000	**1.0000**	0.0011	**0.8384**	0.0089	**0.1948**	0.0133	**0.1567**
	β_2	1.0249	0.0000	1.0532	0.0000	1.0731	0.0000	1.0608	0.0000
	β_3	-0.0002	**0.8244**	-0.0002	**0.8912**	-0.0000	**0.9265**	0.0010	**0.7140**
	β_4	-0.0303	0.0000	-0.0657	0.0000	-0.0964	0.0000	-0.1169	0.0000
	β_5	-0.0003	**0.9150**	-0.0111	**0.0755**	-0.0168	0.0301	-0.0158	**0.1511**
Nota 3	β_0	0.0111	0.0000	0.0230	0.0000	0.0360	0.0000	0.0500	0.0000
	β_1	0.0228	0.0136	0.0478	0.0021	0.0782	0.0010	0.0886	0.0014
	β_2	1.0180	0.0000	1.0374	0.0000	1.0367	0.0000	1.0153	0.0000
	β_3	-0.0047	**0.2233**	-0.0122	**0.1451**	-0.0158	**0.0671**	-0.0138	**0.1888**
	β_4	-0.0369	0.0005	-0.0666	0.0000	-0.0861	0.0010	-0.0918	0.0029
	β_5	-0.0020	**0.8563**	-0.0255	**0.1679**	-0.0338	**0.2372**	-0.0498	**0.1651**
Nota 4	β_0	0.0081	0.0000	0.0172	0.0000	0.0270	0.0000	0.0371	0.0000
	β_1	0.0002	**0.9647**	-0.0002	**0.8647**	-0.0004	**0.4110**	-0.0154	**0.1909**
	β_2	1.0304	0.0000	1.0579	0.0000	1.0850	0.0000	1.0816	0.0000
	β_3	-0.0010	**0.4987**	-0.0025	**0.4070**	-0.0020	**0.5820**	0.0004	**0.6795**
	β_4	-0.0229	0.0000	-0.0455	0.0000	-0.0545	0.0000	-0.0602	0.0002
	β_5	-0.0042	**0.3406**	-0.0102	**0.1025**	-0.0131	**0.1441**	-0.0122	**0.3175**
Nota 5	β_0	0.0133	0.0000	0.0273	0.0000	0.0423	0.0000	0.0577	0.0000
	β_1	0.0041	**0.5543**	0.0193	**0.1440**	0.0348	**0.0636**	0.0622	0.0025
	β_2	1.0072	0.0000	1.0070	0.0000	0.9728	0.0000	0.9402	0.0000
	β_3	-0.0037	**0.1887**	-0.0097	0.0336	-0.0147	0.0280	-0.0165	**0.0668**
	β_4	0.0101	**0.2670**	-0.0304	**0.0771**	-0.0342	**0.1206**	-0.0455	0.0357
	β_5	0.0000	**1.0000**	0.0026	**0.8353**	0.0111	**0.5026**	0.0272	**0.2204**
Nota 6	β_0	0.0151	0.0000	0.0322	0.0000	0.0495	0.0000	0.0658	0.0000
	β_1	0.0004	**0.9655**	0.0037	**0.7794**	-0.0001	**0.8109**	0.0011	**0.6573**
	β_2	1.0275	0.0000	1.0532	0.0000	1.0801	0.0000	1.0786	0.0000
	β_3	-0.0021	**0.5929**	-0.0069	**0.2508**	-0.0080	**0.3512**	-0.0111	**0.3384**
	β_4	-0.0299	0.0004	-0.0622	0.0000	-0.0893	0.0000	-0.1149	0.0000
	β_5	-0.0079	**0.1153**	-0.0180	0.0335	-0.0276	0.0393	-0.0361	0.0387
Nota 7	β_0	0.0158	0.0000	0.0323	0.0000	0.0490	0.0000	0.0658	0.0000
	β_1	0.0000	**1.0000**	0.0001	**0.9885**	0.0024	**0.8668**	0.0036	**0.8567**
	β_2	1.0222	0.0000	1.0382	0.0000	1.0621	0.0000	1.0447	0.0000
	β_3	-0.0014	**0.4829**	-0.0029	**0.4050**	-0.0037	**0.4203**	-0.0014	**0.6950**
	β_4	-0.0418	0.0000	-0.0884	0.0000	-0.1218	0.0000	-0.1519	0.0000
	β_5	-0.0020	**0.6998**	-0.0100	**0.2826**	-0.0142	**0.2060**	-0.0153	**0.3132**

