UNIVERSIDADE FEDERAL DO MARANHÃO

Curso de Ciência da Computação

José Vinícius Reis de Almeida

Desenvolvimento de um modelo não linear do sucesso futuro de curto prazo de cientistas da física usando regressão simbólica

São Luís - MA

2023

José Vinícius Reis de Almeida

Desenvolvimento de um modelo não linear do sucesso futuro de curto prazo de cientistas da física usando regressão simbólica

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

Reis de Almeida, José Vinícius.
Desenvolvimento de um modelo não linear do sucesso futuro de curto prazo de cientistas da física usando regressão simbólica / José Vinícius Reis de Almeida. - 2023.
34 f.

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

1. Aprendizado de máquina. 2. Predições científicas. 3. Regressão simbólica. I. de Abreu Batista Júnior, Antônio. II. Título.

José Vinícius Reis de Almeida

Desenvolvimento de um modelo não linear do sucesso futuro de curto prazo de cientistas da física usando regressão simbólica

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

Trabalho aprovado em 09 de Janeiro de 2024.

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

Prof. Dr. André Borges Cavalcante
Examinador Externo
Universidade Federal do Maranhão

Prof. Dr. Luciano Reis Coutinho
Examinador Interno

Profa. Msc. Yonara Costa Magalhães
Examinadora Interna

São Luís - MA
2023

Dedico aos meus pais, por todo o amor e cuidado e por ensinarem a mim o valor do esforço e da dedicação.

Agradecimentos

A Deus, pela saúde e pelo dom da vida.

Ao meu orientador, o professor Antônio de Abreu Batista Júnior, pela confiança, paciência, encorajamento e valiosos conselhos dados durante a graduação e, principalmente, na elaboração deste trabalho.

À minha noiva, pelas palavras de perseverança e incentivo, e também pela atenção dada a mim nos bons e maus momentos.

Aos professores Rodrigo Gustavo de Souza e Lindalva Martins Maia Maciel por direta ou indiretamente terem contribuído na minha carreira acadêmica e profissional.

Aos professores do curso de Ciência da Computação da Universidade Federal do Maranhão pelo conhecimento e dedicação.

Por fim, agradeço a todas as pessoas que me ajudaram durante esta caminhada.

"Conhecimento não é aquilo que você sabe, mas o que você faz com aquilo que sabe."
(Aldous Huxley)

Resumo

A predição do sucesso científico de curto prazo de físicos, medido pelo número de artigos com 3 citações do cientista nos próximos anos, é uma tarefa importante para diversos atores, como instituições de pesquisa, agências de fomento e empresas. Para esta tarefa, modelos de redes neurais artificiais são capazes de fornecer resultados precisos, mas são complexos e difíceis de interpretar. Modelos de regressão linear múltipla, por outro lado, são mais fáceis de interpretar, mas pressupõem que a relação entre variáveis independentes e dependentes é linear, o que nem sempre é o caso. Neste trabalho, propomos um novo modelo não linear do índice I3 futuro de físicos, baseado em regressão simbólica. O modelo foi treinado e testado com um conjunto de dados que inclui informações sobre o desempenho acadêmico, a experiência de pesquisa e as características pessoais de físicos. O modelo proposto obteve um RMSE menor que o do modelo linear, indicando que é superior em termos de precisão.

Palavras-chave: aprendizado de máquina; regressão simbólica; predições científicas

Abstract

Predicting the short-term scientific success of physicists, measured by the number of articles with 3 citations of the scientist in the next three years, is an important task for a variety of stakeholders such as research institutions, funding organizations and companies. For this task, artificial neural network models can provide accurate results, but they are complex and difficult to interpret. Multiple linear regression models, on the other hand, are easier to interpret, but they assume that the relationship between independent and dependent variables is linear, which is not always the case. In this work, we propose a new non-linear model of the future I3 index of physicists based on symbolic regression. The model was trained and tested on a dataset containing information about the academic performance, research experience and personal characteristics of physicists. The proposed model achieved a lower RMSE than the linear model, suggesting that it is superior in terms of accuracy.

Keywords: machine learning; symbolic regression; scientific predictions

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

IA	*Inteligência Artificial*
RNA	*Rede Neural Artificial*
RS	*Regressão Simbólica*

Sumário

1 Introdução

Com o crescimento exponencial da produção científica e do contingente de pesquisadores, a predição precisa do impacto científico de cientistas emerge como um dos maiores desafios da pesquisa moderna. Este problema impacta a eficiência da pesquisa, a tomada de decisões e a avaliação científica, atraindo a atenção de cientistas de diversas áreas (XIA; LI; LI, 2023).

Nas últimas décadas, numerosas tentativas foram realizadas (ACUNA; ALLESINA; KORDING, 2012a; KUPPLER, 2022; HOU; WU; XIE, 2022; KUMAR; BHOWMICK; PAIK, 2023; LAURANCE et al., 2013) visando ao desenvolvimento de modelos de previsão precisos.

Contudo, a pesquisa tem tendido a se concentrar em modelos de redes neurais artificiais, em detrimento de outros modelos não lineares. Um problema adicional é que modelos de redes neurais artificiais são complexos e de difícil compreensão quanto à maneira como tomam suas decisões.

