XI SINGEP
Simpósio Internacional de Gestão, Projetos, Inovação e Sustentabilidade
International Symposium on Management, Project, Innovation and Sustainability
ISSN: 2317-8302
CYRUS Institute of Knowledge MAKE A DIFFERENCE
CIK 11th INTERNATIONAL CONFERENCE

ANÁLISE PREDITIVA DE DADOS: UMA ANÁLISE BIBLIOMÉTRICA DE ARTIGOS PUBLICADOS ENTRE 2018 E 2023

PREDICTIVE DATA ANALYSIS: A BIBLIOMETRIC ANALYSIS OF ARTICLES PUBLISHED BETWEEN 2018 AND 2023

MARCELO NEVES GONÇALVES
UNIVERSIDADE PRESBITERIANA MACKENZIE

LEONARDO FERNANDO CRUZ BASSO
UNIVERSIDADE PRESBITERIANA MACKENZIE

ANÁLISE PREDITIVA DE DADOS: UMA ANÁLISE BIBLIOMÉTRICA DE ARTIGOS PUBLICADOS ENTRE 2018 E 2023

Objetivo do estudo
O estudo tem como objetivos realizar uma análise bibliométrica abrangente, identificar principais tendências e áreas de foco, avaliar desenvolvimentos metodológicos, e identificar lacunas e oportunidades de pesquisa.

Relevância/originalidade
Essa análise bibliométrica, espera fornecer uma visão panorâmica das tendências e desenvolvimentos na análise preditiva de dados ao longo do período e a compreensão desses padrões e áreas de foco contribuirá para identificar novas oportunidades de pesquisa

Metodologia/abordagem
foi realizada uma análise bibliométrica no período de 2018 a abril de 2023, utilizando uma amostra final de 491 artigos em 278 periódicos acadêmicos, que serviram como base para uma análise quantitativa desenvolvida por meio de contagens de frequência e co-citações.

Principais resultados
A academia está focando suas pesquisas em algoritmos de aprendizagem de máquina, mais adequados quando há complexidade e não linearidade nos dados, permitindo a criação de modelos mais sofisticados, capazes de se ajustar automaticamente aos dados, mas menos transparentes no processamento.

Contribuições teóricas/metodológicas
Embora os algoritmos de aprendizagem de máquina têham a vantagem de serem capazes de se ajustar automaticamente aos dados, sem a necessidade de especificar explicitamente uma equação de regressão ou modelo, seu processamento não é transparente para alcançar os resultados

Contribuições sociais/para a gestão
A formação de pessoas que atuam com ciência de dados deverá adaptar-se com maiores exigencias no trato com dados para treinamento e ajuste adequado dos modelos que podem ser mais suscetíveis a overfitting, quando não generaliza bem para novos dados.

Palavras-chave: Aprendizagem de máquina, inteligência artificial, Redes Neurais, Algoritmos, Sobreajuste

PREDICTIVE DATA ANALYSIS: A BIBLIOMETRIC ANALYSIS OF ARTICLES PUBLISHED BETWEEN 2018 AND 2023

Study purpose
The study aims to perform a comprehensive bibliometric analysis, identify key trends and focus areas, assess methodological developments, and identify research gaps and opportunities.

Relevance / originality
This bibliometric analysis is expected to provide a bird's-eye view of trends and developments in predictive data analytics over time, and understanding these patterns and areas of focus will help identify new research opportunities.

Methodology / approach
A bibliometric analysis was carried out from 2018 to April 2023, using a final sample of 491 articles in 278 academic journals, which served as the basis for a quantitative analysis developed through frequency counts and co-citations.

Main results
The academy is focusing its research on machine learning algorithms, more suitable when there is complexity and non-linearity in the data, allowing the creation of more sophisticated models, capable of automatically adjusting to the data, but less transparent in the processing.

