International Journal of Innovation

e-ISSN: 2318-9975
https://doi.org/10.5585/2023.24056
Recebido: 13 mar. 2023 - **Aprovado:** 18 jun. 2023

Processo de Avaliação: Double Blind Review
Editora Chefe: Priscila Rezende da Costa
Coeditora: Isabel Cristina Scafuto
Editora Científica: Vânia Maria Jorge Nassif
Editora Assistente: Angelica Pigola
***Seção**: Artigo*

Check for updates

MACHINE LEARNING:UMA ANÁLISE BIBLIOMÉTRICA

Emerson Martins [1]
Napoleão Verardi Galegale [2]

Cite as – American Psychological Association (APA)

Martins, E., & Galegale, N. V. (2023, set./dez.). Machine learning: uma análise bibliométrica. *International Journal of Innovation - IJI*, São Paulo, *11*(3), 1-40, e24056. https://doi.org/10.5585/2023.24056

Resumo

Objetivo: Apresentar uma visão dos artigos científicos publicados nos últimos dez anos sobre o tema aprendizado de máquina, do inglês *machine learning* (ML), com ênfase nos algoritmos preditivos.
Método/abordagem: Análise bibliométrica, com apoio do protocolo PRISMA, para avaliar autores, universidades e países, quanto a produtividade, citações bibliográficas e focos sobre o tema, com amostra de 773 artigos das bases de dados Scopus e *Web of Science*, no período de 2013 a maio/2023.
Originalidade/valor: Há ausência de estudos na literatura que consolidem artigos relacionados a ML e *Big Data*. A pesquisa contribui para cobrir tal lacuna, favorecendo o delineamento de ações e pesquisas futuras.
Principais resultados: Foram identificados no corpus bibliométrico de ML: autores mais citados e com maior número de publicações, países e universidades mais produtivas, periódicos com maior número de publicações e citações, áreas de conhecimento com maior número de publicações e artigos de maior prestígio. Nos temas e domínios de ML, foram identificados: principais coocorrências de palavras-chaves, temas emergentes (agrupados em cinco *clusters*) e nuvem de palavras por título e por resumo. Os estudos sobre impacto da aquisição de dados e análise preditiva representam oportunidades para pesquisas futuras.
Contribuições teóricas/metodológicas: O protocolo PRISMA possibilitou a identificação e análises quantitativa e qualitativa relevantes dos artigos, consolidando o conhecimento científico sobre o tema.
Contribuições sociais/gerenciais: Facilidade de compreender a maturidade das pesquisas sobre ML e *Big Data* por parte de gestores de empresas e pesquisadores, quanto à viabilidade de investimentos para se obter vantagens competitivas com tais tecnologias.

Palavras-chave: aprendizado de máquina, análise de *big data*, análise bibliométrica, predição

MACHINE LEARNING: A BIBLIOMETRIC ANALYSIS

Abstract

Objective: Present an overview of scientific articles published in the last ten years on the topic of

[1] Mestre em Gestão e Tecnologia em Sistemas Produtivos (CEETEPS) e Pesquisador do Grupo de Pesquisa em Gestão Estratégica da TI (CEETEPS/CNPq). CEETEPS – Centro Estadual de Educação Tecnológica Paula Souza / São Paulo (SP) – Brasil. emerson.martins@cpspos.sp.gov.br
[2] Doutor em Controladoria e Contabilidade (FEA/USP), Mestre em Engenharia da Produção (POLI/USP), Professor e Pesquisador da UPEP/CEETEPS e da FEA/PUC-SP, líder do Grupo de Pesquisa em Gestão Estratégica da TI (CEETEPS/CNPq) e Consultor de Empresas.CEETEPS – Centro Estadual de Educação Tecnológica Paula Souza / São Paulo (SP) – Brasil. napoleao.galegale@cpspos.sp.gov.br

(cc) BY-NC-SA

machine learning (ML), with an emphasis on predictive algorithms.
Method/approach: Bibliometric analysis, with support from the PRISMA protocol, to evaluate authors, universities and countries, regarding productivity, bibliographic citations and focuses on the topic, with a sample of 773 articles from the Scopus and Web of Science databases, from 2013 to May /2023.
Originality/value: There is an absence of studies in the literature that consolidate articles related to ML and Big Data. The research contributes to covering this gap, favoring the design of future actions and research.
Main results: The following were identified in the ML bibliometric corpus: most cited authors with the greatest number of publications, most productive countries and universities, journals with the greatest number of publications and citations, areas of knowledge with the greatest number of publications, and the most prestigious articles. In the ML themes and domains, the following were identified: main co-occurrences of keywords, emerging themes (grouped into five clusters), and word clouds by title and abstract. Studies on the impact of data acquisition and predictive analysis represent opportunities for future research.
Theoretical/methodological contributions: The PRISMA protocol enabled the identification and relevant quantitative and qualitative analyzes of articles, consolidating scientific knowledge on the topic.
Social/managerial contributions: Ease of understanding the maturity of research on ML and Big Data by company managers and researchers, regarding the feasibility of investments to obtain competitive advantages with such technologies.

Keywords: machine learning, Big Data analysis, bibliometric analysis, prediction.

APRENDIZAJE AUTOMÁTICO: UN ANÁLISIS BIBLIOMÉTRICO

Resumen

Objetivo: Presentar un panorama de artículos científicos publicados en los últimos diez años sobre el tema de aprendizaje automático (ML en Inglés), con énfasis en algoritmos predictivos.
Método/enfoque*:* Análisis bibliométrico, con apoyo del protocolo PRISMA, para evaluar autores, universidades y países, en cuanto a productividad, citaciones bibliográficas y enfoques en el tema, con una muestra de 773 artículos de las bases de datos Scopus y Web of Science, del 2013 a mayo/2023.
Originalidad/valor: Existe una ausencia de estudios en la literatura que consoliden artículos relacionados con ML y Big Data. La investigación contribuye a cubrir este vacío, favoreciendo el diseño de futuras acciones e investigaciones.
Principales resultados: En el corpus bibliométrico de ML se identificaron: autores más citados con mayor número de publicaciones, países y universidades más productivos, revistas con mayor número de publicaciones y citaciones, áreas de conocimiento con mayor número de publicaciones y las más prestigiosas. artículos. En los temas y dominios de ML, se identificaron lo siguiente: principales co-ocurrencias de palabras clave, temas emergentes (agrupados en cinco grupos) y nubes de palabras por título y resumen. Los estudios sobre el impacto de la adquisición de datos y el análisis predictivo representan oportunidades para futuras investigaciones.
Contribuciones teóricas/metodológicas: El protocolo PRISMA permitió la identificación y análisis cuantitativos y cualitativos relevantes de artículos, consolidando el conocimiento científico sobre el tema.
Contribuciones sociales/gerenciales: Facilidad de comprensión de la madurez de la investigación sobre ML y Big Data por parte de directivos e investigadores de empresas, en cuanto a la viabilidad de inversiones para obtener ventajas competitivas con dichas tecnologías.

Palabras clave: machine learning, análisis de Big Data, análisis bibliométrico, predicción.

(cc) BY-NC-SA

1 INTRODUÇÃO

Nos últimos anos ocorreram vários avanços tecnológicos, como o surgimento de conceitos de *big data*, além dos benefícios acumulados com a ciência de dados para a sociedade (CHEN *et al*, 2014). Assim como o capital humano e as máquinas, os dados surgiram como um recurso essencial para gerar prosperidade na sociedade. Embora o processamento de dados tenha iniciado com métodos tradicionais de extrair, transformar e tratar os dados por meio de sistemas de gestão empresarial, de acordo com Hu *et al* (2014), essas técnicas não são escaláveis, especialmente dado ao enorme aumento no volume de dados. *Big data*, portanto, evolui, à medida que as empresas percebem que, para obter vantagem competitiva o investimento na análise de dados é igualmente importante junto com os produtos, serviços, processos e tecnologia (MISHRA *et al*, 2018).

Esta necessidade de evolução ocorre também devido a ocorrência de dados não estruturados, os quais não podem ser processados diretamente com as ferramentas tradicionais, ou seja, precisam de técnicas especiais de tratamento de dados e processamento de informações, como *Natural Language Processing* (NLP) e o *Machine Learning* (ML). Atualmente o processamento das informações para geração de conhecimento tornou-se vital para os tomadores de decisão, particularmente em algumas áreas importantes como a previsão de vendas de produtos ou serviços no varejo, nas quais variáveis externas como o tempo ou economia global podem afetar a decisão de consumo das pessoas (KRAWCZYK, 2016). Além disso, esta revolução demandou a integração em nuvem, da Internet das Coisas (IoT), *Blockchain* e do *Big Data Analysis* (BDA) (GILL *et al*, 2019).

O *Big Data* e o *ML* tem sido amplamente utilizados pelas organizações devido as necessidades crescentes de negócio e serviços para enfrentar os desafios globais na obtenção de vantagem competitiva. Este novo modelo multiplicou a demanda por ferramentas analíticas para resolver problemas de negócios complexos em vários domínios, incluindo mercado financeiro, marketing, saúde, cadeia de suprimentos e a predição de vendas. Diante deste cenário surge o *Business Analytics*, que é a aplicação de técnicas utilizando ferramentas de análise de *Big Data* conhecidas como Ciência de Dados para tomada de decisões (CHEN *et al*, 2014).

A pesquisa no domínio do ML teve um salto significativo nos últimos anos (ATHMAJA *et al*, 2017). Consequentemente, vários estudos realizaram pesquisas bibliométricas para resumir o conhecimento existente no campo do ML. Por exemplo,

(cc) BY-NC-SA

(ANTONOPOULOS *et al*, 2020) realizaram uma revisão das perspectivas no setor de energia renovável. Da mesma forma (SHARMA *et al*, 2020) revisaram a aplicação de análises do ML no contexto agrícola. Apesar dessas importantes tentativas de sintetizar a literatura existente, é observado que a literatura sobre o surgimento das tecnologias mais recentes, como inteligência artificial (IA), ML e *Big Data*, parecem fragmentadas (CHANDRA e VERMA, 2021). Os diferentes aspectos do ML e seu escopo para pesquisas futuras não foi considerado. Há uma necessidade evidente de pesquisas para fornecer uma compreensão abrangente do passado, presente e futuro de pesquisa a respeito da utilização do ML. Portanto, esta pesquisa considera esta lacuna em estudos bibliométricos e estende o levantamento bibliométrico do impacto da utilização do ML nas organizações. Este estudo considera três questões de pesquisa para abordar as lacunas de pesquisas mencionadas anteriormente: (1) qual é o foco da presente pesquisa sobre ML? (2) quais são os principais temas e domínios em ML e sua evolução? e (3) qual o escopo para pesquisas futuras, seja do ponto de vista acadêmico ou do mercado?

Este artigo fornece uma visão geral bibliográfica em consonância com Batistic e Van (2019), assim como Sahoo (2021). Da mesma forma, o artigo também é uma generalização de estudos bibliométricos contidos na literatura, como a análise da cadeia de suprimentos realizada por Mishra *et al*. (2018), *Smart Cities* realizado por Kousis e Tjortjis (2021) e a análise bibliométrica das cadeias de suprimentos sustentáveis realizado por (BUI et al, 2021).

2 MÉTODO

Este artigo apresenta uma análise aprofundada da citação e publicação de tendências na análise de ML entre 2013 e maio/2023. Este período foi escolhido com base na disponibilidade de acesso aos dados das bases de dados *Web of Science* e *Scopus*. Os autores, instituições, países e periódicos significativos são apresentados. Os principais temas discutidos são destacados, e os artigos são classificados em cinco grupos bibliográficos com base nas palavras-chaves que ocorrem com frequência. Esta abordagem ilustra os principais temas presentes nos artigos examinados, assim como a relação dos autores com as palavras-chaves. Os tópicos que ocorrem com frequência são indicados por meio de análise de nuvem de palavras e a análise da estrutura das citações são realizadas por grupo, para destacar os temas emergentes.

A análise bibliométrica é usada para destacar os principais autores, instituições/universidades e países em termos de suas contribuições no respectivo campo. Os padrões de colaboração entre autores, instituições e os países também são analisados (BATISTIC E VAN, 2019).