Neste trabalho, apresentamos um modelo de regressão simbólica, não linear nas entradas, que pode prever o sucesso futuro de curto prazo de cientistas da física, medido pelos seus índices I3[1], com precisão semelhante aos modelos existentes, mas com a vantagem de ser interpretável.

Nossa hipótese é que um modelo de regressão simbólica não linear pode prever os índices I3 futuros com mais precisão do que modelos de regressão linear. Para testar esta hipótese, desenvolvemos um modelo de regressão simbólica não linear, uma rede neural e um modelo de regressão linear e comparamos a precisão, a interpretabilidade e a robustez dos três modelos usando os mesmos dados de publicações científicas.

Objetivos

O objetivo geral deste trabalho consiste em desenvolver um modelo de regressão não linear para predizer o índice I3 futuro de cientistas da física usando regressão simbólica. Mais especificamente, pretende-se:

- Desenvolver um modelo que seja capaz de prever o índice I3 com uma precisão comparável à de uma rede neural;

[1] O índice I3 é um indicador do sucesso futuro de um cientista, medido pelo número de publicações com, no mínimo, 3 citações cada, contados nos três anos seguintes ao ano da predição.

- Comparar a precisão do modelo de regressão simbólica não linear com a precisão de um modelo de regressão linear;
- Identificar as características do modelo de regressão simbólica não linear que estão associadas à sua precisão.

2 Fundamentação Teórica

Os principais conceitos utilizados ao longo do estudo são explorados neste capítulo, incluindo informações sobre o conjunto de dados dos cientistas da física e as técnicas empregadas para realizar a predição do sucesso futuro destes cientistas. A escolha dos algoritmos levou em consideração as características do problema em questão, no qual se pretende realizar predições sobre o sucesso de curto prazo de pesquisadores, medido pelo índice I3, a partir de características do cientista.

2.1 Problemas de regressão

Quando desejamos encontrar correlações entre variáveis dependentes e independentes, prevendo uma resposta a partir de um conjunto de variáveis de entrada, estamos diante de um problema de regressão (ACEBES et al., 2022). Nessa seção são detalhadas algumas técnicas aplicadas na modelagem de problemas dessa natureza.

Regressão linear

Graficamente representada por um reta que ajusta um conjunto de pontos, a regressão linear busca estabelecer uma relação linear não determinística (FAHRMEIR et al., 2013) entre uma variável dependente, a resposta, e uma ou mais variáveis independentes, as preditoras (MAULUD; ABDULAZEEZ, 2020). Na regressão linear simples, o modelo possui apenas uma variável preditora para a resposta. Quando a regressão envolver um conjunto de variáveis independentes, o modelo será de regressão linear múltipla, sendo representado pela expressão a seguir (UYANIK; GÜLER, 2013):

$$y = \beta_0 + \beta_i x_i + \cdots + \beta_n x_n + \epsilon$$

Em que:

- y é a variável dependente;
- x_i e x_n são variáveis independentes;
- β_0 e β_i representam os parâmetros do modelo, respectivamente, o intercepto e o declive, ou coeficiente de regressão;
- ϵ é o erro , ou seja, a variação entre o valor real de y e o valor predito pelo modelo.

Um conjunto de pressupostos deve ser observado para que o modelo seja construído corretamente.

- **Linearidade** : o modelo é limitado a relações lineares, sendo as predições o resultado de uma combinação linear das variáveis independentes;
- **Homoscedasticidade**: a variação dos erros em torno da reta de regressão não depende dos valores da variável independente;
- **Normalidade** : os resíduos do modelo devem ter uma distribuição normal;
- **Independência**: os resíduos devem ser independentes entre si;
- **Variáveis independentes fixas**: as variáveis dependentes são tratadas como "constantes"e não como variáveis estatísticas;
- **Ausência de multicolinearidade**: Não deve existir uma alta correlação entre as variáveis independentes.

A regressão linear pode não performar adequadamente quando o conjunto de variáveis independentes é muito extenso, pois essa condição aumentar o risco de haver alguma correlação entre as variáveis, levando ao problema da multicolinearidade. Algumas abordagens tem sido utilizadas com sucesso para contornar o problema (DUZAN; SHARIFF, 2015; SHRESTHA, 2020).

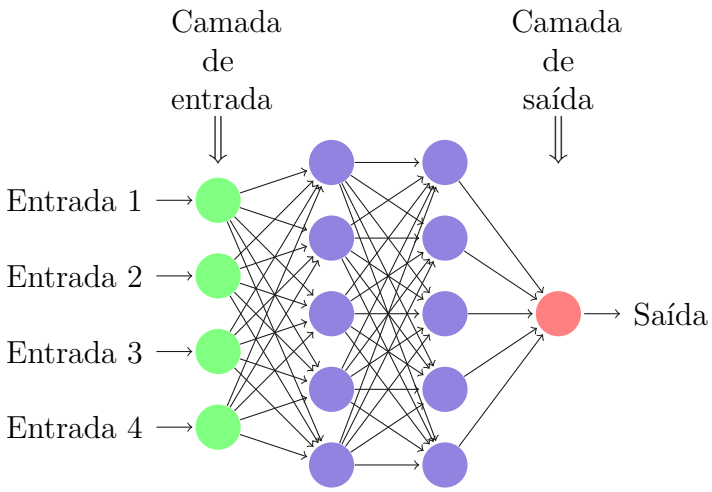
Redes neurais

Uma rede neural artificial (RNA) ,ou simplesmente rede neural, é um modelo de aprendizado de máquina inspirado no funcionamento do cérebro humano (MCCULLOCH; PITTS, 1943). É constituída por uma rede de nós interligados (neurônios) em uma estrutura de camadas, que utilizam conexões ponderadas para analisar e enviar informações. A Figura 1 ilustra o esquema de uma RNA que contém quatro camadas: a camada de entrada , contendo quatro neurônios, a camada de saída com um neurônio e duas camadas intermediárias, com cinco neurônios cada.