Theoretical / methodological contributions
Although machine learning algorithms have the advantage of being able to automatically fit the data, without the need to explicitly specify a regression equation or model, their processing is not transparent in achieving the results.

Social / management contributions
The training of people who work with data science must adapt to greater demands in dealing with data for training and proper adjustment of models that may be more susceptible to overfitting, when it does not generalize well to new data.

Keywords: Machine learning, Artificial Intelligence, Neural networks, Algorithms, Overfitting

ANÁLISE PREDITIVA DE DADOS: UMA ANÁLISE BIBLIOMÉTRICA DE ARTIGOS PUBLICADOS ENTRE 2018 E 2023

1 Introdução

Nos últimos anos, a análise preditiva de dados emergiu como um campo de estudo e aplicação de extrema relevância, alimentado pelo aumento exponencial na disponibilidade e na diversidade dos dados. Esse cenário tem possibilitado uma visão mais aprofundada e perspicaz das tendências, comportamentos e padrões ocultos subjacentes aos conjuntos de dados, permitindo a previsão de eventos futuros com base em análises estatísticas e algoritmos de aprendizado de máquina. O presente estudo busca explorar a evolução e os progressos alcançados no campo da análise preditiva de dados ao longo do período de cinco anos compreendidos entre 2018 e 2023, por meio de uma abordagem bibliométrica. A análise bibliométrica é uma ferramenta que permite mapear a produção científica, identificar padrões de colaboração entre pesquisadores e instituições, e compreender as tendências que moldam um determinado campo de pesquisa. A seleção desse intervalo de tempo não é mero acaso, mas sim um reflexo da rápida transformação tecnológica e da crescente importância da análise preditiva de dados nesse período. O surgimento de novas técnicas, aprimoramento de algoritmos e o desenvolvimento de ferramentas mais poderosas criaram um ambiente propício para uma análise aprofundada da produção acadêmica que permeia esse domínio. Ao traçar uma amostra bibliométrica de artigos publicados, este estudo visa identificar as principais áreas de concentração e exploração dentro da análise preditiva de dados, que podem incluir, por exemplo, a aplicação de técnicas preditivas em setores específicos, como saúde, finanças, marketing e logística, ou a investigação de algoritmos inovadores para aprimorar a precisão das previsões. Ao mesmo tempo, pretende-se mapear as conexões entre diferentes pesquisadores e instituições, destacando as colaborações mais significativas que contribuíram para o desenvolvimento do campo. Além disso, a análise aprofundada das palavras-chave e dos termos mais frequentes nos artigos fornecerá insights sobre as tendências emergentes, bem como sobre o uso continuado de conceitos e metodologias estabelecidos. Isso permitirá uma visão panorâmica das transformações conceituais que ocorreram na análise preditiva de dados ao longo do período em estudo. Em última análise, este estudo bibliométrico pretende não apenas fornecer um instantâneo da produção acadêmica no campo da análise preditiva de dados, mas também contribuir para a compreensão geral das trajetórias de pesquisa, das áreas de foco predominantes e das direções futuras que esse campo está tomando. À medida que a análise preditiva de dados continua a moldar indústrias e a orientar decisões estratégicas, essa análise bibliométrica poderá servir como um guia para pesquisadores, profissionais e tomadores de decisão que desejem explorar e entender as complexidades desse domínio.

A análise preditiva se concentra em extrair padrões, tendências e relações ocultas a partir de conjuntos de dados, permitindo prever eventos futuros com um grau razoável de precisão. Assim, a pesquisa dessa temática torna-se relevante ao realizar uma análise bibliométrica e para entender as principais tendências, áreas de foco e contribuições ao longo do período estabelecido. Nesse contexto, este estudo aborda o seguinte problema de pesquisa: Quais são as principais tendências, áreas de foco e desenvolvimentos na análise preditiva de dados, conforme evidenciado pela produção acadêmica em artigos publicados em periódicos entre os anos de 2018 e 2023?