Adicionalmente foi utilizado o protocolo PRISMA-P, cujo objetivo é apoiar os pesquisadores a melhorarem o relato de revisões sistemáticas e meta-análises, filtrando o número de publicações com maior relevância ao tema pesquisado (MOHER *et al*., 2015).

Conforme mencionado na introdução deste artigo, foi observado a ausência de estudos na literatura que consolidem artigos relacionados a ML e *Big Data* com ênfase em algoritmos preditivos, desta forma, esta pesquisa realiza uma revisão da literatura separando os temas-chaves em agrupamentos temáticos, fornecendo uma agenda para pesquisas futuras com menos subjetividade. Registros bibliográficos foram acessados a partir do banco de dados da *Web Of Science* e *Scopus*, em várias disciplinas como ciência da computação, engenharia, ciência de decisão, ciências sociais, gestão de negócios e matemática. (SAHOO, 2021; MISHRA *et al*, 2018; KOUSIS e TJORTJIS, 2021; BUI *et al*, 2021) sugerem que os estudos bibliométricos são sintetizadores imparciais de conteúdo literário.

Para Levy e Ellis (2006), conhecer o atual estágio do corpo de conhecimentos sobre um determinado tema é o primeiro passo em um projeto de pesquisa. Assim um estudo bibliométrico é útil para (LEVY; ELLIS, 2006):

- Ajudar o pesquisador no dimensionamento e compreensão do corpo de conhecimento referente a um determinado assunto, incluindo identificar pesquisas que já foram realizadas, o que falta pesquisar, quais são as lacunas;
- Prover embasamento teórico para o estudo proposto;
- Apresentar as devidas justificativas para a condução do estudo, e qual a contribuição original para o corpo de conhecimento e/ou teoria;
- Contribuir para melhor definir e estruturar o método de pesquisa, objetivos e questões para o estudo proposto.

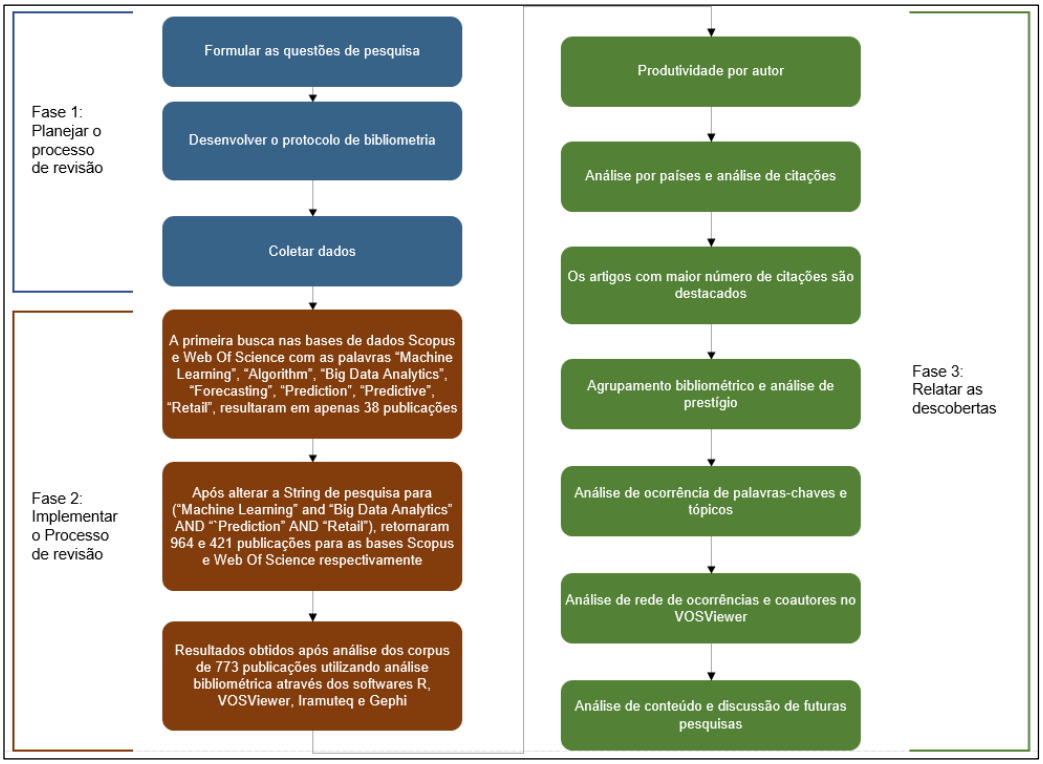
Levy e Ellis (2006), descrevem um estudo bibliométrico por meio de um processo. Os autores adotam a definição de processo como “sequência de passos e atividades”. Para alcançar esses resultados, os autores definem três fases principais: Entrada; Processamento; e

BY-NC-SA

Saída. Na fase “entrada” estão as informações preliminares que serão processadas, adotada como Fase 1 nesta pesquisa. Na fase “processamento” deve ser aplicado um protocolo que filtre o número de publicações conforme o tema da pesquisa, chamada de Fase 2 nesta pesquisa. Por fim na fase “saída” serão gerados os relatórios com a síntese dos resultados, identificado como Fase 3 nesta pesquisa. Estas três fases são detalhadas na Figura 1.

Figura 1

Desenho da Pesquisa



Fonte: Resultados da Pesquisa.

2.1 Planejar o processo de revisão - Fase 1

Isso é feito formulando as questões de pesquisa e coletando dados da *Web of Sciense* e *Scopus*. A pesquisa revelou publicações em periódicos de renome, fornecendo informações valiosas a respeito de *Big Data* e ML em vários países. Assim, uma busca foi realizada nos bancos de dados para recuperar os registros de publicações usando as constantes de pesquisa “*Machine Learning*", "*Algorithm*" e "*Big Data Analytics”.* Inicialmente algumas combinações com palavras-chaves foram adicionadas a pesquisa como: "*Forecasting*", "*Prediction*", "*Predictive*" e “*Retail*”, mas os registros bibliométricos foram limitados a apenas 38 publicações, desta forma o termo “*Algorithm*” foi excluído da *string* de pesquisa, mantendo

somente: ("*Machine Learning*" AND "*Big Data Analytics*" AND "*Prediction*" AND "*Retail*"), com este método foi possível extrair 964 publicações da base *Scopus* e 421 publicações da *Web of Sciense* com publicações a partir do ano de 2013 até o mês de maio/2023, quando a pesquisa foi executada.

As palavras chaves selecionadas, foram obtidas a partir de uma análise inicial dos artigos por meio do software *Gephi* utilizando o atributo “Total Link Strength” o qual indica a coocorrências de palavras-chaves que ocorrem com maior frequência, conforme descrito na seção 3.2.4.

Por meio de uma análise descritiva no corpus dos artigos, assim como uma análise quantitativa das publicações foi avaliado as tendências de citações para os anos de 2013 a maio/2023. Seguido por uma análise dos principais autores prolíficos e dos principais países com publicações em *Big Data* e ML.

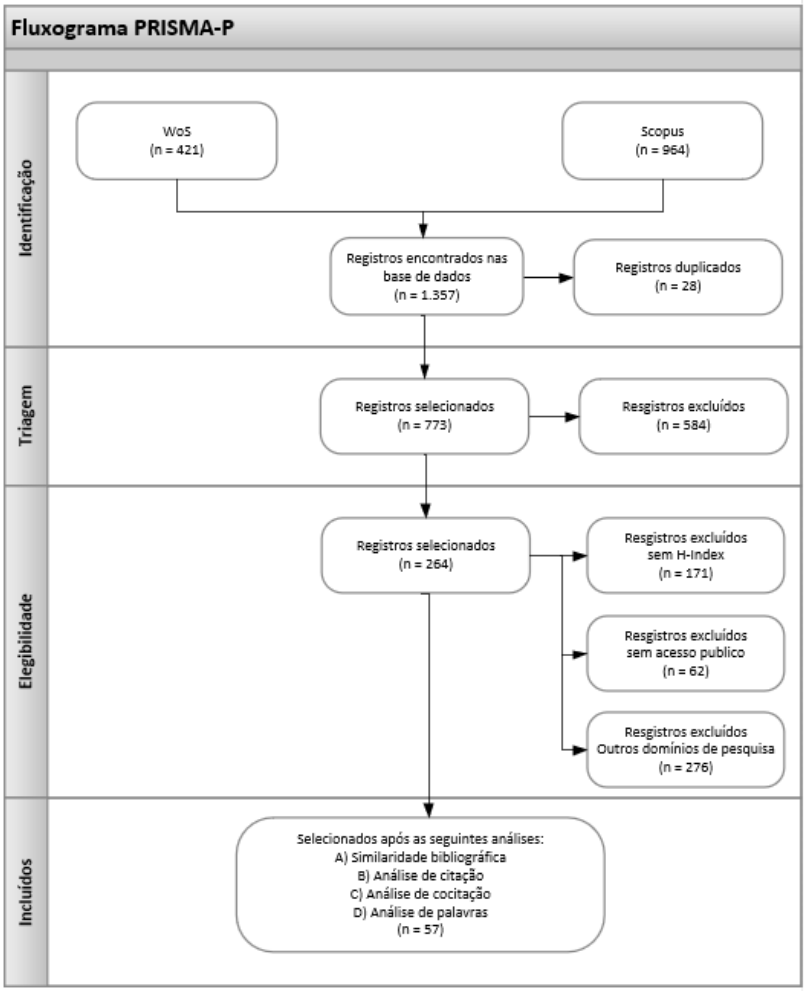
2.2 Implementação do protocolo bibliométrico - Fase 2

A Figura 2, apresenta o fluxograma do processo de seleção das publicações científicas em cada uma das quatro etapas previstas pelo protocolo PRISMA-P: Identificação, triagem, elegibilidade e documentos incluídos para análise.

(cc) BY-NC-SA

Figura 2

Fluxograma PRISMA-P



Fonte: Resultados da Pesquisa.

Na etapa de triagem, foram removidas 28 publicações em duplicidade, 53 Livros, 81 capítulos de livros, 272 artigos de conferências, 172 revisões e 6 materiais editoriais, totalizando 584 registros excluídos, desta forma resultaram 773 registros selecionados para a próxima etapa.

Na etapa de elegibilidade, foram excluídos 171 artigos sem H-Index, 62 artigos cujo acesso não era público e 276 artigos cuja pesquisa estava relacionado a algum dos seguintes domínios: Medicina, Engenharia e Arquitetura, Educação, Psicologia, Agricultura e Biociências, Artes e Humanidades, Bioquímica, Geociências e Logística.

Foram selecionadas 264 publicações para seguirem para a última etapa, na qual foram aplicadas as seguintes técnicas:

(a) Similaridade bibliográfica. Ajuda a identificar um conjunto de publicações com a maior semelhança bibliográfica, medida em termos de número de referências compartilhadas. Essa semelhança reflete o grau de similaridade na pesquisa e uma possível semelhança nas direções futuras de pesquisa. Neste artigo, a similaridade bibliográfica é realizada para autores, instituições / universidades e periódicos para extrair *insights* como ilustrado subseção 3.2.

(b) Análise de citação. A análise de tendência de citação é realizada na subseção 3.1 para avaliar a contribuição em termos de quantos documentos estão se referindo ao artigo e/ou citando-o. Esta técnica identifica os principais autores, instituições e países em termos de citações. Além disso, também é utilizado a métrica *PageRank* para medir o prestígio do artigo em periódicos de renome.

(c) Análise de cocitação. A análise de cocitação é realizada na subseção 3.2.2 para identificar semelhanças entre os títulos das publicações e agrupá-las em diferentes temas/tópicos com base em sua estrutura conceitual. A análise de cocitação é complementada com uma análise de palavras para identificar coocorrências de palavras-chave.

(d) Análise de palavras. É conduzida para visualizar a frequência de ocorrência de um determinado autor e palavras-chave de índice sobre um tema de pesquisa na subseção 3.2.4. Também é necessário analisar a evolução da mudança temática ao longo do tempo para identificar tópicos emergentes e aqueles que estão saturados.

Após as análises eliminatórias acima, foram mantidas 57 publicações para análise qualitativa.