Os neurônios são as unidades de processamento de uma RNA e produzem um comportamento inteligente a partir seguinte dinâmica (MCCULLOCH; PITTS, 1943):

- A entrada recebe sinais
- É aplicado um peso a cada sinal, determinado o seu impacto na saída da unidade
- Realiza-se a soma ponderada dos sinais, produzindo um nível de atividade

Figura 1 – Modelo de Rede Neural.



Fonte: Acervo do autor

- A unidade de processamento gera uma resposta de saída específica quando o nível de atividade ultrapassa um limiar pré-determinado

Uma característica importante da RNA é o fato de serem aproximadores universais de funções (KOLMOGOROV, 1956; CYBENKO, 1989), sendo capazes de representar e realizar operações de uma função específica, entretanto, são ineficientes na tarefa de desvendar de forma adequada a natureza e os atributos da função em questão.

Regressão simbólica

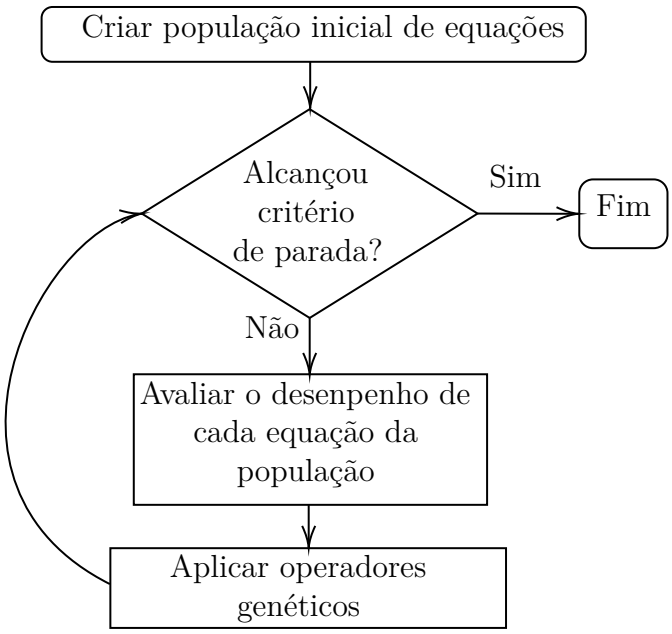
Diferente de técnicas que ajustam os parâmetros para uma equação de uma determinada forma, a exemplo da regressão linear e da rede neural artificial, a regressão simbólica (KOZA, 1994), por outro lado, busca ao mesmo tempo por equações e parâmetros em um espaço de expressões matemáticas (SCHMIDT; LIPSON, 2009), através de estratégias de programação genética. A equação (ou expressão) final gerada pode descrever relacionamentos não lineares com precisão próxima à de outros modelos de difícil interpretação, além de possibilitar, com maiores detalhes, a compreensão do seu comportamento por seres humanos. (FRANCA et al., 2023).

Dado um conjunto de pares de entradas e saídas $\{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$, onde X_1, X_2 e X_n é um subconjunto formado pelas variáveis independentes, também conhecidas como características, e y_1, y_2 e y_n é um subconjunto formato pela variável dependente, ou variável alvo, matematicamente, a regressão simbólica quer encontrar uma função $f'(x)$, que gera um conjunto de saída próximo à uma função $f(x)$ desconhecida, ou seja, para cada valor de (X_i, y_i), $f'(X_i) \approx y_i$.

A Figura 2 ilustra, de forma simplificada, o processo usado pela regressão simbólica

para encontrar uma expressão.

Figura 2 – Fluxograma do algoritmo de programação genética.



Fonte: Acervo do autor

Inicialmente, é criada uma população inicial de equações randomicamente, combinando operações algébricas, constantes e funções. Caso alguma dessas equações atenda aos critérios de parada, o processo é finalizado. Caso contrário, o desempenho de cada equação é avaliado usando os dados de treino para calcular sua medida de aptidão, que indica quais equações são mais precisas. Aplicam-se, então, operadores genéticos de mutação e cruzamento, gerando novas equações. O processo finaliza quando um número máximo de iterações é atingido ou quando se alcança um nível desejado na precisão. Árvores de expressão são comumente usadas para representar as funções geradas durante o processo. Os nós dessas árvores estão associados a operadores e a quantidades de operandos deste operador correspondem ao número de ramificações que saem de cada nó.

3 Trabalhos Relacionados

Neste capítulo apresentamos os trabalhos relacionados à predição do sucesso científico, usando diversas abordagens envolvendo algoritmos de IA.

Acuna, Allesina e Kording (2012b) apresentaram um modelo linear do sucesso futuro de cientistas da vida usando regressão linear. Entretanto, foi demonstrado que o modelo favorece cientistas com mais experiência, atribuindo pontuações mais altas a eles do que a cientistas no início de carreira.