O estudo tem como objetivos realizar uma análise bibliométrica abrangente, buscando identificar e coletar uma ampla gama de artigos acadêmicos relacionados à análise preditiva de dados publicados em periódicos de impacto relevantes durante o período de 2018 a 2023, e extrair informações bibliométricas, como número de publicações, autores mais produtivos, afiliações institucionais e redes de colaboração, para fornecer uma visão geral da comunidade de pesquisa nesse campo; identificar principais tendências e áreas de foco, buscando analisar

os tópicos e palavras-chave mais frequentes nos artigos para identificar as principais tendências e áreas de foco na análise preditiva de dados, e avaliar a evolução dessas tendências ao longo do período de estudo, destacando quais conceitos e métodos ganharam destaque e quais podem ter perdido relevância; avaliar desenvolvimentos metodológicos, buscando investigar as abordagens metodológicas mais utilizadas na análise preditiva de dados, incluindo técnicas de aprendizado de máquina, mineração de dados, modelagem estatística e outras, e analisar como essas metodologias têm evoluído e quais novas abordagens têm surgido no campo ao longo dos anos; e identificar lacunas e oportunidades de pesquisa, buscando identificar áreas de pesquisa menos exploradas ou emergentes na análise preditiva de dados, que possam representar oportunidades para futuras investigações, e discutir as possíveis implicações práticas e teóricas dos resultados obtidos, indicando direções potenciais para o desenvolvimento futuro da pesquisa nesse campo.

Por meio dessa análise bibliométrica, espera-se fornecer uma visão panorâmica das tendências e desenvolvimentos na análise preditiva de dados ao longo do período de 2018 a 2023. A compreensão desses padrões e áreas de foco contribuirá para uma apreciação mais profunda do estado atual da pesquisa nesse campo, e para a identificação de oportunidades para avanços futuros nas pesquisas envolvendo a análise preditiva de dados.

2 Referencial Teórico

Análises preditivas com uso de inteligência artificial (IA) estão sendo desenvolvidas em uma ampla gama de campos e indústrias. Algumas das principais áreas em que os estudos de análises preditivas com uso de IA estão sendo desenvolvidos incluem Finanças, em que a IA pode ser usada para prever preços de ações, identificar fraudes e melhorar as decisões de investimento; Saúde, onde a IA pode ser usada para prever a probabilidade de desenvolver certas doenças, melhorar diagnósticos e personalizar tratamentos; Transporte, em que a IA pode ser usada para prever padrões de tráfego, otimizar logística e remessa e melhorar a segurança de veículos autônomos; Marketing, em que a IA pode ser usada para prever o comportamento do consumidor, personalizar a publicidade e otimizar as estratégias de preços; Energia, onde a IA pode usada para prever a demanda de energia, otimizar o consumo de energia e melhorar a eficiência das fontes de energia renováveis; Manufatura, em que a IA pode ser usada para prever falhas de equipamentos, otimizar cadeias de suprimentos e melhorar o controle de qualidade; Agricultura, onde a IA pode ser usada para prever o rendimento das colheitas, otimizar a irrigação e melhorar o manejo de pragas; Educação, em que a IA pode ser usada para prever o desempenho do aluno, personalizar o aprendizado e melhorar os resultados educacionais; e para o Governo, onde a IA pode ser usada para prever padrões de crime, melhorar a segurança pública e aumentar a segurança nacional, por exemplo. No geral, os estudos de análises preditivas com uso de IA estão sendo desenvolvidos em vários campos e indústrias, com o objetivo de melhorar a tomada de decisões, a eficiência e a precisão (Borz et al., 2022; Gondia et al., 2023; Kamepalli & Rao, 2019; Kumar et al., 2020; Patel & Parikh, 2020).

