International Journal of Innovation

e-ISSN: 2318-9975
https://doi.org/10.5585/2023.24056
Received: 13 mar. 2023 - **Approved:** 18 June 2023

Evaluation Process: Double Blind Review
Editor in Chief: Priscila Rezende da Costa
Coeditor: Isabel Cristina Scafuto
Scientific Editor: Vânia Maria Jorge Nassif
Assistant Editor: Angelica Pigola
***Section**: Article*

Check for updates

MACHINE LEARNING: A BIBLIOMETRIC ANALYSIS

Emerson Martins [1]
Napoleão Verardi Galegale [2]

Cite as – American Psychological Association (APA)

Martins, E., & Galegale, N. V. (2023, Sept./Dec.). Machine learning: a bibliometric analysis. *International Journal of Innovation - IJI*, São Paulo, *11*(3), 1-37, e24056. https://doi.org/10.5585/2023.24056

Abstract

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

[1] Master in Management and Technology in Production Systems (CEETEPS) and Researcher at the IT Strategic Management Research Group (CEETEPS/CNPq). CEETEPS – State Center for Technological Education Paula Souza / São Paulo (SP) – Brazil. emerson.martins@cpspos.sp.gov.br
[2] PhD in Controllership and Accounting (FEA/USP), Master in Production Engineering (POLI/USP), Professor and Researcher at UPEP/CEETEPS and FEA/PUC-SP, leader of the IT Strategic Management Research Group (CEETEPS/CNPq) and Business Consultant. CEETEPS – State Center for Technological Education Paula Souza / São Paulo (SP) – Brazil. napoleao.galegale@cpspos.sp.gov.br

MACHINE LEARNING: UMA ANÁLISE BIBLIOMÉTRICA

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.

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.

1 INTRODUCTION

In recent years, several technological advances have occurred, such as the emergence of Big Data, along with the two benefits accruing from data science to society (Chen *et al.*, 2014). As well as an adequate information security policy (Galegale *et al.*, 2017), human capital, and machines, data will emerge as an essential resource to generate prosperity in society. Although data processing has begun with traditional methods of extracting, transforming, and treating data through business management systems, according to Hu *et al.* (2014), these techniques are not scaled, especially due to the enormous increase in data volume. Big Data, therefore, evolves as companies realize that, to obtain competitive advantage, investment in data analysis is equally important, along with products, services, processes, and technology (Mishra *et al.*, 2018).

This need for evolution also occurs due to the occurrence of unstructured data, which cannot be processed directly with traditional tools, or they require special data processing and information processing techniques, such as Natural Language Processing (NLP) and Machine Learning (ML). Nowadays, the processing of information for the generation of knowledge has become vital for decision-makers, particularly in some important areas such as the anticipation of sales of products or services, where external variables such as time or the global economy can affect consumer decisions. das pessoas (KRAWCZYCK, 2016). Also, this revolution demands new integration of the Internet of Things (IoT), Blockchain, and Big Data Analysis (BDA) (GILL *et al.*, 2019).

Big Data and ML have been widely used by organizations due to the growing needs of businesses and services to face global challenges in obtaining competitive advantage. This new model is multiplied on demand by analytical tools to solve complex business problems in

various domains, including financial markets, marketing, health, supply chain, and sales prediction (MARTINS & GALEGALE, 2023). From this scenario, Business Analytics emerges, which is the application of techniques using Big Data analysis tools known as Data Science for decision-making (CHEN *et al.*, 2014).

Research in the ML domain has made a significant leap in recent years (ATHMAJA *et al.*, 2017). Consequently, several studies will carry out bibliometric research to summarize the existing knowledge in the ML field. For example, Antonopoulos *et al.* (2020) review perspectives in the renewable energy sector. Likewise, Sharma *et al.* (2020) review the application of ML analysis in the agricultural context.