Enfoque diferente foi dado por Sarigöl et al. (2014), em que foi proposto um classificador *Random Forest* para prever com alta precisão se um artigo seria bastante citado cinco anos após sua publicação. O conjunto de dados da análise continha mais de 100.000 publicações da área de ciência da computação. Os resultados sugeriram que entre a centralidade do autor e o sucesso das citações existe uma forte dependência estatística. Todavia,este é um modelo que envolve a classificação e, possivelmente, não poderia ser adotado para tarefas de regressão com a mesma performance.

Gafarov et al. (2021), por outro lado, propuseram um modelo de rede neural complexa para prever o sucesso pessoal com base na atividade nas redes sociais. O modelo foi treinado em um conjunto de dados de dados de usuários de redes sociais. Os resultados do modelo mostraram que ele é capaz de prever o sucesso pessoal com precisão. No entanto, o modelo é difícil de ser adaptado para físicos, pois é baseado em dados de redes sociais.

Diferentemente dos estudos anteriores que usaram redes neurais artificiais, e outros modelos, que são complexos e difíceis de interpretar, neste trabalho usamos regressão simbólica, um método que usa regras matemáticas para gerar modelos do sucesso futuro de físicos mais simples e interpretáveis.

4 Configuração experimental

4.1 Conjunto de dados

Utilizou-se neste estudo o conjunto de dados *APS*, mantido pela *American Physical Society*[1]. Ele cobre mais de um século de artigos da física publicados em *Physical Review*, uma coleção de periódicos revisados por pares publicando artigos científicos originais a partir de todas as áreas da física interdisciplinar, pura e aplicada.

O conjunto de dados contém registros de artigos publicados nos periódicos de Physical Review de 1893 a 2018, cada um identificado com um rótulo numérico único. Para cada artigo, as seguintes informações estão disponíveis: o título do artigo, a data de publicação, os nomes e afiliações de cada um dos autores, e a lista dos rótulos numéricos de artigos sendo citados (lista de referências).

O experimento foi realizado dentro do ambiente *Jupyter Notebook*[2]. Da biblioteca *scikit-learn* [3], foram usadas A classe *LinearRegression* e *MLPRegressor*, respectivamente, na construção dos modelo de regressão linear múltipla e rede neural. Na regressão simbólica, utilizou-se a biblioteca ITEA[4].

4.2 Configuração do modelo de regressão linear

```
LinearRegression(*,fit_intercept=True,
            copy_X=True, n_jobs=None, positive=False)
```

1. `fit_intercept=True` : permite calcular o intercepto do modelo;
2. `copy_X=True` : impede que o conjunto das variáveis independentes seja sobrescrito;
3. `n_jobs=None` : o número de tarefas usadas para realizar a computação, o valor está definido para 1;
4. `positive=False` : os coeficientes da reta de regressão podem ser positivos ou negativos.

[1] https://journals.aps.org/datasets
[2] https://jupyter.org/
[3] https://scikit-learn.org
[4] https://itea-python.readthedocs.io

4.3 Configuração da rede neural

```
MLPRegressor(alpha=0.001, epsilon=0.01, hidden_layer_sizes=(1000,),
             learning_rate_init=1e-05, max_iter=10000,
             n_iter_no_change=30)
```

1. `alpha=0.001` : parâmetro que representa a força da regularização L2, que é uma técnica útil para melhorar o desenpenho de modelos de regressão através da penalização de pesos do modelo, reduzindo a sua magnitude;
2. `activation=relu` : a função de ativação da camada oculta escolhida foi a *Rectified Linear Unit function*, devido ao seu bom desempenho na maioria dos testes;
3. `solver=adam` : na otimização dos pesos, optou-se pelo otimizador adam, pois sabe-se que ele performa bem em conjuntos de dados relativamente grandes;
4. `epsilon=0.01` : valor para estabilizar numericamente o otimizador adam;
5. `hidden_layer_sizes=(1000,)` : tamanho da camada oculta e quantidade de camadas, no caso, a rede tem apenas uma camada oculta com 1000 neurônios;
6. `learning_rate_init=1e-05` : taxa inicial de aprendizagem, controlando o tamanho do passo na atualização dos pesos;
7. `max_iter=10000` : número máximo de iterações;
8. `n_iter_no_change=30` : número máximo de épocas, ,caso a melhoria não seja alcançada.

Alguns parâmetros da rede neural foram obtidos através do método de seleção de hiperparâmetros *HalvingRandomSearchCV*, que utilizou validação cruzada para encontrar valores para os seguintes parâmetros: `n_iter_no_change`, `learning_rate_init`, `hidden_layer_sizes`, `epsilon` e `alpha`.