2.3 Relatando as descobertas - Fase 3

Está estruturada em termos de análises descritivas e bibliométricas. A análise descritiva inclui o número total de publicações e citações. As informações foram extraídas por meio da biblioteca "*biblioAnalysis*" função contida no pacote bibliométrico no software R. Esta fase compreende a contribuição e a extensão da colaboração em pesquisa considerando vários autores e países. Uma análise considerando autores mais citados também é conduzida para entender as principais pesquisas dos autores mais citados.

Foi realizada uma análise bibliométrica das publicações usando o software R, por meio do RStudio 2022.07.2 *Build* 576, para identificar ligações bibliográficas (entre autores),

BY-NC-SA

cocitações e coocorrências usando o pacote "*bibliometrix"* v4.0.1 em R, que contém a função predefinida "*biblioNetwork*". A ferramenta *Gephi* v0.10 foi usada para realizar análise de prestígio. VOSviewer v1.6.17 foi usado para mapear coocorrências de palavras-chave, colaboração das universidades e análise de cocitação de periódicos. Os temas emergentes e agrupamento dos títulos foram identificados usando a análise de nuvem de palavras por meio do *Iramuteq* v0.7 *alpha 2*. A análise de estrutura conceitual foi realizada também por meio do "*bibliometrix*").

3 RESULTADOS

Os resultados da análise descritiva e bibliométrica são demonstrados nesta seção.

3.1 Análise descritiva e de citações das publicações por periódicos

Uma análise descritiva dos arquivos exportados "*BibTex*" de 2013 a maio/2023 foi conduzida e exibida na Tabela 1, utilizando a função "*summary*" da biblioteca "*bibliometrix*". Após consolidar as publicações da base de dados Scopus (964 documentos) e WoS (421 documentos), foram eliminadas 28 publicações duplicadas, resultando 1.357 publicações.

Das 1.357 publicações, 773 são artigos de pesquisa, 53 são livros, 81 capítulos de livros, 272 artigos de conferências, 172 revisões e 6 materiais editoriais.

A distribuição de frequência das palavras-chave por autor é 3.727, o que implica que estas palavras-chave são frequentemente utilizadas pelos autores em publicações de ML e BDA. A distribuição de palavras-chave extraídas dos artigos de periódicos no domínio é 4.407. O número de autores foi 4.205, com 5.233 aparições, incluindo autoria única e aparições de vários autores. Dos 4.205 autores, 89 autores publicaram artigos com um único autor, enquanto os 4.116 autores restantes publicaram artigos com vários autores, indicando alto grau de colaboração de pesquisa nos artigos publicados. O número de autoria única de documentos é de apenas 92, enquanto o restante de 1.265 é documentos com autoria múltipla. O número de documentos por autor, ou seja, a proporção do número total de documentos (1.357) para o número total de autores (4.205), é 0,323. A proporção recíproca desta métrica, o número de autores por documento (4.205/1.357), é 3,10, enquanto o número de coautores por documento é 3,86. O índice de colaboração, ou seja, a proporção do número total de autores em documentos de autoria múltipla para o número de documentos de autoria múltipla (4.116/1.265), é de 3,25, indicando, assim, que para um documento de autoria múltipla, há

(cc) BY-NC-SA

aproximadamente três autores. Este achado corrobora um índice robusto na rede de colaboração.

Tabela 1

Resumo da análise descritiva dos registros

Descrição	Resultados
MAIN INFORMATION ABOUT DATA	
Timespan	2013:2023
Documents	1.357
Sources (Journals, Books, etc)	828
Average citations per documents	21.58
DOCUMENT CONTENTS	
Keywords Plus (ID)	4.407
Author's Keywords (DE)	3.727
AUTHORS	
Authors	4.205
Author Appearances	5.233
Authors of single-authored documents	89
Authors of multi-authored documents	4.116
AUTHORS COLLABORATION	
Single-authored documents	92
Documents per Author	0.323
Authors per Document	3.10
Co-Authors per Documents	3.86
Collaboration Index	3.25

Fonte: Resultados da Pesquisa

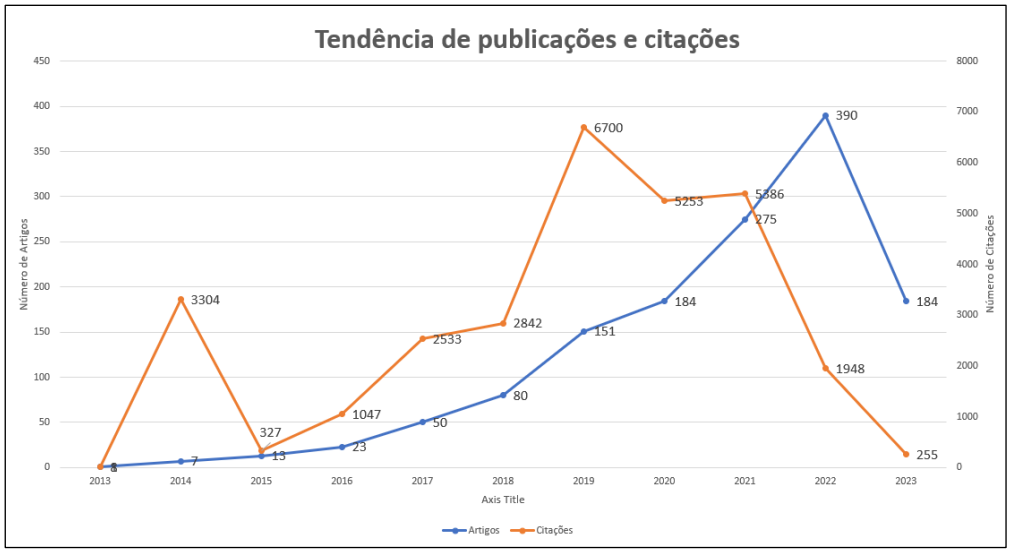
A média de citações por documento é de 21,58, o que implica que os artigos dos periódicos são citados em uma média de quase 22 vezes.

Por meio da Figura 3 foi obtida uma perspectiva do número de artigos no domínio ML e BDA publicados ao longo de 10 anos de estudo, identificando a tendência de publicações (medida em termos de número de artigos) e a tendência de citações para o período de 2013 a maio/2023.

(cc) BY-NC-SA

Figura 3

Tendência de publicações

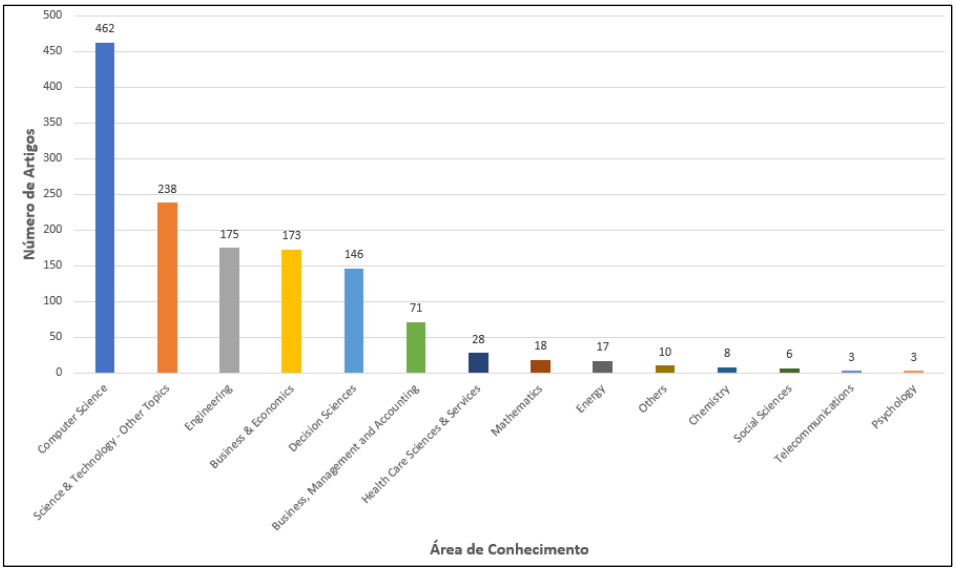


Fonte: Resultados da Pesquisa.

A partir da Figura 4, Ciência da Computação é considerado o domínio predominante com 34% dos artigos nessa categoria, seguido por Ciência e Tecnologia (18%), Engenharia e Negócios (13%), Ciências de Decisão (11%), Gestão Empresarial (5%) e Saúde (2%). Outras áreas com uma contribuição marginal para ML e BDA incluem Química e Psicologia. Essas áreas possuem escopo de pesquisa mais interdisciplinar.

Figura 4

Publicações por área de conhecimento



Fonte: Resultados da Pesquisa

3.1.2 Análise por autor

Considerando o alto grau de colaboração, identificamos os autores mais citados em termos do número de publicações totais (NP), assim como o número de citações (TC), e o número de citações por publicação (C/P), como mostrado na Tabela 2, Yunhao Liu (abreviado como Liu Y) da China foi considerado um nome frequente e familiar com um alto valor de C/P de 294,7, seguido por Yogesh Kumar Dwivedi (Dwivedi Y.) do Reino Unido com um valor de C/P de 137,6. Os próximos autores mais produtivos em termos de C/P foram Kar A. da Polônia e Wang Y. da China. Esses autores são renomados estudiosos do campo de ML contribuindo com o estado da arte em artigos de pesquisa, promovendo aos pesquisadores e profissionais um grande arcabouço de conhecimento. Para analisar a produtividade dos autores em termos de citações totais (TC) ao longo do período analisado, foi extraído da biblioteca '*Bibliometrix'* o gráfico apresentado na Figura 5. A função *AuthorProdOverTime* calcula e plota a produtividade dos autores em termos do número de publicações e o total de citações por ano.

O h-index é medido em termos do número 'h' de publicações com citações mínimas de vezes 'h'. O g-index indica o número 'g' de artigos com pelo menos citações 'g^2'. O m-index do artigo é calculado com a razão entre o h-index e o número de anos desde que a primeira publicação do autor foi realizada. Os índices 'h', 'g' e 'm' são apresentados como medidas de citação e produtividade, com Wang Y., Wang X. e Chen G. tendo os maiores valores de h-index (6; 5 e 5) respectivamente, g-index (7; 5 e 5) e m-index (1,2; 0,625 e 0,8333333). Além disso, os resultados mostram que para Yunhao Liu, 4 artigos foram citados 4 vezes, e os 6 principais artigos foram citados em pelo menos 6^2 (36 vezes). Uma vez Yunhao Liu está ativo desde 2015 (8 anos ativos de publicação), seu m-index é (4/8) = 0,5.