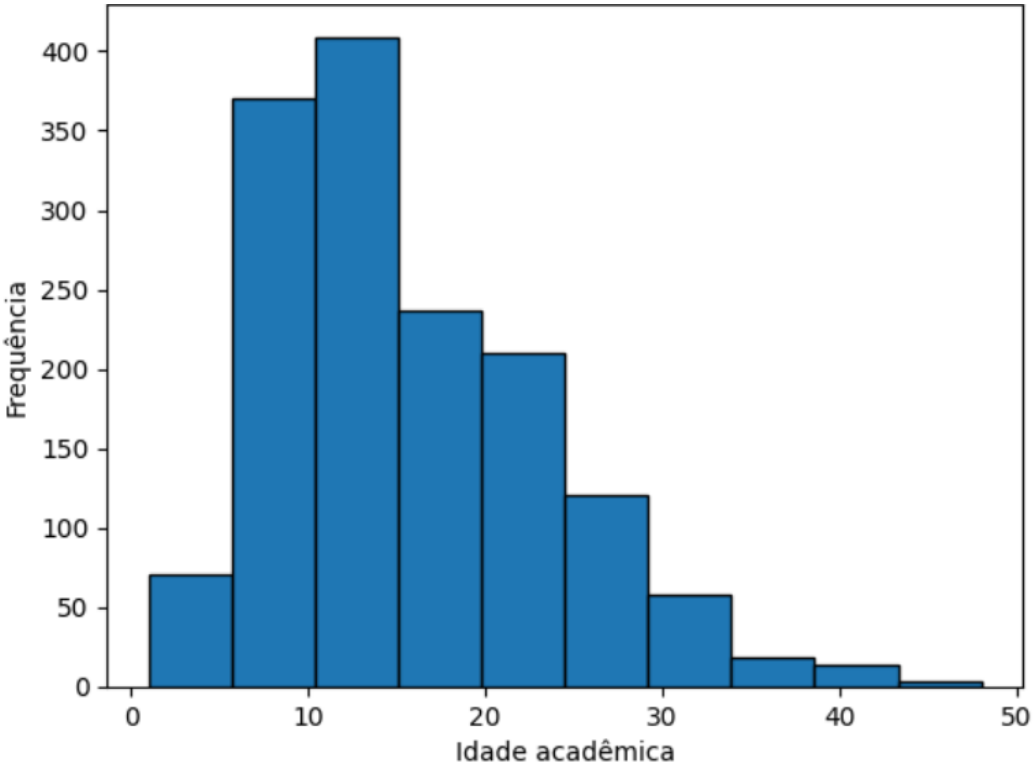
A Figura 1 apresenta o histograma dos dados por idade acadêmica. Os coeficientes de regressão *ElasticNet* para $\hat{h}_{t+\tau}^{(i)} = \beta_0 + \beta_1\sqrt{n_t^{(i)}} + \beta_2 h_t^{(i)} + \beta_3 y_t^{(i)} + \beta_4 j_t^{(i)} + \beta_5 k_t^{(i)}$, considerando a idade acadêmica, e seus valores-p correspondentes, são apresentados na Tabela 5. Nesses resultados, observa-se que os valores-p ficaram abaixo de 0,05 para a maioria dos casos,

tendo maior significância estatística do que a divisão dos dados por notas dos programas. Os valores altos estão em negrito na tabela, correspondendo ao β_3 para pesquisadores juniors e β_5 para os seniors, e não serão considerados na análise.

Tabela 5 – Coeficientes de regressão *ElasticNet* para $\hat{h}_{t+\tau}^{(i)} = \beta_0 + \beta_1\sqrt{n_t^{(i)}} + \beta_2 h_t^{(i)} + \beta_3 y_t^{(i)} + \beta_4 j_t^{(i)} + \beta_5 k_t^{(i)}$, considerando a idade acadêmica, e seus valores-p correspondentes.