4.4 Configuração do modelo de regressão simbólica

```
ITEA_regressor(expolim=(0, 2), fitness_f='rmse', gens=900,
        labels=array(['co', 'co_5', 'h', 'h_5', 'h_3', 'c',
        'c_5', 'c_3', 'i_10_5', 'i_5_5', 'i_3_5', 'v_5',
        'p_5', 'p_3', 'y2y1_dif'], dtype=object),
        max_terms=7, popsize=15, random_state=39,
        simplify_method='simplify_by_coef',
        tfuncs={'exp': <ufunc 'exp'>,
                'id': <function <lambda> at 0x0000017E634F5A80>,
                'log': <ufunc 'log'>,
                'sqrt.abs': <function <lambda> at 0x0000017E60F7A200>},
        tfuncs_dx={'exp': <ufunc 'exp'>,
                'id': <function <lambda> at 0x0000017E634F5940>,
                'log': <function <lambda> at 0x0000017E634F59E0>,
                'sqrt.abs': <function <lambda> at 0x0000017E634F58A0>},
        verbose=10)
```

1. `expolim=(0, 2)` : tupla especificando os limites dos expoentes para a expressão final gerada;
2. `fitness_f='rmse'` : *String* com o método de ajuste das expressões;
3. `gens=900` : número de gerações do processo evolucionário;
4. `labels=array(['co', 'co_5', 'h', 'h_5', 'h_3', 'c', 'c_5', 'c_3', 'i_10_5', 'i_5_5', 'i_3_5', 'v_5', 'p_5', 'p_3', 'y2y1_dif'], dtype=object)` : rótulos dos dados usados no processo evolucionário e na construção da expressão final;
5. `max_terms=7` : o número máximo de termos definidos na expressão;
6. `popsize=15` : o tamanho da população;
7. `random_state=39` : o estado randômico (esse valor foi igual durante o treinamento dos três modelos);
8. `simplify_method='simplify_by_coef'` : define de que forma a expressão final será simplificada, nesse caso, foi escolhida a simplificação por coeficientes;
9. `tfuncs` : um dicionário com o conjunto de funções escolhidas para a geração da população inicial;
10. `tfuncs_dx` : a derivada das funções elencadas no parâmetro `tfuncs`;

11. `verbose=10` : apenas mostra iinformações do processo na tela.

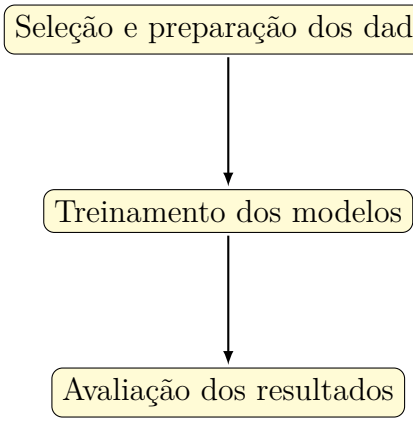
Com estes parâmetros foi gerado o seguinte modelo:

```
Final expression:
 -0.566*sqrt.abs(i_5_5^2) +
0.089*sqrt.abs(co * co_5^2 * c_3) +
-0.114*sqrt.abs(co_5^2 * h_5 * h_3 * co_53) +
0.0*sqrt.abs(h * h_5^2 * c * c_3^2 * i_10_5 * co_51^2 * co_52) +
0.615*sqrt.abs(h_5 * co_53^2) +
2.624*sqrt.abs(i_5_5) +
0.001*sqrt.abs(co * co_5^2 * h_5 * c_3 * co_53^2) +
-1.145
```

5 Metodologia

A metodologia considerada adequada para a avaliação dos modelos é abordada neste capítulo. Na Figura 3 é mostrado um fluxograma da metodologia utilizada para comparar o desempenho de modelos de aprendizado de máquina quanto à precisão.

Figura 3 – Fluxograma da metodologia.



Fonte: Acervo do autor

5.1 Seleção e preparação dos dados

Após selecionar o conjunto de dados *APS* para a comparação, definiremos as características que serão empregadas no treinamento do modelo. Neste trabalho, apresentamos um método para prever o desempenho futuro dos cientistas. As características do cientista selecionadas para prever seu desempenho são descritas na Tabela 1. Elas fornecem informações sobre a produtividade, a influência e a colaboração dos cientistas. O desempenho futuro é medido pelo número de artigos com três citações cada que o cientista terá nos três anos seguintes ao momento da predição (índice-I_3).

Pré-processamento

O pré-processamento desempenha um papel crucial na utilização de algoritmos de extração de dados. Além de melhorar a a precisão e confiabilidade dos resultados, minimiza o tempo necessário para a extração de conhecimentos mais profundos sobre as características dos dados a extrair. Com o propósito de implementar estas propriedades, foram realizadas as seguintes operações:

(a) Remoção dos registros incorretos das colunas referentes aos atributos `ano da primeira publicação do cientista na base de dados` e `ano da última publicação do`

Tabela 1 – Características do cientistas usadas para predição do seu desempenho futuro.

Atributo	Descrição
$y1$	Ano da primeira publicação do cientista na base de dados
$y2$	Ano da última publicação do cientista na base de dados
co	Número de coautores diferentes do cientista no período que vai desde a primeira publicação até o ano da predição
co_5	Número de coautores diferentes nos últimos cinco anos
h	Índice-h do cientista no ano da predição
h_5	Índice-h do cientista considerando somente as publicações dos últimos cinco anos
h_3	Índice-h do cientista considerando somente as publicações dos últimos três anos
c	Total de citações recebidas pelo cientista até o ano da predição
c_5	Total de citações recebidas pelo cientista considerando somente as publicações dos últimos cinco anos
c_3	Total de citações recebidas pelo cientista considerando somente as publicações dos últimos três anos
i_{10_5}	Total de artigos com 10 citações considerando somente as publicações dos últimos 5 anos
i_{5_5}	Total de artigos com 5 citações considerando somente as publicações dos últimos 5 anos
i_{3_5}	Total de artigos com 3 citações considerando somente as publicações dos últimos 5 anos
v_5	Número de veículos de publicação diferentes em que publicou nos últimos 5 anos
p_5	Total de artigos publicados nos últimos cinco anos
p_3	Total de artigos publicados nos últimos três anos

Fonte: Conjunto de dados *APS*

`cientista na base de dados`. Na primeira foram removidos valores inferiores a 1893, e na última, valores superiores a 2018.