(cc) BY-NC-SA

Tabela 2

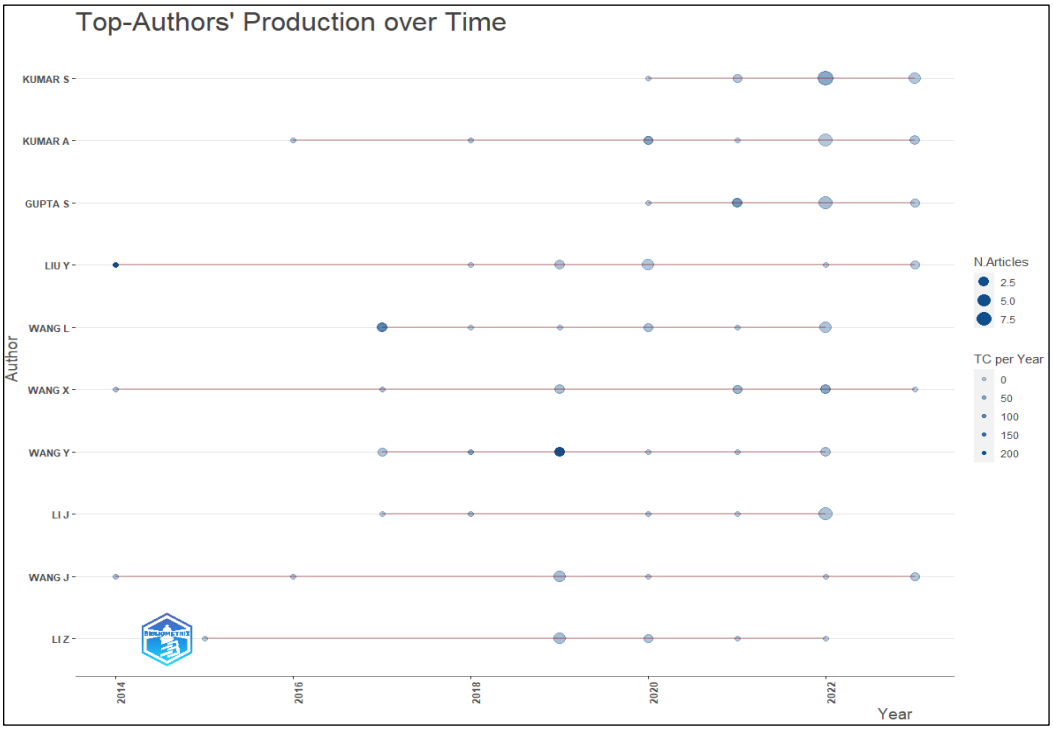
Os dez autores mais produtivos em publicações e citações

Author	Country	NP	TC	C/P	h-index	g-index	m-index	PY-start
CHEN G	China	5	98	19,6	5	5	0,8333333	2016
DWIVEDI Y	United Kingdom	5	688	137,6	4	5	1,3333333	2019
GUPTA S	India	4	62	15,5	3	4	1,5000000	2020
KAR A	Poland	3	311	103,7	3	3	0,6000000	2017
LI Z	China	7	51	7,3	4	7	0,5714286	2015
LIU Y	**China**	**6**	**1768**	**294,7**	**4**	**6**	**0,5000000**	**2014**
WANG J	Japan	7	216	38,8	4	7	0,5000000	2014
WANG L	China	6	553	92,2	4	6	0,8000000	2017
WANG X	China	5	224	44,8	5	5	0,6250000	2014
WANG Y	China	7	697	99,6	6	7	1,2000000	2017

Fonte: Resultados da Pesquisa

Figura 5

Função AuthorProdOverTime da biblioteca Bibliometrix



Fonte: Resultados da Pesquisa.

3.1.3. Análise por país

Para identificar os países com maior número de citações, é usado a métrica de citações por publicação (C/P), na qual indica que o Reino Unido, Canada e China são os três países que mais possuem citações com C/P de 55.9, 42.2 e 32.7 respectivamente conforme Tabela 3. Em termos de número de publicações (NP), os três principais países incluem India, China e Estados Unidos, enquanto Singapura e Corea são os menos citados.

Tabela 3

Os dez países mais produtivos em publicações e citações

Country	NP	TC	C/P
India	186	2502	13,4
China	154	5043	32,7
United States	110	2721	24,7
United Kingdom	65	3632	55,9
Germany	36	425	11,8
Canada	33	1393	42,2
Italy	33	425	12,9
Australia	30	626	20,9
Singapore	30	422	14,1
Korea	29	425	14,6

Fonte: Resultados da Pesquisa

3.1.4 Análise por universidade

Foi realizada uma análise por universidade, exibida na Tabela 4. As universidades mais produtivas são: Universidade de Michigan, Universidade Estadual da Pensilvânia e a Universidade King Saud, em termos de números de publicações (NP) e citações (TC). A contribuição da Universidade de Hong Kong e da Universidade da Carolina do Sul é mais escasso.

(cc) BY-NC-SA

Tabela 4

Publicações por universidades

Short-name	University	Country	NP	TC
UNIV MICHIGAN	University of Michigan	United States	15	232
PENN STATE UNIV	The Pennsylvania State University	United States	11	105
KING SAUD UNIV	King Saud University	Saudi Arabia	10	107
SWANSEA UNIVERSITY	Swansea University	United Kingdom	9	1217
UNIV WEST ENGLAND	University of the West of England, Bristol	England	8	156
THE HONG KONG POLYTECHNIC UNIVERSITY	The Hong Kong Polytechnic University	Hong Kong	8	417
TSINGHUA UNIVERSITY	Tsinghua University	China	7	2115
SEJONG UNIV	Sejong University	South Korea	7	142
UNIV SOUTH CAROLINA	University of South Carolina	United States	7	70
CITY UNIV HONG KONG	City University of Hong Kong	China	7	14

Fonte: Resultados da Pesquisa.

3.1.5 Análise por periódico

Os 10 principais periódicos em termos de número de publicações (NP) e total de citações (TC) estão resumidos na Tabela 5. "*IEEE Access"* e "*International Journal of Information Management"* aparecem com 34 e 12 publicações respectivamente e com 1662 e 827 citações, demonstrando que tais periódicos possuem grande relevância no domínio de ML e BDA. Por meio de uma análise descritiva em conjunto com uma análise bibliométrica e de rede, a próxima seção abordará a identificação de acoplamento bibliográfico, cocitações, e coocorrências de tópicos.

Tabela 5

Publicações por periódico

Periódico	NP	TC
IEEE Access	34	1662
International Journal of Information Management	12	827
Journal of Big Data	11	55
International Journal of Production Research	11	297
Annals Of Operations Research	10	132
Sustainability (Switzerland)	9	75
Decision Support Systems	5	139
Journal of Business Research	5	73
Computers and Industrial Engineering	5	37
Industrial Management and Data Systems	5	10

Fonte: Resultados da Pesquisa.

3.2 Análise bibliométrica e de rede

Foi realizada uma análise de acoplamento por autor e uma rede de cocitações para analisar a colaboração e dependência mútua de pesquisa das citações entre autores, universidades e periódicos. Forma identificados os temas emergentes e tópicos relevantes utilizando análise de coocorrências de palavras-chave e agrupamento bibliométrico dos documentos. Além disso, uma análise de série temporal de *clusters* é apresentada para reconhecer temas emergentes com escopo para pesquisas futuras. Os resultados foram validados por um gráfico de coocorrências de palavras chaves, junto com uma nuvem de palavras dos títulos, resumos e temas. Foi realizada uma análise de rede de cocitação para identificar a direção de novas pesquisas com base em artigos anteriores com alto número de citações.

3.2.1 Acoplamento bibliométrico de autores

Foi desenvolvido o acoplamento bibliométrico entre os autores para identificar a colaboração entre os mesmos para analisar informações de ML e BDA, conforme apresentado na Tabela 6 e Figura 6. Em termos de acoplamento bibliográfico, Wang, Y. e Dwivedi estão em *clusters* diferentes, porém é por meio destes dois autores que os demais grupos conseguem se relacionar em termos de colaboração. Ao mesmo tempo, Rana, N. P. e Tamilmani, K. estão

(cc) BY-NC-SA

no mesmo *cluster* e são colaboradores ativos.

Tabela 6

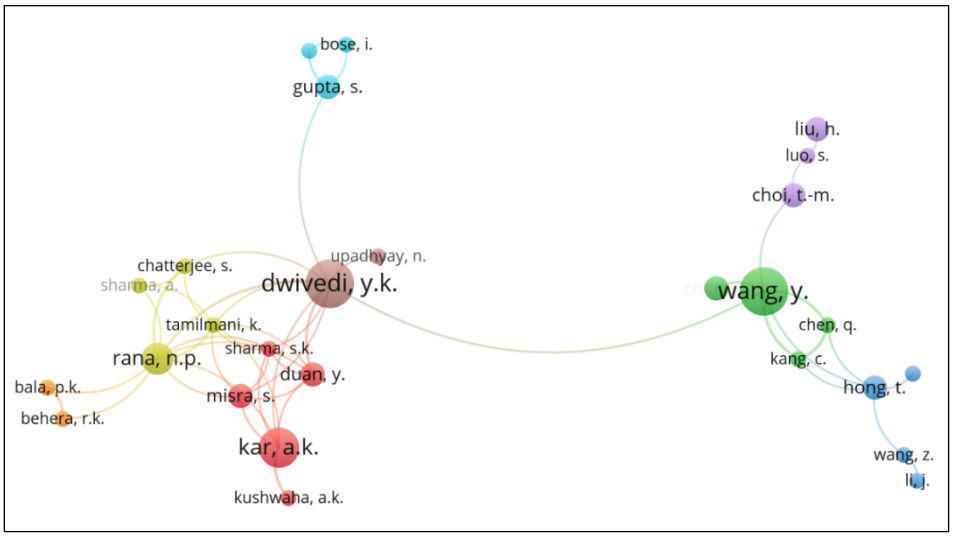
Os principais autores em termos de acoplamento bibliográfico

Autor	Total Link Strength
Rana, N. P.	13
Dwivedi, Y. K.	12
Tamilmani, K.	9
Kar, A. K.	8
Wang, Y.	8
Duan, Y.	7
Misra, S.	6
Sharma, S. K.	6
Chen, Q.	5
Kang, C.	5

Fonte: Resultados da Pesquisa

Figura 6

Acoplamento por Autor



Fonte: Resultados da Pesquisa

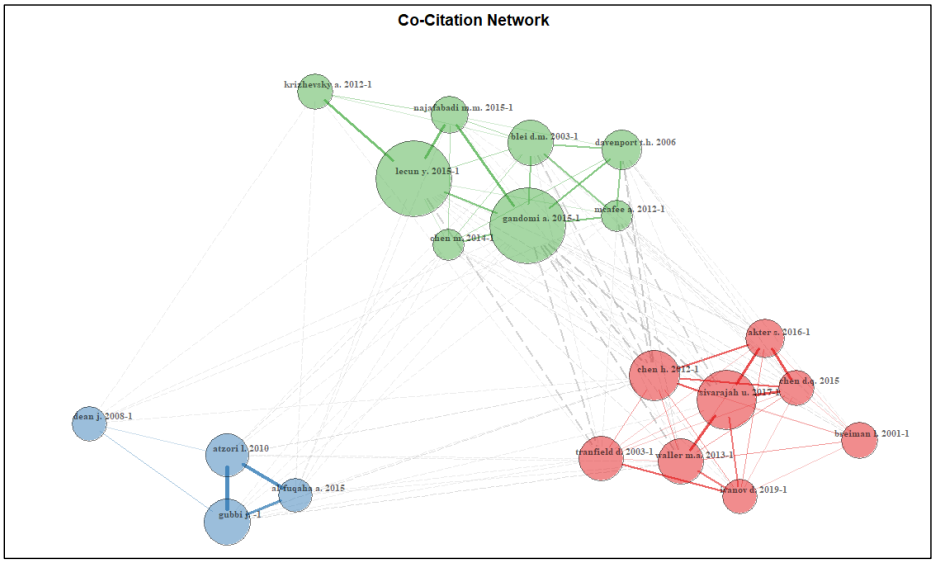
3.2.2 Análise de rede de cocitação

Foi elaborada uma rede de cocitação utilizando o software R para analisar a força de citação entre os principais autores. Na Figura 7 é apresentada, por meio de círculos, o número

de citações por autor, ou seja, quanto maior o tamanho do círculo, maior será a força de citação. Além disso, o número ao lado do ano indica a força da citação. Por exemplo, Chen M. (2014) é denotado como "Chen M. 2014-1", devido a sua força de citação ser igual ao "Bleid D.M. 2003-1". Com relação às citações locais, ou seja, a frequência de dois autores sendo citados no mesmo artigo, são indicados por linhas sólidas. Em contrapartida, as citações globais mútuas entre dois diferentes artigos, quando ambos são citados em um terceiro artigo, são indicados por linhas tracejadas. Verifica-se que Gandomi (2015) e Sivarajah (2017) são frequentemente citados e são indicados por grandes círculos verde e vermelho respectivamente. As linhas sólidas indicam as citações locais dos autores, por exemplo, Waller (2013) tem sido frequentemente citado por Sivarajah (2017) no mesmo artigo, como fica evidente pela linha sólida vermelha. As linhas pontilhadas, indicam citações globais, por exemplo, Tranfield (2003) e Chen (2012) são frequentemente citados pelo terceiro Gandomi (2015), e, portanto, uma linha tracejada é desenhada de Tranfield e Chen para Gandomi, indicando a citação mútua global.

Figura 7

Rede de cocitação



Fonte: Resultados da Pesquisa.