Modelo	β	β para $\tau = 1$	p-values $\tau = 1$	β para $\tau = 2$	p-values $\tau = 2$	β para $\tau = 3$	p-values $\tau = 3$	β para $\tau = 4$	p-values $\tau = 4$
Juniors	β_0	0.0077	0.0000	0.0160	0.0000	0.0249	0.0000	0.0347	0.0000
	β_1	0.0127	0.0020	0.0343	0.0000	0.0625	0.0000	0.0911	0.0000
	β_2	1.0253	0.0000	1.0515	0.0000	1.0643	0.0000	1.0415	0.0000
	β_3	-0.0000	**0.8389**	0.0004	**0.7216**	0.0014	**0.5276**	0.0040	**0.3486**
	β_4	-0.0372	0.0000	-0.0875	0.0000	-0.1276	0.0000	-0.1620	0.0000
	β_5	-0.0065	**0.3143**	-0.0253	0.0077	-0.0341	0.0066	-0.0437	0.0025
Seniors	β_0	0.0142	0.0000	0.0296	0.0000	0.0448	0.0000	0.0599	0.0000
	β_1	-0.0037	**0.4207**	-0.0274	0.0000	-0.0476	0.0000	-0.0618	0.0000
	β_2	1.0174	0.0000	1.0472	0.0000	1.0658	0.0000	1.0645	0.0000
	β_3	-0.0059	0.0000	-0.0084	0.0054	-0.0100	0.0078	-0.0103	**0.0706**
	β_4	-0.0229	0.0000	-0.0380	0.0000	-0.0425	0.0000	-0.0442	0.0003
	β_5	0.0000	**1.0000**	-0.0007	**0.9158**	-0.0023	**0.7997**	-0.0011	**0.9242**

Figura 1 – Histograma da idade acadêmica.



Para pesquisadores juniors, o coeficiente β_1, referente ao número de publicações ($\sqrt{n}$), possui um impacto positivo e aumenta com o aumento dos horizontes de previsão, com valores $0,0127$, $0,0343$, $0,0625$ e $0,0911$ para 1, 2, 3 e 4 anos, respectivamente. Isso indica que, a partir do ano atual, quanto maior o horizonte de predição, maior a influência

desse coeficiente para prever o índice-h. Já para pesquisadores seniors, a influência negativa se torna mais forte com o aumento dos horizontes de previsão

O índice-h atual (h_t), β_2, tem um coeficiente positivo e acima de 1 para todos os grupos, indicando que esse coeficiente tem um forte impacto para fazer a predição.

Para os pesquisadores seniors, os valores de β_3, idade acadêmica (y_t), foram negativos, próximos a $0,01$. Isso pode indicar que, para esses pesquisadores, o índice-h tenha um crescimento mais lento com o passar dos anos, indicando uma estabilidade de produção.

O número de artigos em revistas distintas j_t, afeta negativamente o índice-h, em todas as situações, refletindo potenciais reduções na produtividade ao longo do tempo. O mesmo acontece com as publicações em periódicos de alto impacto (k_t) para pesquisadores juniors.

Para comparar os modelos *ElasticNet* obtidos neste estudo com o modelo obtido por Acuna, Allesina and Kording (2012) foram considerados os horizontes 1 e 4 para comparar com 1 e 5 do último (Equações 5.1 e 5.2). Observa-se que o número de publicações ($\sqrt{n}$) é menor para predição de 1 ano e aumenta para a predição de 4 anos, como acontece no modelo de Acuna para 1 e 5 anos, respectivamente, embora seus valores sejam maiores. Observa-se uma maior importância em relação ao coeficiente do parâmetro h_t (índice-h atual), para os diferentes horizontes de predição. Os valores de y_t são negativos para seniors, como o modelo de Acuna, apesar serem maiores. Vale lembrar que o modelo de Acuna usou dados de pesquisadores de ciências da vida.

$$h_{+1} = 0.76 + 0.37\sqrt{n} + 0.97h - 0.07y + 0.02j + 0.03q \tag{5.1}$$

$$h_{+5} = 4 + 1.58\sqrt{n} + 0.86h - 0.35y + 0.06j + 0.2q \tag{5.2}$$

6 CONCLUSÕES

Neste estudo, levantou-se a hipótese de que o modelo proposto por Acuna, Allesina and Kording (2012) para prever o índice-h dos autores poderia ser efetivamente aplicado a professores dos programas de pós-graduação brasileiros em Ciência da Computação, com os devidos ajustes e treinamento usando dados históricos. Os resultados experimentais deste estudo confirmaram essa hipótese, demonstrando que o modelo *ElasticNet*, adaptado ao contexto específico, teve um desempenho satisfatório e serviu como *baseline*. O modelo mostrou tendências consistentes nos valores dos coeficientes para diferentes horizontes de previsão, alcançando um RMSE de 0.0167, 0.0268, 0.0357 e 0.0439 para o horizonte de 1, 2, 3 e 4 anos, respectivamente, para todos os professores.