(b) Criação de uma nova característica: `diferença entre o ano da última e da primeira publicação`.

(c) Criação de um novo conjunto de dados com as características constantes na Tabela 1 e o índice I3.

(d) Separação dos atributos de entrada e saída.

Normalização dos dados

Os dados foram normalizados usando a função `StandardScaller` da biblioteca *scikit-learn*. Essa função transforma os dados de modo que a média seja 0 e o desvio padrão seja 1. Isso é importante para garantir que os dados estejam na mesma escala e que os modelos de aprendizado de máquina sejam treinados de forma adequada.

Divisão dos dados de treino e teste

Os dados foram divididos em dois conjuntos, treino e teste, usando a abordagem *holdout* (KOHAVI, 1995). A divisão foi feita em uma proporção de 70% para treino e 30% para teste. Os dados de treino foram usados para treinar os modelos, e os dados de teste foram usados para avaliar o desempenho dos modelos.

A abordagem *holdout* é uma técnica simples e eficaz de divisão dos dados de treino e teste. Ela garante que os dados de teste sejam independentes dos dados de treino, o que é importante para obter uma avaliação precisa do desempenho dos modelos.

5.2 Treinamento dos modelos

O método de treinamento é baseado em uma abordagem supervisionada padrão. Essa abordagem consiste em fornecer ao modelo um conjunto de dados de treinamento composto por pares de entradas e saídas desejadas. O modelo aprende a mapear entradas para saídas desejadas observando esses pares.

$$\text{Função de perda} = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i)$$

Em que:

- y_i é a saída desejada para a entrada i.
- $\hat{y}_i$ é a saída estimada do modelo para a entrada i.
- $\ell(y_i, \hat{y}_i)$ é a função de perda para a entrada i.

A função de perda mede a diferença entre as saídas do modelo e as saídas desejadas. O otimizador é um algoritmo que atualiza os parâmetros do modelo de modo a minimizar a função de perda.

5.3 Avaliação dos resultados

O desempenho dos modelos foi avaliado usando as métricas RMSE, MAE e R^2, que são métricas comumente usadas em contextos de problemas envolvendo regressão.

RMSE (Root Mean Square Error)

O RMSE é uma medida da distância média entre as previsões do modelo e os valores reais. É calculado pela seguinte fórmula:

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

Em que:

- y_i é o valor real da amostra i
- $\hat{y}_i$ é a previsão do modelo para a amostra i
- n é o número de amostras

MAE (Mean Absolute Error)

O MAE é uma medida da diferença média entre as previsões do modelo e os valores reais. É calculado pela seguinte fórmula:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$$

Em que:

- y_i é o valor real da amostra i
- $\hat{y}_i$ é a previsão do modelo para a amostra i
- n é o número de amostras

R^2 (Coeficiente de determinação)

O R^2 é uma medida da variância explicada pelo modelo. É calculado pela seguinte fórmula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Em que:

- y_i é o valor real da amostra i
- $\hat{y}_i$ é a previsão do modelo para a amostra i
- $\bar{y}$ é a média dos valores reais

6 Resultados

A Tabela 2 fornece as métricas de avaliação dos três modelos de regressão: rede neural artificial, regressão simbólica e regressão linear múltipla. O repositório que contém os algoritmos utilizados no processo pode ser consultado no Apêndice A deste trabalho. Os valores das métricas foram calculados usando um conjunto de teste independente dos dados usados para treinar os modelos.

Tabela 2 – Resultados da avaliação do desempenho dos modelos.

Modelo	**RMSE**	**R²**	**MAE**
Rede neural artificial	1.22	0.48	0.62
Regressão simbólica	1.24	0.46	0.79
Regressão linear múltipla	1.28	0.45	0.79

Ao comparar o modelo de regressão simbólica não linear com o de regressão linear múltipla, usando a rede neural como parâmetro, obtemos o seguinte:

Na comparação do RMSE, a rede neural teve o menor valor (1.22), seguida pela regressão simbólica (1.24) e, por último, a regressão linear múltipla (1.28). Isso indica que a rede neural e a regressão simbólica não linear têm um desempenho melhor na previsão dos dados. Quanto ao R^2, a rede neural e a regressão simbólica obtiveram valores próximos , ambos são maiores que o modelo linear (0.45). Isso sugere que ambos os modelos explicam uma maior proporção da variância nos dados em comparação com a regressão linear múltipla. No que se refere ao MA, a rede neural teve o menor valor (0.62), seguido pela regressão simbólica e regressão linear múltipla (0.79). Isso indica que tanto a equação linear quanto a equação simbólica tem o mesmo desempenho na previsão dos dados. Em resumo, os resultados apontam que o modelo de regressão simbólica alcançou resultados comparáveis à rede neural, o mais preciso dos três modelos.