3.2.4 Análise de coocorrências de palavras-chaves

Na Figura 8, as coocorrências de palavras-chaves medidas em termos de força total de ligação são consideradas elevadas entre o "*Machine Learning"* e "*Supply Chains"* (círculos

violetas), enquanto o termo como "*Big Data*" ocorre frequentemente com "*Data Analytics*", "*Advanced Analytics*" e "*Data Handling*" (círculos vermelhos), indicando que "*Advanced Analytics*" e "*Data Handling*" estão surgindo em áreas de pesquisa em "*Big Data*". Os termos "*Artificial Intelligence*", "*Learning Systems*" e "*Deep Learning*" são temas com coocorrências frequentes (círculos amarelos). Além disso, os termos "*Big Data Analytics*", "*Decision Making*" e "*Supply Chain Management*" estão correlacionados (círculos azuis), enquanto os termos "*Data Mining*", "*Forecasting*", "*Sales*", "*Commerce*" e "*Predictive Analytics*" estão correlacionados (círculos verdes). As palavras-chaves que ocorrem com frequência em um determinado tema são identificadas pelos círculos que possuem a mesma cor.

Coocorrências entre temas (indicados por linhas tracejadas de cores diferentes como violeta, vermelha, amarela e verde) são encontrados entre o "*Big Data*" e todos os demais grupos, indicando uma forte ligação.

As principais palavras-chaves por autor, medido pelo "*Total Link Strength*" são: *Machine Learning*, *Big Data*, Big *Data Analytics*, *Artificial Intelligence*, *Data Mining*, *Deep Learning*, *Data Analytics*, *Analytics*, *Internet of Things* e *Prediction.*

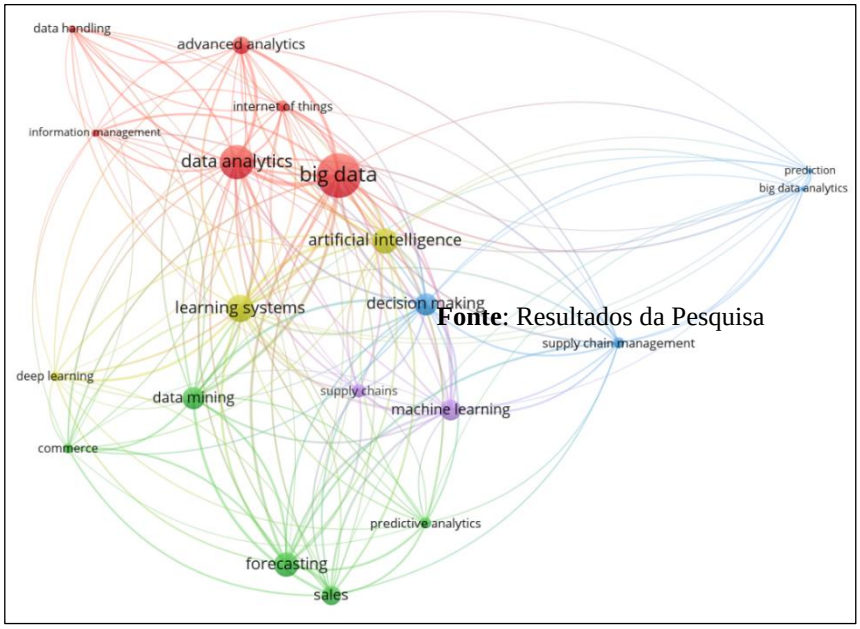
As principais palavras-chaves de índice medido pelo "*Total Link Strength*" são: *Big Data*, *Data Analytics*, *Learning* Systems, *Artificila Intellingence*, *Forecasting*, *Decision Making*, *Data Mining*, *Machine Learning*, *Sales* e *Advanced Analytics.*

O gráfico de coocorrências de palavras-chaves apresentado na Figura 8, também pode ser interpretado para destacar os tópicos específicos que aparecem com frequência, assim como os tópicos baseados em palavras-chaves gerais que possuem um escopo de cobertura maior. O tamanho dos círculos indica a frequência em que determinada palavra-chave ocorre. É observado que tópicos sobre *Big Data*, *Data Analytics*, *Artificial Intelligence* e *Machine Learning* são recorrentes. Ao mesmo tempo, *Deep Learning* e *Data Handling* são tópicos menos frequentes, ou seja, com menor número de artigos, mas com alto interesse para pesquisas futuras. Além disso, o agrupamento bibliográfico é realizado para compreender as diferentes áreas de pesquisas emergentes, fornecendo *insights* em direção de futuras pesquisas.

(cc) BY-NC-SA

Figura 8

Coocorrência de palavras-chaves



Fonte: Resultados da Pesquisa

Fonte: Resultados da Pesquisa.

3.2.5 Agrupamento bibliográfico por meio de mapa de estrutura conceitual

O agrupamento bibliográfico é adotado para agrupar a pesquisa no domínio de ML e *Big Data* em diferentes áreas temáticas. Temas semelhantes são agrupados no mesmo *cluster*. Além disso, a densidade dos clusters podem servir como medida da extensão da pesquisa realizada na área temática em particular. Os agrupamentos considerados densos são considerados áreas saturadas para pesquisa, enquanto *clusters* com dados escassos são considerados como tendo espaço para pesquisas futuras. A análise do *cluster* bibliográfico é realizada por meio do mapa de estrutura conceitual utilizando a escala multidimensional (MCA), cujos gráficos de dendrograma estão representados na Figura 9. Os temas estão agrupados em 5 grupos (*clusters*).

Cluster 1: *Machine Learning Techniques*, *Decision Makes*, *Sales* e *Forecasting*

Estudos, que tratam de ML estão no cluster mais denso, destacado em verde na Figura 9. Conforme relatam L´Heureux *et al* (2017), a revolução do *Big Data* promete transformar a forma como vivemos, trabalhamos e pensamos, permitindo a otimização de processos, capacitando a descoberta de *insights* e melhorando a tomada de decisões. A realização desse grande potencial depende da capacidade de extrair valor desses dados massivos por meio da

(cc) BY-NC-SA

análise de dados; o aprendizado de máquina é fundamental por causa de sua capacidade de aprender com os dados e fornecer *insights*, decisões e previsões baseadas em dados. No entanto, as abordagens tradicionais de aprendizado de máquina foram desenvolvidas em uma era diferente e, portanto, são baseadas em várias suposições, como o conjunto de dados se encaixando inteiramente na memória, o que infelizmente não é mais verdadeiro neste novo contexto.

A inteligência artificial (IA) existe há mais de seis décadas e vem amadurecendo com o tempo. A ascensão do superpoder da computação e das tecnologias de *Big Data* parecem ter potencializado a IA nos últimos anos. A nova geração de IA, está se expandindo rapidamente e voltou a ser um tópico atraente para pesquisa. Duan *et al.* (2019), investigam os desafios associados ao uso e impacto de sistemas baseados em IA revitalizados para a tomada de decisões e oferecem um conjunto de propostas de pesquisa para pesquisadores de sistemas de informação (SI).

Mesmo com mais de duas décadas de desenvolvimento contínuo, a aprendizagem de dados desequilibrados ainda é um foco intenso de pesquisa. Com a expansão do aprendizado de máquina e mineração de dados, combinado com a chegada da era do B*ig Data*, foi possível obter uma visão mais profunda sobre a natureza da aprendizagem desequilibrada, ao mesmo tempo que enfrenta novos desafios emergentes. Métodos de nível de dados e de algoritmo estão constantemente sendo melhorados e abordagens híbridas ganham popularidade crescente. As tendências recentes se concentram em analisar não apenas a desproporção entre os aprendizados, mas também outras dificuldades embutidas na natureza dos dados. Novos problemas da vida real motivam os pesquisadores para se concentrar em eficiência computacional, adaptável e métodos em tempo real. Krawczyk (2016), discuti questões e desafios em aberto que precisam ser resolvidos para desenvolver ainda mais o campo de aprendizagem desequilibrada. Foram identificadas algumas áreas vitais de pesquisa neste tópico, cobrindo todo o espectro de aprendizagem de dados desequilibrados: classificação, regressão, agrupamento, fluxos de dados, análise de *Big Data* e aplicativos, por exemplo, em mídia social e visão computacional.

Cluster 2: *Business Intelligence* (BI) e *Big Data Analytics* (BDA)

Estudos que examinam BI e BDA estão no cluster em vermelho mostrado na Figura 9. Os resultados indicam que muitas pesquisas emergentes aceitas e publicadas em periódicos se

enquadram nesta categoria. Este *cluster* é, portanto, altamente dominante. Os estudos neste grupo temático examinam a necessidade de adoção BI e técnicas de BDA. Mais notavelmente, Chen, Mao e Liu (2014) revisam os antecedentes e o estado da arte do *Big Data* focando nas quatro fases da cadeia de valor do BD: geração de dados, aquisição de dados, armazenamento de dados e análise de dados. Foram analisados artigos neste domínio em termos de tópicos de pesquisa emergentes, como também os maiores pesquisadores e as contribuições mais importantes. Além disso, BD apresenta uma característica única, comparando com dados tradicionais, ele é comumente não estruturado, necessitando de mais análise em tempo real, conforme relato por (HU *et al*, 2014).

A importância do *Big Data* na melhoria do desempenho de uma empresa é corroborada no estudo de Choi *et al*. (2018), que explorou as grandes técnicas de análise relacionadas a dados, identificando seus pontos fortes e fracos, bem como as principais funcionalidades. Desta forma, foram discutidas estratégias de análise de BD para superar os respectivos desafios computacionais e de dados.

Cluster 3: IoT, *Data Handling* e *Cloud Computing*

Estudos em IoT, *Data Handling* e *Cloud Computing* foram amplamente pesquisados, marcado em violeta na Figura 9. Gill, Tuli e Xu (2019), exploram como os três paradigmas emergentes (*Blockchain*, IoT, e IA), influenciarão os futuros sistemas de computação em nuvem e propuseram um modelo conceitual para a futurologia da nuvem para explorar a influência de paradigmas e tecnologias emergentes na evolução da computação em nuvem.

Calatayud, Mangan e Christopher (2019), exploram como a cadeia de abastecimento do futuro será autônoma e terá capacidades preditivas, trazendo ganho de eficiência em um ambiente cada vez mais complexo e incerto. O estudo é realizado através de uma revisão sistemática e multidisciplinar da literatura, revisando 126 artigos publicados no período de 1950-2018.

Ahmadi *et al* (2019), relatam que IoT é um ecossistema que integra objetos físicos, software e hardware para interagir com outros objetos. O envelhecimento da população, escassez de recursos de saúde e custos médicos crescentes tornam as tecnologias baseadas em IoT necessárias, as quais podem ser adaptadas para enfrentar estes desafios na área da saúde, sendo assim, esta revisão sistemática da literatura foi realizada para determinar a principal área de aplicação de IoT em saúde.

(cc) BY-NC-SA

Cluster 4: *Activity Recognition*, *Health*, *Medicine*, *Service Quality*, *Therapy*, *Heart Failure*, *Classification* e *Risk*.

Artigos que investigam as novas tecnologias relacionadas à área de saúde, são marcados em azul na Figura 9.

Com o crescimento do *Big Data* na área biomédica e de saúde, está sendo possível realizar análises precisas dos benefícios dos dados médicos de forma antecipada para detecção de doenças. Contudo, a precisão da análise é reduzida quando a qualidade dos dados médicos está incompleta. Além disso, diferentes regiões exibem características de certas doenças regionais, que podem enfraquecer a previsão de surtos de doenças. Desta forma, Chen *et al* (2017), simplificam os algoritmos de aprendizado de máquina para previsão eficaz de doenças crônicas e surto de doenças frequentes em comunidades.