3 Metodologia

Com o intuito de alcançar os objetivos desse estudo, foi realizada uma análise bibliométrica no período de 2018 a abril de 2023, utilizando uma amostra final de 491 artigos em 278 periódicos acadêmicos, que serviram como base para uma análise quantitativa desenvolvida por meio de contagens de frequência e co-citações. Os artigos foram obtidos e analisados com base nas bases de dados SCOPUS e Web of Science (WoS), e com o auxílio de softwares como R e Biblioshiny, e a verificação de leis de análise bibliométrica-chave, como Zipf (Zipf, 1949), (Bradford, 1985) e Lotka (Lotka, 1926).

4 Análise dos resultados

Na análise preditiva de dados, a escolha entre métodos tradicionais, como a análise de regressão linear/logística, e algoritmos de aprendizagem de máquina depende da natureza dos

dados, do problema em questão e da complexidade das relações que se deseja modelar. Ambas as abordagens têm suas próprias vantagens e limitações, e a seleção da técnica adequada deve ser baseada na compreensão das características do problema e dos objetivos da análise.

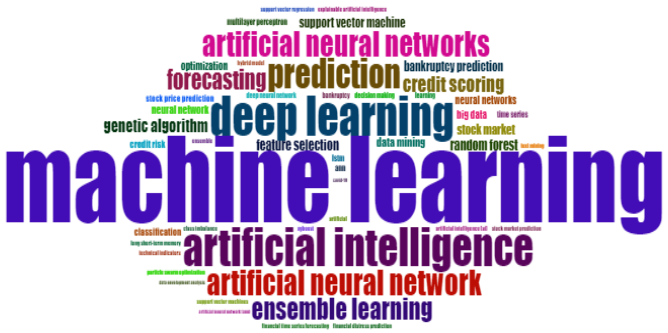

Figura 1Mapa de Palavras-chave indicadas nos artigos da amostra do estudo
Fonte: dados gerados pelo software Biblioshiny a partir da amostra de dados do estudo.

A análise das palavras-chave utilizadas pelos artigos que fazem parte da amostra pesquisada demonstra que o principal interesse da academia refere-se a modelagem de relações mais complexas, com realização de estudos que envolvem algoritmos de aprendizagem de máquina, *deep learning*, ou redes neurais, como pode ser observado no mapa de palavras da Figura 1, acima, e na figura 2 abaixo, em que fica evidente a tendencia de lidar com conjuntos de dados complexos e realizando tarefas mais sofisticadas, como classificação e regressão não linear, nas pesquisas.

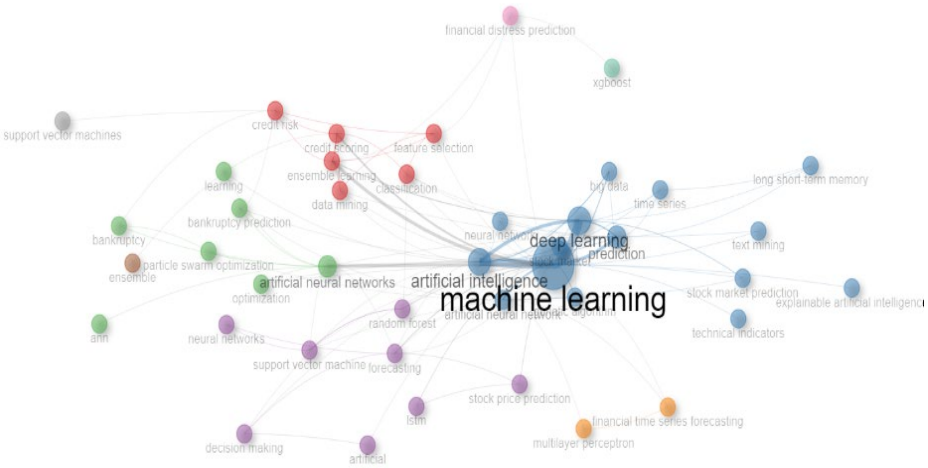

Figura 2 Redes de Palavras-chave dos estudos da amostra
Fonte: dados gerados pelo software Biblioshiny a partir da amostra de dados do estudo.