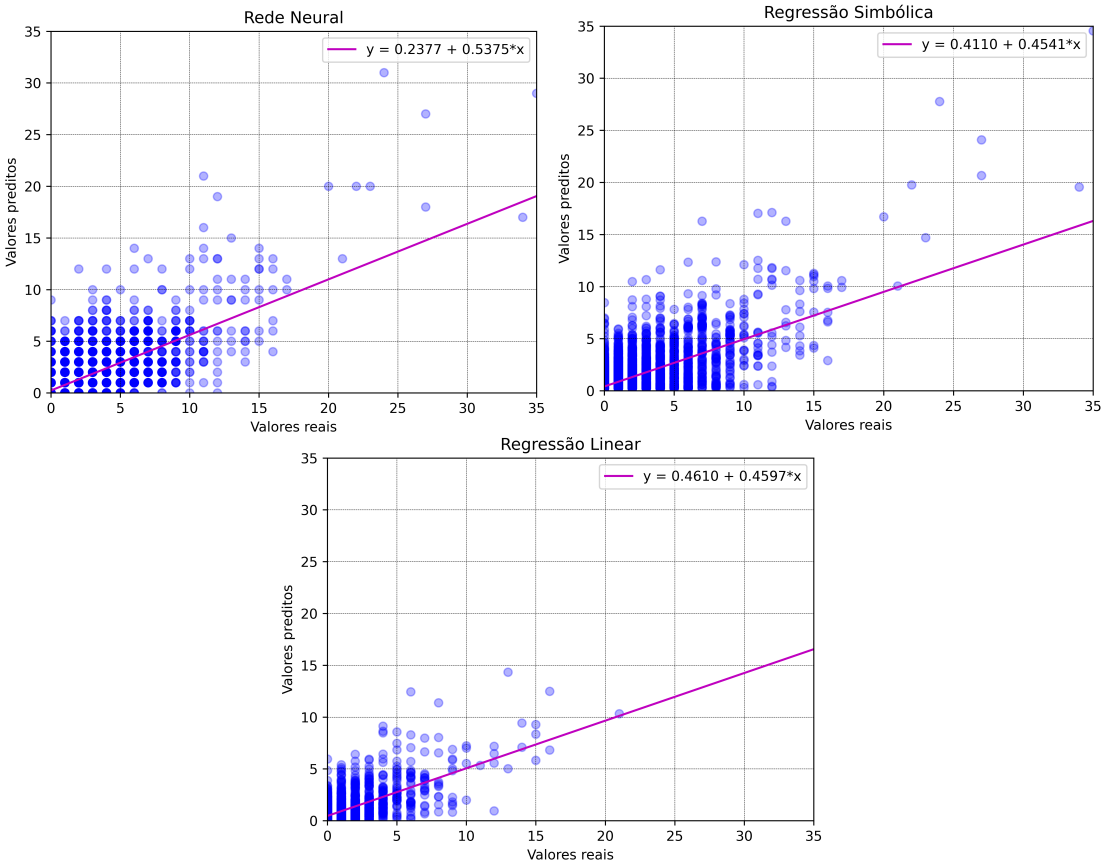
Em outra análise utilizou-se a regressão linear entre valores reais e preditos da variável dependente. Os resultados encontram-se na Tabela 3.

Tabela 3 – Coeficientes da equação resultante da regressão entre valores reais e preditos para três modelos de aprendizado de máquina

Coeficiente	**Rede Neural**	**Regressão simbólica**	**Regressão Linear**
a	0.54	0.45	0.46
b	0.24	0.41	0.46

Os coeficientes a e b representam, respectivamente, o coeficiente angular e o intercepto das reta de regressão simples $y = ax + b$ para os três modelos analisados: rede neural, regressão simbólica e regressão linear. Um valor de $a = 1$ e $b = 0$ indicaria um ajuste perfeito entre o valor real e o valor esperado. Se tomarmos essa informação como parâmetro para a análise, a rede neural obteve a melhor aproximação dos dados, seguida pela regressão simbólica. Comparando os modelos de regressão interpretáveis, a regressão simbólica tem desempenho melhor que a regressão linear. Os gráficos da Figura 4 ilustram as previsões dos modelos.

Figura 4 – Representação gráfica da regressão entre valores reais e preditos.



Fonte: Acervo do autor

7 Discussões e Conclusão

Neste estudo, desenvolveu-se um modelo não linear para prever o sucesso futuro de curto prazo para físicos. Para construir o modelo, utilizou a regressão simbólica para inferir uma equação a partir das relações entre as variáveis no conjunto de dados APS, mantido pela American Physical Society. A acurácia do modelo foi medida usando as métricas MAE, RMSE e R^2. Os resultados mostram que o modelo de regressão simbólica não linear supera o modelo linear, demonstrando ser capaz de capturar as relações não lineares entre as variáveis que influenciam o sucesso de curto prazo de físicos. No entanto, é importante notar que esses resultados são baseados em um conjunto de dados específico que inclui , na maioria dos registros, pesquisadores dos Estados Unidos e podem não refletir a realidade dos pesquisadores do Brasil.

A aplicação da regressão simbólica ao desenvolvimento de um modelo não linear do sucesso futuro de curto prazo de cientistas da física é um avanço significativo na pesquisa sobre a previsão do sucesso científico. Isso ocorre porque é a primeira vez que essa metodologia é usada para estudar esta temática nesse contexto. Os achados deste estudo são, portanto, inéditos e fornecem novas informações sobre os fatores que contribuem para o sucesso científico de curto prazo de cientistas da física.