Razavian *et al* (2015), apresentaram uma nova abordagem para a saúde da população, na qual modelos preditivos baseados em dados são aprendidos com base nos resultados de diabetes tipo 2. A abordagem permite a avaliação de risco a partir de dados de sinistros eletrônicos prontamente disponíveis em grandes populações. O modelo proposto revela os fatores de risco em estágio inicial e final. Foram coletados: reclamações, registros de farmácia, utilização de serviços de saúde e resultados laboratoriais de 4,1 milhões de indivíduos entre 2005 e 2009, em um conjunto inicial de 42.000 variáveis que juntas descrevem o estado de saúde completo e histórico de cada Indivíduo. O aprendizado de máquina foi então usado para aprimorar metodicamente o conjunto de variáveis preditivas e modelos de ajuste que preveem o início da diabetes tipo 2

Cluster 5: Data *Acquisition* e *Predictive Analytics*

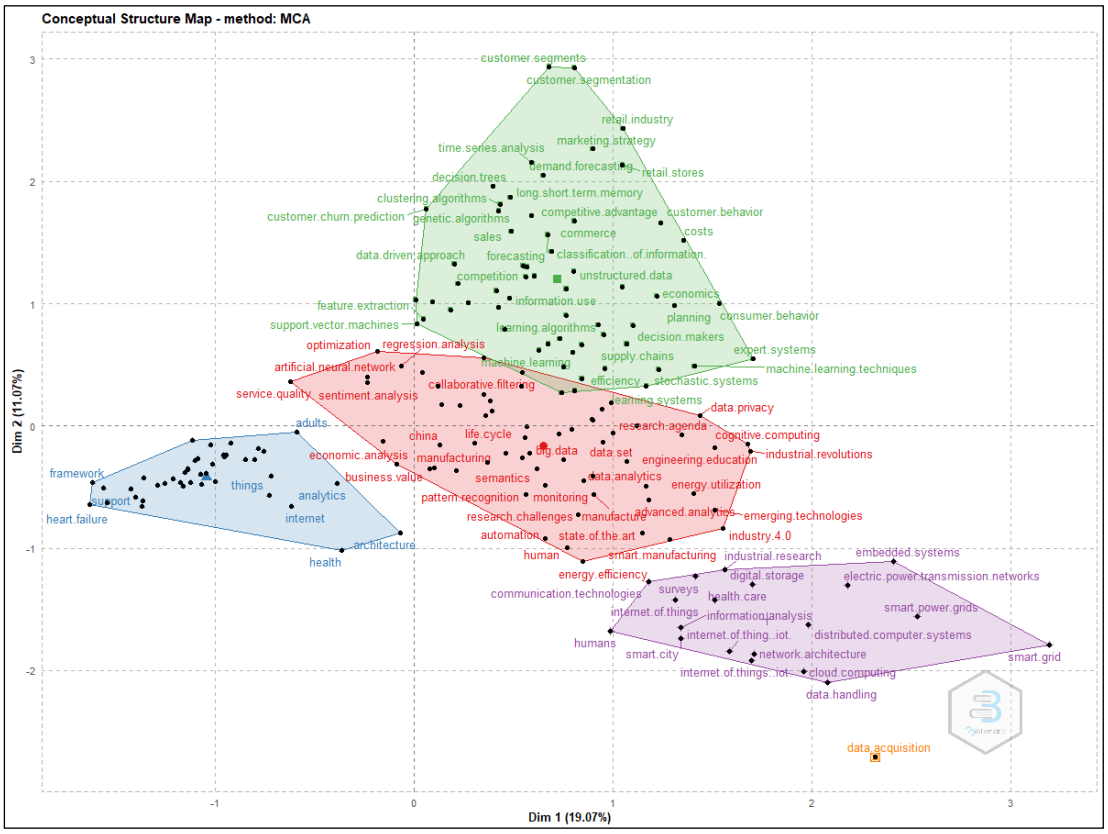
Estudos com o objetivo de analisar o impacto da aquisição de dados e da análise preditiva (marcado em laranja na Figura 9), são escassos, ou seja, demonstram que estudar o impacto da classificação e análises preditivas representam um amplo escopo para pesquisas futuras.

Tomar decisões adequadas é, de fato, um fator chave para ajudar as empresas que enfrentam os desafios das cadeias de abastecimento. Nguyen *et al* (2021), relatam duas abordagens baseadas em dados que permitem tomar melhores decisões na gestão da cadeia de fornecimento. É sugerido o método *Long Short Term Memory* (LSTM) baseado em rede de previsão de dados de série temporal multivariada e um método LSTM *Autoencoder*

combinado com a classe de algoritmo suporte de máquina de vetor (SVM) para detecção de anomalias nas vendas.

Figura 9

Mapa de estrutura conceitual (grupos de tópicos iniciais k = 5)



Fonte: Resultados da Pesquisa

Desta forma, os *clusters* 1, 2 e 3 são densos, ou seja, continuarão a emergir no futuro. Os *cluster* 4 e 5 estão com poucas pesquisas, mas são áreas de pesquisas emergentes. Uma visão geral dos 5 *clusters* é apresentada na Tabela 7.

(cc) BY-NC-SA

Tabela 7

Visão geral dos 5 clusters

Cluster	Foco Central	Principais temas explorados	Total Artigos	Artigo mais citado em cada cluster		
				Título do Artigo	Principal autor	Total de citações
1	ML, Decision Makes	Machine Learning Techniques, Decision Makes, Sales, Forecasting	353	Learning from imbalanced data open challenges and future directions	Krawczyk (2016)	1788
2	BI, Analytics	Business intelligence, Big Data Analytics	239	Big data: A survey	Chen *et al.* (2014)	4227
3	IoT	IoT, Data Handling, Cloud Computing	102	Transformative effects of IoT, Blockchain and Artificial Intelligence on cloud computing: Evolution, vision, trends and open challenges	Gill *et al.* (2019)	265
4	Activity Recognition	Activity Recognition, Health, Medicine, Service Quality, Therapy, Heart Failure, Classification, Risk.	48	Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case	Ivanov (2020)	1550
5	Data Acquisition	Data Acquisition, Predictive Analytics	31	Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews	Korfiatis *et al* (2019)	109

Fonte: Resultados da Pesquisa

A evolução dos grupos de temas em forma de linha do tempo foi traçada para 10 anos, em 5 intervalos (2013–2014, 2015–2016, 2017-2018, 2019-2020, 2021-2023), como mostrado na Figura 10. O *Cluster* 1 possui um agrupamento altamente denso com interesse contínuo e crescente, ou seja, indicando um aumento de artigos publicados no domínio de ML, confirmando seu domínio. O *Cluster* 2 também é considerado significativo devido ao crescimento na utilização de BI e BDA.

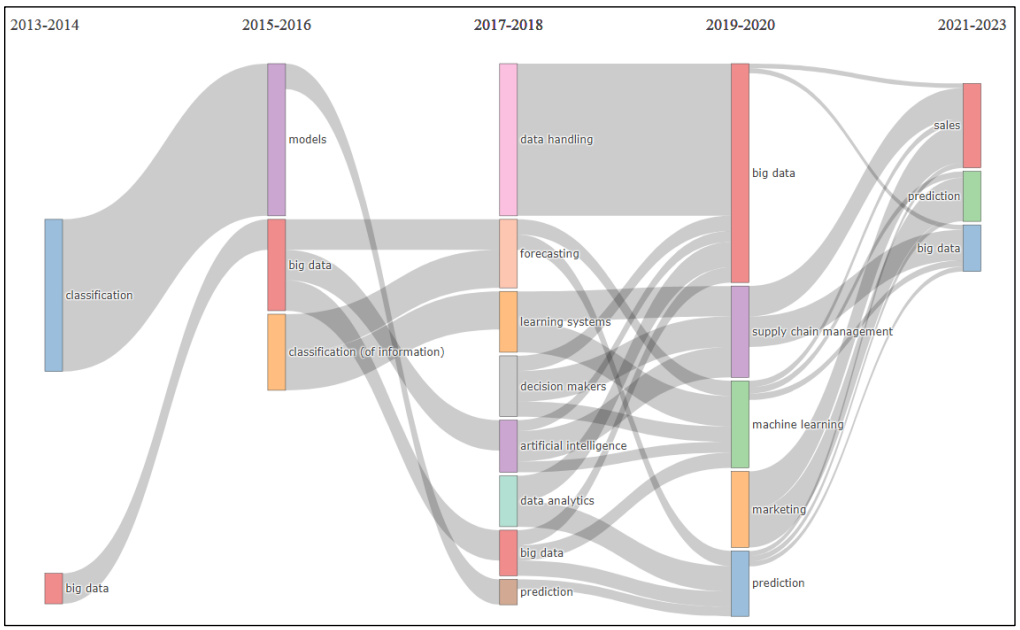
Os *Clusters* 3 e 4 continuam ganhando tração devido a utilização de dispositivos IoT, *Cloud Computing*, assim como a detecção de doenças de forma antecipada. No entanto, no *Cluster* 5 estão os tópicos com escassa cobertura indicando baixo interesse nos temas dos artigos, ou seja, demonstram que estudar o impacto da classificação e análises preditivas representam uma área com baixa exploração acadêmica.

A Figura 10 ilustra o mapa de evolução temática para os anos de 2013-2023. O período de 2013 a 2014 gira em torno do tema "*Big Data"* e métodos de classificação para

tratar a grande quantidade de dados. Gradualmente, outros temas evoluíram no período 2015 a 2016, por exemplo "*Learning Systems*" e "*Classification (of information)*", indicando a necessidade de utilizar sistemas de aprendizagem para classificar as informações. O período de 2017 a 2018 se destacam os termos *"Artificial Intelligence", "Decision Making", "Deep Learning"* e *"Prediction"*, indicando a busca de mecanismos para apoiar a indústria na tomada de decisão e a predição de eventos. Em 2019 a 2020, os termos "*Prediction*", "*Machine Learning*", "*Big Data*" e "*Decision Making*", ganham tração, indicando, assim, a demanda por utilização de métodos de aprendizado de máquina para tomada de decisão. No período de 2021 a 2023, fica evidenciado que os termos IoT, *Big Data*, *Decision Making* e *Prediction* continuam dominando os temas dos artigos acadêmicos.

Figura 10

Mapa de evolução temática para os anos 2013-2023



Fonte: Resultados da Pesquisa

3.2.6 Coocorrências de títulos

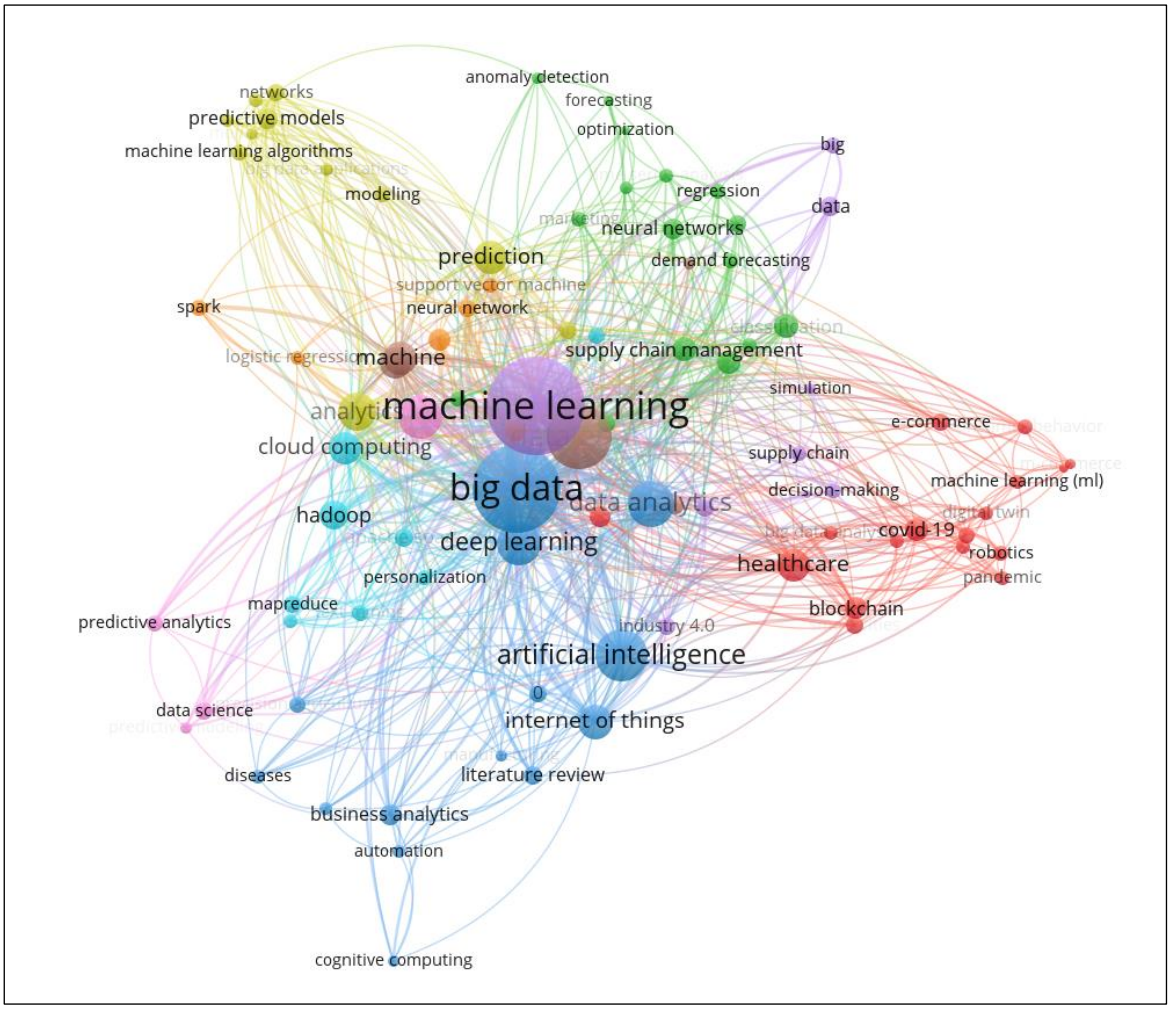
Na Figura 11 são analisadas as palavras que ocorrem com frequência e de forma concomitantes nos artigos, formando uma rede. As palavras com maior ocorrência incluem *Machine Learning*, *Big Data*, *Data Analytics*, *Deep Learning*, *Artificial Intelligence*. Outros domínios de pesquisa também são identificados como *Healthcare*, *Blockchain*, *Covid-19* e *Supply Chain Management*. Na verdade *Machine Learning* tem sido frequentemente usado