5 Conclusões/Considerações finais

As análises por meio de regressão linear ou logística e os algoritmos de aprendizagem de máquina representam abordagens distintas para a interpretação e previsão de padrões em conjuntos de dados. Enquanto ambas as técnicas buscam extrair informações valiosas, elas

diferem em termos de metodologia e aplicabilidade. A regressão linear e logística são técnicas estatísticas tradicionais frequentemente usadas para entender a relação entre variáveis dependentes e independentes. A regressão linear é aplicada quando a variável dependente é contínua e o objetivo é estabelecer uma relação linear entre as variáveis. Por outro lado, a regressão logística é empregada quando a variável dependente é categórica e assume dois valores, sendo útil para modelar a probabilidade de uma determinada categoria ocorrer.

Já os algoritmos de aprendizagem de máquina, como aqueles utilizados em técnicas de aprendizado supervisionado (por exemplo, árvores de decisão, redes neurais, k-vizinhos mais próximos), são capazes de lidar com conjuntos de dados complexos e realizar tarefas mais sofisticadas, como classificação e regressão não linear. Ao contrário das abordagens estatísticas tradicionais, os algoritmos de aprendizagem de máquina têm a capacidade de capturar padrões não lineares e de alta dimensionalidade nos dados, adaptando-se melhor a problemas com múltiplas variáveis de entrada e saída (Gondia et al., 2023; Kamepalli & Rao, 2019; Kumar et al., 2020).

Uma das principais diferenças entre essas abordagens é a flexibilidade. Enquanto a regressão linear ou logística pode ser apropriada para cenários onde a relação entre variáveis é bem compreendida e linear, os algoritmos de aprendizagem de máquina são mais adequados quando há complexidade e não linearidade nos dados, permitindo a criação de modelos mais sofisticados (Kamepalli & Rao, 2019; Patel & Parikh, 2020). Além disso, os algoritmos de aprendizagem de máquina têm a vantagem de serem capazes de se ajustar automaticamente aos dados, sem a necessidade de especificar explicitamente uma equação de regressão ou modelo. No entanto, eles também podem exigir mais dados para treinamento e ajuste adequado, e podem ser mais suscetíveis a overfitting (quando o modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados).

6 Referencias

Borz, S. A., Forkuo, G. O., Oprea-Sorescu, O., & Proto, A. R. (2022). Development of a Robust Machine Learning Model to Monitor the Operational Performance of Fixed-Post Multi-Blade Vertical Sawing Machines. *Forests*, *13*(7). https://doi.org/10.3390/f13071115

Bradford, S. C. (1985). Sources of information on specific subjects 1934. *Journal of Information Science*, *10*(4), 176–180. https://doi.org/10.1177/016555158501000407

Gondia, A., Moussa, A., Ezzeldin, M., & El-Dakhakhni, W. (2023). Machine learning-based construction site dynamic risk models. *Technological Forecasting and Social Change*, *189*. https://doi.org/10.1016/j.techfore.2023.122347

Kamepalli, S., & Rao, B. S. (2019). Recent applications of machine learning: A survey. *International Journal of Innovative Technology and Exploring Engineering*, *8*(6 C2), 263–267.

Kumar, G., Thakur, K., & Ayyagari, M. R. (2020). MLEsIDSs: machine learning-based ensembles for intrusion detection systems—a review. *Journal of Supercomputing*, *76*(11), 8938–8971. https://doi.org/10.1007/s11227-020-03196-z

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–323. http://www.jstor.org/stable/24529203

Patel, H., & Parikh, A. (2020). Predicting possible fraud in India using machine learning: An empirical comparison between model for better prediction. *International Journal of Advanced Science and Technology*, *29*(5), 5204–5217.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human eoclogy* (Addison-Wesley Press., Ed.; 1st ed., Vol. 1). Addison-Wesley Press.