É importante ressaltar que o uso da equação gerada no modelo deve ser feito com cautela. Os dados utilizados para gerar a equação podem conter vieses, o que pode gerar resultados tendenciosos. Portanto, o modelo deve ser considerado como um instrumento de apoio, mas não como o único elemento determinante para a tomada de decisões.

Em trabalhos futuros, planeja-se o algoritmo *Deep Symbolic Regression* (PETERSEN et al., 2019), que mescla a interpretabilidade da regressão simbólica com a precisão do aprendizado profundo de redes neurais. Propõe-se ainda, fazer uso outras bases de dados que contenham, além de indicadores bibliométricos (empregados no conjunto de dados deste estudo), indicadores alternativos, tais como: fator de impacto social, índice de menção e índice de visualização. Por fim, pretende-se ampliar o estudo de interpretabilidade de algoritmos de regressão simbólica, combinando métodos de explicação locais e globais.

Referências

ACEBES, F.; POZA, D.; GONZáLEZ-VARONA, J. M.; LóPEZ-PAREDES, A. Stochastic earned duration analysis for project schedule management. *Engineering*, v. 9, p. 148–161, 2022. ISSN 2095-8099. Citado na página 14.

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

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. Citado na página 18.

CYBENKO, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, Springer, v. 2, n. 4, p. 303–314, 1989. Citado na página 16.

DUZAN, H.; SHARIFF, N. S. B. M. Ridge regression for solving the multicollinearity problem: review of methods and models. *Journal of Applied Science*, ANSI net, 2015. Citado na página 15.

FAHRMEIR, L.; KNEIB, T.; LANG, S.; MARX, B. *Regression: Models, Methods and Applications*. [S.l.]: Springer Berlin Heidelberg, 2013. Citado na página 14.

FRANCA, F. de; VIRGOLIN, M.; KOMMENDA, M.; MAJUMDER, M.; CRANMER, M.; ESPADA, G.; INGELSE, L.; FONSECA, A.; LANDAJUELA, M.; PETERSEN, B. et al. Interpretable symbolic regression for data science: Analysis of the 2022 competition. *arXiv preprint arXiv:2304.01117*, 2023. Citado na página 16.

GAFAROV, F. M.; NIKOLAEV, K. S.; USTIN, P. N.; BERDNIKOV, A. A.; ZAKHAROVA, V. L.; REZNICHENKO, S. A. A complex neural network model for predicting a personal success based on their activity in social networks. *Eurasia Journal of Mathematics, Science and Technology Education*, ERIC, v. 17, n. 10, 2021. Citado na página 18.

HOU, L.; WU, Q.; XIE, Y. Does early publishing in top journals really predict long-term scientific success in the business field? *Scientometrics*, v. 127, n. 11, p. 6083–6107, Nov 2022. Citado na página 12.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 1137–1143. ISBN 1558603638. Citado na página 25.

KOLMOGOROV, A. N. On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. *Transl., Ser. 2, Am. Math. Soc.*, v. 17, p. 369–373, 1956. Citado na página 16.

KOZA, J. R. Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, v. 4, n. 2, p. 87–112, Jun 1994. Citado na página 16.

KUMAR, D.; BHOWMICK, P. K.; PAIK, J. H. Researcher influence prediction (resip) using academic genealogy network. *Journal of Informetrics*, v. 17, n. 2, p. 101392, 2023. Citado na página 12.

KUPPLER, M. Predicting the future impact of computer science researchers: Is there a gender bias? *Scientometrics*, v. 127, n. 11, p. 6695–6732, Nov 2022. Citado na página 12.

LAURANCE, W. F.; USECHE, D. C.; LAURANCE, S. G.; BRADSHAW, C. J. A. Predicting Publication Success for Biologists. *BioScience*, v. 63, n. 10, p. 817–823, 10 2013. Citado na página 12.

MAULUD, D.; ABDULAZEEZ, A. M. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, v. 1, n. 4, p. 140–147, 2020. Citado na página 14.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, p. 115–133, 1943. Citado na página 15.

PETERSEN, B. K.; LANDAJUELA, M.; MUNDHENK, T. N.; SANTIAGO, C. P.; KIM, S. K.; KIM, J. T. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. *arXiv preprint arXiv:1912.04871*, 2019. Citado na página 29.

SARIGÖL, E.; PFITZNER, R.; SCHOLTES, I.; GARAS, A.; SCHWEITZER, F. Predicting scientific success based on coauthorship networks. *EPJ Data Science*, Springer, v. 3, p. 1–16, 2014. Citado na página 18.

SCHMIDT, M.; LIPSON, H. Distilling free-form natural laws from experimental data. *Science*, v. 324, n. 5923, p. 81–85, 2009. Citado na página 16.

SHRESTHA, N. Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, v. 8, n. 2, p. 39–42, 2020. Citado na página 15.

UYANIK, G. K.; GÜLER, N. A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, Elsevier, v. 106, p. 234–240, 2013. Citado na página 14.

XIA, W.; LI, T.; LI, C. A review of scientific impact prediction: tasks, features and methods. *Scientometrics*, v. 128, n. 1, p. 543–585, Jan 2023. Citado na página 12.

Apêndice - Repositório contendo os algoritmos utilizados na geração dos modelos

<https://github.com/jvinicius7/sucesso-futuro>