(cc) BY-NC-SA

com *Big Data* e *Data Analytics*, uma vez que ML depende de um conjunto de dados "*Data Set"* para ser treinado e aplicado.

Figura 11

Coocorrência de títulos



Fonte: Resultados da Pesquisa.

4 DISCUSSÃO

Para abordar a primeira questão de pesquisa, ou seja, o foco da pesquisa sobre ML, foi conduzida uma análise descritiva no corpus bibliométrico. Os resultados são resumidos nos 7 itens abaixo:

(a) Principais autores: Em termos de número de citações e resultados de acoplamento bibliográfico ilustrados nas Tabelas 2 e 6, foi descoberto que Yunhao Liu da China

é um nome frequentemente citado, seguido por Yogesh K. Dwivedia do Reino Unido. Em termos de número de publicações (NP), verifica-se que Zhenhua Li da China, J.W. Wang. do Japão e Yi Wang da China são os autores mais produtivos. Nripendra P. Rana, Yogesh K. Dwivedia, Kuttimani Tamilmani, Arpan Kumar Kar, Yi Wang, Yanqing Duan, Santosh Misra, Sujeet Kumar Sharma, Qixin Chen, e Chongqing Kang são os 10 principais autores em termos de acoplamento bibliográfico.

(b) Países-chaves: Os países mais produtivos para o domínio de ML em termos de número total de publicações (NP) são EUA, Índia, China e Reino Unido, seguido pela Alemanha e Canada conforme ilustrado na Tabela 3. Para citações por publicação, Polonia, Egito e Singapura são os três principais países com a métrica C/P mais elevada, enquanto China, EUA e Reino Unido aparecem com o maior número de citações totais TC

(c) Principais universidades: As 3 universidades mais produtivas são *University of Michigan*, *The Pennsylvania State University* e *King Saud University* conforme demostrado na Tabela 4. Em termos de colaboração de pesquisa, o *Indian Institute of Technology*, *Swansea University* e *Copenhagen Business School* estão engajados em uma ampla colaboração de pesquisa no domínio de ML.

(d) Principais periódicos: Os 10 periódicos com maior número de publicações e citações são: *IEEE Access*, *International Journal of Information Management*, *Journal of Big Data*, *International Journal of Production Research*, *Annals Of Operations Research*, *Sustainability (Switzerland)*, *Decision Support Systems*, *Journal of Business Research*, *Computers and Industrial Engineering* e *Industrial Management and Data Systems.* Baseado em acoplamento bibliográfico, os 10 principais periódicos são: *IEEE Access*, *International Journal of Information Management*, *International Journal of Production*, *Research Annals of Operations Research*, *Sustainability (Switzerland)*, *Industrial Management and Data Systems*, *British Journal of Management*, *Journal of Business Research*, *Journal of Enterprise Information Management* e *International Journal of Production Economics.*

(e) Áreas de conhecimento: Na Figura 4, é evidenciado que as áreas de conhecimento com maior número de publicações são: *"Computer Science", "Engineering", "Decision Sciences", "Social Sciences", "Business, Management and Accounting"* e *"Mathematics"*.

(cc) BY-NC-SA

(f) Análise da rede de cocitação por Autor: É inferida a partir da Figura 6, no qual (DWIVEDI *et al*, 2021) e (WANG *et al*, 2019) são altamente citados, estes autores são responsáveis pela ligação entre as duas redes de cocitação. Do lado esquerdo é evidenciado que os autores Sharma S.K., Misra S, Duan Y. e Kar A.K, citam com frequência o Dwivedi Y.K. e do lado direito Chen Q., Kang C. e Choi T.-M. citam com frequência o Wang Y.

(g) Análise de prestígio: Os três artigos de maior prestígio são (LIU Y. *et al*, 2020), (LI Z. *et al*, 2020) e (WANG Y. *et al*, 2019). No entanto o artigo de (CHEN M. *et al*, 2014) recebeu o maior número de citações.

Para a segunda questão de pesquisa, ou seja, a análise temática da evolução do ML, foi realizado um agrupamento de coocorrências de palavras-chaves, com os seguintes resultados:

(a) Principais coocorrências de palavras-chaves: "*Artificial Intelligence"* (AI), "*Learning Systems"* e "*Deep Learning"*, estão frequentemente correlacionadas, enquanto o termo "*Big Data*" frequentemente ocorre conjuntamente com "*Data Analytics*", "*Internet of Things*", "*Advanced Analytics*". Os termos "*Data Mining*", "*Predictive Analytics*", "*Forecasting*", "*Commerce*", e "*Sales*" são frequentemente correlacionados. Os termos "*Decision Making*", "*Big Data Analytics*", "*Prediction*" e "*Supply Chain Management*" estão relacionados diretamente. Apesar do termo "*Machine Learning*" apresentar uma relação direta com "*Supply Chains*", é observado no VOSViewer, que ML tem uma forte coocorrência com todos os demais grupos: "*Big Data*", "IA", "*Decision Making*", "*Data Mining*" e "*Learning Systems*".

(b) Informações dos *Clusters*: A partir da Figura 15 e da Tabela 7 os temas emergentes foram agrupados em cinco *clusters* bibliográficos, sendo o *Cluster* 1 (*Machine Learning Techniques, Decision Makes, Sales* e *Forecasting*), que apresenta maior densidade, seguido pelo *Cluster* 2 (BI e BDA), *Cluster* 3 (IoT, *Data Handling* e *Cloud Computing*), *Cluster* 4 (*Activity Recognition, Health, Medicine, Service Quality, Therapy, Heart Failure, Classification* e *Risk*) e o *Cluster* 5 (*Data Acquisition* e *Predictive Analytics*), que apresenta artigos diversos e possui baixa exploração em análises preditivas.

(c) Análise da nuvem de palavras por título e por resumo: *Big*, *Learn*, *Machine*, *Prediction*, *Analysis*, *Intelligence* e *Retail* são palavras frequentes nos títulos das publicações, enquanto *Big*, *Learn*, *Prediction*, *Algorithm*, *Model*, *Method*, *Analysis* e *Technique* são palavras frequentes nos resumos o que corrobora os temas cobertos pelos periódicos.

Para a terceira questão de pesquisa, ou seja, a identificação de futuras áreas de pesquisa, teóricas e/ou práticas, os resultados da análise de *cluster* e da análise de cocitação evidenciaram que DBA, ML e *Deep Learning* estão essencialmente ligados para resolver problemas de previsão de negócios e tomadas de decisão e áreas de aplicação como mercado de ações, marketing e gestão na cadeia de suprimentos. O papel da computação em nuvem e da IoT também são citados para servir como infraestrutura e gerar uma grande quantidade de dados a partir de sensores e atuadores.

CONCLUSÃO

O presente estudo apresentou uma análise bibliométrica sobre ML, considerando artigos de periódicos da *Scopus* e *Web Of Science* no período de 2013 a maio/2023.

Em termos de contribuições teóricas, os resultados alcançados podem auxiliar futuros pesquisadores a identificar temas emergentes para pesquisa e potenciais colaborações. Em primeiro lugar, o estudo examinou o foco da atual utilização do ML. O foco ilustrou as principais contribuições em termos de autores, universidades, periódicos e países para o domínio do ML. Em segundo lugar, destacou as principais áreas temáticas, agrupando-as bibliograficamente em cinco *clusters. S*obre o escopo para pesquisas futuras, observou-se que em estudos anteriores, como Batistic e Van (2019), foi adotado o mesmo protocolo bibliométrico para estudar o impacto das técnicas de BDA nas empresas. Este estudo estende a pesquisa existente para estreitar o foco para aplicações preditivas de grandes bases de dados em um contexto de ML. A literatura existente sobre análise de ML também foi detalhada, com um mapa de evolução temática indicando os temas emergentes. Assim, esses temas fornecem direcionamento para futuras pesquisas no domínio do ML.

Com relação às contribuições práticas, o estudo forneceu uma visão dos diferentes temas publicados nos últimos dez anos no domínio do ML. Esta evolução temática evidencia que ML está se tornando um domínio procurado por pesquisadores e profissionais de sistema

BY-NC-SA

de informação. Hoje, os dados são novas comodities e esta pesquisa pode auxiliar empresas que desejam investir na adoção de técnicas de ML para obter vantagem competitiva por meio de um diagnóstico mais assertivo. Equipes de pesquisa e desenvolvimento podem adotar este protocolo bibliométrico com pequenos ajustes na *string* de pesquisa, podendo aprofundar futuras pesquisas, recuperando documentos relevantes como pontos de verificação de referência para outras abordagens relacionadas ao *Big Data* e ML.

Embora o estudo revele algumas descobertas interessantes e forneça *insights* úteis, também existem algumas limitações. Primeiro, a amostra de dados foi limitada aos bancos de dados *Scopus* e *Web of Science* devido a disponibilidade de acesso para extração de artigos relevantes dos últimos dez anos. Em segundo lugar, uma combinação específica de palavras-chave foi utilizada para análise bibliométrica, a qual pode ser ajustada para derivar diferentes percepções. Além disso, o período para a extração pode ser variado para revelar diferentes tendências de publicações e citações.

A necessidade de análise preditiva é encontrada não só no setor corporativo para diagnóstico, mas também como uma área de pesquisa emergente. O principal motivador para a pesquisa neste campo é a necessidade de desenvolver ferramentas altamente precisas com alta capacidade de previsão recursos obtidos em diferentes segmentos, como bancos e serviços financeiros, marketing, cadeia de suprimentos, gestão de pessoas e predição de vendas.