Também foram realizados teste estratificados por nota CAPES dos programas e por idade acadêmica dos pesquisadores. Para cada grupo, o valor de RMSE se manteve na mesmo ordem de grandeza, com poucas variações. Também foi feita uma análise dos valores dos coeficientes obtidos. Nesta análise, alguns coeficientes apresentaram valores-p altos, indicando que os valores obtidos não possuem significância estatística, e, portanto, não contribuem na predição. Para esses casos, futuramente o modelo será treinado novamente sem esses coeficientes.

A simplicidade e interpretabilidade do modelo *ElasticNet* podem torná-lo uma ferramenta valiosa para pesquisadores e agências que precisam avaliar uma área científica ou um Programa de Pós-Graduação, fornecendo informações de alto nível para uma melhor tomada de decisão.

REFERÊNCIAS

ABBASI, A.; HOSSAIN, L.; OWEN, C. Exploring the relationship between research impact and collaborations for information science. *In*: **2012 45th Hawaii International Conference on System Sciences**. [*S.l.: s.n.*], 2012. p. 774–780.

ACUNA, D. E.; ALLESINA, S.; KORDING, K. P. Predicting scientific success. **Nature**, Nature Publishing Group UK London, v. 489, n. 7415, p. 201–202, 2012.

BOURDIEU, P. **Science of science and reflexivity**. [*S.l.: s.n.*]: Polity Press, 2004.

CAPES. **Plataforma Sucupira**. 2024. Accessed on 26.05.2024. Available at: <https://sucupira-v2.capes.gov.br/sucupira4/>.

DONG, Y.; JOHNSON, R. A.; CHAWLA, N. V. Can scientific impact be predicted? **IEEE Transactions on Big Data**, v. 2, n. 1, p. 18–30, 2016.

FITZGERALD, C. *et al.* Temporal dynamics of faculty hiring in mathematics. **Humanities and Social Sciences Communications**, Springer Nature, v. 10, n. 1, p. 247, 2023.

FORTUNATO, S. *et al.* Science of science. **Science**, v. 359, n. 6379, p. eaao0185, 2018. Available at: <https://www.science.org/doi/abs/10.1126/science.aao0185>.

HIRSCH, J. E. Does the h index have predictive power? **Proceedings of the National Academy of Sciences**, v. 104, n. 49, p. 19193–19198, 2007. Available at: <https://www.pnas.org/doi/abs/10.1073/pnas.0707962104>.

JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in R**. [*S.l.: s.n.*]: Springer Publishing Company, Incorporated, 2014. ISBN 1461471370.

MURRAY, D. *et al.* Unsupervised embedding of trajectories captures the latent structure of scientific migration. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 120, n. 52, p. e2305414120, 2023.

PRICE, D. J. D. S. **Little science, big science**. [*S.l.: s.n.*]: Columbia university press, 1963.

PRIEM, J.; PIWOWAR, H. A.; ORR, R. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. **arXiv preprint arXiv:2205.01833**, 2022. ArXiv:2205.01833 [cs.DL]. Available at: <https://doi.org/10.48550/arXiv.2205.01833>.

SINATRA, R. *et al.* Quantifying the evolution of individual scientific impact. **Science**, v. 354, n. 6312, p. aaf5239, 2016. Available at: <https://www.science.org/doi/abs/10.1126/science.aaf5239>.

SUGIMOTO, C. R. *et al.* Scientists have most impact when they're free to move. **Nature**, Nature Publishing Group, v. 550, p. 29–31, 2017.

WAPMAN, K. H. *et al.* Quantifying hierarchy and dynamics in us faculty hiring and retention. **Nature**, Nature Publishing Group, v. 610, n. 7930, p. 120–127, 2022.

WEN, J.; WU, L.; CHAI, J. Paper citation count prediction based on recurrent neural network with gated recurrent unit. *In*: **2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)**. [*S.l.: s.n.*], 2020. p. 303–306.

WU, Z. *et al.* Predicting long-term scientific impact based on multi-field feature extraction. **IEEE Access**, v. 7, p. 51759–51770, 2019.

ZHANG, A. *et al.* **Dive into Deep Learning**. [*S.l.: s.n.*]: Cambridge University Press, 2023. <https://D2L.ai>.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, Oxford University Press, v. 67, n. 2, p. 301–320, 2005.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, Oxford University Press, v. 67, n. 2, p. 301–320, 2005.