CONTRIBUIÇÃO DOS AUTORES

Contribuição	Martins, E.	Galegale, N. V.
Contextualização	60%	40%
Metodologia	40%	60%
Software	60%	40%
Validação	60%	40%
Análise formal	60%	40%
Investigação	60%	40%
Recursos	60%	40%
Curadoria de dados	60%	40%
Original	60%	40%
Revisão e edição	60%	40%
Visualização	60%	40%
Supervisão	40%	60%
Administração do projeto	40%	60%
Aquisição de financiamento	---	---

REFERÊNCIAS

Ahani A., Nilashi M., Ibrahim O., Sanzogni L., Weaven S., (2019) - Market segmentation and travel choice prediction in Spa hotels through TripAdvisors online reviews https://doi.org/10.1016/j.ijhm.2019.01.003

Ahmadi H., Arji G., Shahmoradi L., Safdari R., Nilashi M., Alizadeh M., (2019) - The application of internet of things in healthcare a systematic literature review and classification. https://doi.org/10.1007/s10209-018-0618-4

Ali M.A.M., Bashar A., Rabbani M.R., Abdulla Y., (2020) - Transforming Business Decision Making with Internet of Things IoT and Machine Learning ML. https://doi.org/10.1109/dasa51403.2020.9317174

Alonso-Betanzos A., Bolon-Canedo V., (2018) - Big-Data Analysis, Cluster Analysis, and Machine-Learning Approaches. https://doi.org/10.1007/978-3-319-77932-4_37

Antonopoulos I., Robu V., Couraud B., Et Al (2020) - Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. https://doi.org/10.1016/j.rser.2020.109899

Athmaja S.; Hanumanthappa M., Kavitha V., (2017) - A Survey of Machine Learning Algorithms for Big Data Analytics. https://doi.org/10.1109/iciiecs.2017.8276028

Baryannis G., Validi S., Dani S., Antoniou G., (2019) - Supply chain risk management and artificial intelligence state of the art and future research directions. https://doi.org/10.1080/00207543.2018.1530476

Batistic S., Van D.L.P., (2019) - History Evolution and Future of Big Data and Analytics A Bibliometric Analysis of Its Relationship to Performance in Organizations. https://doi.org/10.1111/1467-8551.12340

Bhavnani S.P., Parakh K., Atreja A., Et Al (2017) - 2017 Roadmap for Innovation - ACC Health Policy Statement on Healthcare Transformation in the Era of Digital Health,

(cc) BY-NC-SA

Big Data and Precision Health. https://doi.org/10.1016/j.jacc.2017.10.018

Bilgic E., Cakir O., Kantardzic M., Duan Y., Cao G., (2021) - Retail analytics: store segmentation using Rule-Based Purchasing behavior analysis. https://doi.org/10.1080/09593969.2021.1915847

Böse J.-H., Flunkert V., Gasthaus J., Et Al (2017) - Probabilistic demand forecasting at scale. https://doi.org/10.14778/3137765.3137775

Bui T.D., Tsai F.M., Tseng M.L., Tan R.R., Yu K.D.S., Lim M.K., (2021) - Sustainable supply chain management towards disruption and organizational ambidexterity A data driven analysis. https://doi.org/10.1016/j.spc.2020.09.017

Calatayud A., Mangan J., Christopher M., (2019) - The self-thinking supply chain - Supply Chain Management - Emerald Group Holdings Ltd. - United Kingdom. https://doi.org/10.1108/SCM-03-2018-0136

Cerruela García G., Luque Ruiz I., Gómez-Nieto M., (2016) - State of the art trends and future of bluetooth low energy near field communication and visible light communication in the development of smart cities - Sensors (Switzerland) - MDPI AG – Spain. https://doi.org/10.3390/s16111968

Chandra S. E Verma S., (2021) - Big Data and Sustainable Consumption A Review and Research Agenda – Vision - Sage Publications India Pvt. Ltd – India. https://doi.org/10.1177/09722629211022520

Chang, P.C., Liu, C.H., And Fan, C.Y. (2009) - Data clustering and fuzzy neural network for sales forecasting: A case study in printed circuit board industry. https://doi.org/10.1016/j.knosys.2009.02.005

Chen M., Mao S., Liu Y., (2014) - Big data: A survey - Mobile Networks and Applications. https://doi.org/10.1007/s11036-013-0489-0

Chen M., Hao Y.X., Hwang K., Wang L., Wang L., (2017) - Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. https://doi.org/10.1109/access.2017.2694446

Choi T.-M., Wallace S.W., Wang Y., (2018) - Big Data Analytics in Operations Management. https://doi.org/10.1111/poms.12838

Dinov I.D., Heavner B., Tang M., et al (2016) - Predictive Big Data Analytics A Study of Parkinsons Disease Using Large Complex Heterogeneous Incongruent MultiSource and Incomplete Observations - Plos One - Public Library Science - United States. https://doi.org/10.1371/journal.pone.0157077

Duan Y., Edwards J.S., Dwivedi Y.K., (2019) - Artificial Intelligence for Decision Making In The Era Of Big Data Evolution Challenges And Research Agenda. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Dwivedi Y.K., Hughes L., Ismagilova E., et al (2021) - Artificial Intelligence AI Multidisciplinary perspectives on emerging challenges opportunities and agenda for research practice and policy. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

George G., Osinga E., Lavie D., Scott B., (2016) - Big data and data science methods for management research. https://doi.org/10.5465/amj.2016.4005

Gill S. S., Tuli S., Xu M., et al, (2019) - Transformative effects of IoT Blockchain and Artificial Intelligence on cloud computing Evolution vision trends and open challenges. https://doi.org/10.1016/j.iot.2019.100118

Gupta N., Ahuja N., Malhotra S., Bala A., Kaur G., (2017) - Intelligent heart disease prediction in cloud environment through ensembling - Expert Systems – Wiley – India. https://doi.org/10.1111/exsy.12207

Hashimoto D.A., Rosman G., Rus D., Meireles O.R., (2018) - Artificial Intelligence in

BY-NC-SA

Surgery Promises and Perils - Annals of Surgery - Lippincott Williams & Wilkins - United States. http://dx.doi.org/10.1097/SLA.0000000000002693

Hassija V., Chamola V., Saxena V., Jain D., Goyal P., Sikdar B., (2019) - A Survey on IoT Security Application Areas Security Threats and Solution Architectures. https://doi.org/10.1109/access.2019.2924045

Hu H., Wen Y., Chua T-S., Li X., (2014) - Toward scalable systems for big data analytics A technology tutorial - IEEE Access - Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/access.2014.2332453

Kitchens B., Dobolyi D., Li J., Abbasi A., (2018) - Advanced Customer Analytics Strategic Value Through Integration of RelationshipOriented Big Data. https://doi.org/10.1080/07421222.2018.1451957

Kou G., Chao X., Peng Y., Alsaadi F.E., Herrera-Viedma E., (2019) - Machine learning methods for systemic risk analysis in financial sectors. https://doi.org/10.3846/tede.2019.8740

Kousis A. E Tjortjis C., (2021) - Data mining algorithms for smart cities A bibliometric analysis - Algorithms - MDPI AG – Greece. https://doi.org/10.3390/a14080242

Lichman, M. (2013) - UCI Machine Learning Repository. Disponível em: https://archive.ics.uci.edu/ml/datasets/wine

Johnson A.E.W., Ghassemi M.M., Nemati S., Niehaus K.E., Clifton D.A., Clifford G.D., (2016) - Machine Learning and Decision Support in Critical Care. https://doi.org/10.1109/jproc.2015.2501978

Jordan, M.I. E Mitchell, T.M. (2015) - Machine learning: Trends perspectives and prospects. Science, 349:255–260. https://doi.org/10.1126/science.aaa8415

(cc) BY-NC-SA

Ke J., Zheng H., Yang H., Chen X. (2017) - Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. https://doi.org/10.1016/j.trc.2017.10.016

Krawczyk B., (2016) - Learning from imbalanced data open challenges and future directions - Progress in Artificial Intelligence – Springernature – Poland. https://doi.org/10.1007/s13748-016-0094-0

L'heureux A., Grolinger K., Elyamany H.F., Capretz M.A.M., (2017) - Machine Learning with Big Data Challenges and Approaches - IEEE Access - Institute of Electrical and Electronics https://doi.org/10.1109/access.2017.2696365

Levy, Y.; Ellis, T.J. A system approach to conduct an effective literature review in support of information systems research. Informing Science Journal, v.9, p.181-212, 2006. https://doi.org/10.28945/479

Ma C., Zhang H.H., Wang X.F., (2014) - Machine learning for Big Data analytics in plants - Trends in Plant Science - Elsevier Science London – China. https://doi.org/10.1016/j.tplants.2014.08.004

Mishra D., Gunasekaran A., Papadopoulos T., Childe S.J., (2018) - Big Data and supply chain management a review and bibliometric analysis. https://doi.org/10.1007/s10479-016-2236-y

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Stewart, L. A. (2015) - Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews, 4(1). https://doi.org/10.1186/2046-4053-4-1

Moreira Mwl., Rodrigues Jjpc., Kumar N., Saleem K., Illin Iv, (2019) - Postpartum depression prediction through pregnancy data analysis for emotionaware smart systems updates. https://doi.org/10.1016/j.inffus.2018.07.001

(cc) BY-NC-SA

Nguyen H.D., Tran K.P., Thomassey S., Hamad M., (2021) - Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. https://doi.org/10.1016/j.ijinfomgt.2020.102282

Nguyen T., Zhou L., Spiegler V., Ieromonachou P., Lin Y., (2018) - Big data analytics in supply chain management A stateoftheart literature review. https://doi.org/10.1016/j.cor.2017.07.004

Qian T.Q., Zhu S.J., Hoshida Y., (2019) - Use of big data in drug development for precision medicine an update. https://doi.org/10.1080/23808993.2019.1617632

Razavian N., Blecker S., Schmidt A.M., Smith-Mclallen A., Nigam S., Sontag D., (2015) - PopulationLevel Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors https://doi.org/10.1089/big.2015.0020

Sahoo S., (2021) - Big data analytics in manufacturing a bibliometric analysis of research in the field of business management. https://doi.org/10.1080/00207543.2021.1919333

Sharma, R., Kamble, S.S., Gunasekaran, A., Kumar, V., Kumar, A., (2020) - A systematic literature review on machine learning applications for sustainable agriculture supply chain performance - Computers & Operations Research - Pergamon-Elsevier Science Ltd – England. https://doi.org/10.1016/j.cor.2020.104926

Shokouhyar S., Shokoohyar S., Sobhani A., Gorizi A.J., (2021) - Shared mobility in post-COVID era: New challenges and opportunities - Sustainable Cities and Society - Elsevier Ltd https://doi.org/10.1016/j.scs.2021.102714

Silver, D., Huang, A. E Guez, A. (2016) - Mastering the game of go with deep neural networks and tree search - Nature, 529:484–489. https://doi.org/10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K. E Antonoglou, I. (2017) - Mastering the game of go

without human knowledge - Nature, 550:354–359. https://doi.org/10.1038/nature24270

Raschka, S. E Mirjalili, V. (2017) - Python Machine Learning, 2nd Ed.- Packt Publishing, Birmingham, UK, 2 edition.

Trieu V.-H., (2017) - Getting value from Business Intelligence systems A review and research agenda - Decision Support Systems - Elsevier B.V. – Australia. https://doi.org/10.1016/j.dss.2016.09.019

Tzeng G.-H., Shen K.-Y., (2017) - New concepts and trends of hybrid multiple criteria decision making - ISBN 9780367573133

Wanasinghe T.R., Wroblewski L., Petersen B.K., et al (2020) - Digital Twin for the Oil and Gas Industry Overview Research Trends Opportunities and Challenges. https://doi.org/10.1109/access.2020.2998723

Wang D., Liu X., Wang, M., (2013) - A dt-svm strategy for stock futures prediction with big data - IEEE 16th International Conference on Computational Science and Engineering. https://doi.org/10.1109/cse.2013.147

Wang J.L., Zhao P.L., Hoi S.C.H., Jin R., (2014) - Online Feature Selection and Its Applications - IEEE Transactions on Knowledge and Data Engineering - IEEE Computer Soc - United States. https://doi.org/10.1109/tkde.2013.32

Wang W., Gao J.Y., Zhang M.H., et al (2018) - Rafiki Machine Learning as an Analytics Service System - Proceedings of The Vldb Endowment - Assoc Computing Machinery – China. https://doi.org/10.48550/arXiv.1804.06087

Wang Y., Chen Q., Hong T., Kang C., (2019) - Review of Smart Meter Data Analytics Applications Methodologies and Challenges. https://doi.org/10.1109/tsg.2018.2818167

(cc) BY-NC-SA

Xu J., Huang E., Chen C.-H., Lee L.H., (2015) - Simulation optimization A review and exploration in the new era of cloud computing and big data. https://doi.org/10.1142/s0217595915500190

BY-NC-SA