Despite these important attempts to synthesize existing literature, it has been observed that the literature on the emergence of the most recent technologies, such as artificial intelligence (AI), ML, and Big Data, seems fragmented (CHANDRA & VERMA, 2021). The different aspects of ML and its scope for future research were not considered. There is an evident need for research to provide a comprehensive understanding of the past, present, and future of research in respect of the use of ML. Therefore, this research considers this gap in bibliometric studies and extends the bibliometric survey of the impact of the use of ML in organizations. This study considers three research questions to address the research gaps mentioned above: (1) what is the focus of this research on ML?; (2) What are the main topics and domains in ML and its evolution?; and (3) what scope for future research, whether from an academic or market point of view?

This article provides a bibliographic overview in line with Batistic and Van (2019), as well as Sahoo (2021). In the same way, the article is also a generalization of bibliometric studies contained in the literature, such as the analysis of the supply chain carried out by Mishra *et al.* (2018), Smart Cities carried out by Kousis and Tjortjis (2021), and a bibliometric analysis of the chains of sustainable supplies carried out by Bui *et al.* (2021).

2 METHOD

This article presents an in-depth analysis of the citation and publication of trends in the ML analysis between 2013 and May 2023. This period was compiled based on the availability of data in the Web of Science and Scopus databases. The authors, institutions, countries, and significant newspapers are presented. The main topics discussed are highlighted and the articles are classified into five bibliographic groups based on the keywords that occur most

frequently. This approach illustrates the main themes present in the articles examined, as well as the relationship between the authors and the keywords. The topics that occur most frequently are indicated by the analysis of the number of words and the analysis of the structure of the citations is carried out by the group to highlight the emerging themes.

Bibliometric analysis is used to highlight key authors, institutions/universities, and countries in terms of their contributions to the respective field. Collaboration patterns between authors, institutions, and countries are also analyzed (BATISTIC & VAN, 2019).

Additionally, the PRISMA-P protocol was used, whose objective is to support researchers in improving the reporting of systematic reviews and meta-analyses, filtering the number of publications with greater relevance to the researched topic (MOHER *et al.*, 2015).

As mentioned in the introduction of this article, the absence of studies in the literature that consolidate articles related to ML and Big Data with an emphasis on predictive algorithms was observed. In this way, this research carries out a literature review, separating key themes into thematic groupings, and providing an agenda for future research with less subjectivity. Bibliographic records were accessed from the Web Of Science and Scopus databases, in various disciplines, such as computer science, engineering, decision science, social sciences, business management, and mathematics. Several authors (SAHOO, 2021; MISHRA *et al.*, 2018; KOUSIS & TJORTJIS, 2021; BUI *et al.*, 2021) suggest that bibliometric studies are impartial synthesizers of literary content.

For Levy and Ellis (2006), knowing the current state of the body of knowledge on a given topic is the first step in a research project. Thus, according to the authors, a bibliometric study is useful for:

- Help the researcher in sizing and understanding the body of knowledge relating to a given subject, including identifying research that has already been carried out, what remains to be researched, and what the gaps are;
- Provide theoretical basis for the proposed study;
- Present the appropriate justifications for conducting the study, and what is the original contribution to the body of knowledge and/or theory; and
- Contribuir para melhor definir e estruturar o método de pesquisa, objetivos e questões para o estudo proposto.

Levy and Ellis (2006) describe bibliometric study as a process. The authors adopt the definition of a process as a "sequence of steps and activities". To achieve these results, they define three main phases: Entry; Processing; and Output. The "input" phase contains the preliminary information that will be processed, adopted as Phase 1 in this research. In the "processing" phase, a protocol must be applied that filters the number of publications according to the research topic, called Phase 2 in this research. Finally, in the "output" phase, reports will be generated summarizing the results, identified as Phase 3 in this research. These three phases are detailed in Figure 1.

Figure 1

Research Design

Source: Results of this research.

2.1 Plan the review process - Phase 1

This was done by formulating the research questions and collecting data from Web of

Science and Scopus. The search revealed publications in renowned journals, providing valuable information regarding Big Data and ML in several countries. Therefore, a search was carried out in the databases to retrieve publication records using the search constants “Machine Learning”, “Algorithm” and “Big Data Analytics”. Initially, some keyword combinations were added to the search, such as "Forecasting", "Prediction", "Predictive", and “Retail”, but the bibliometric records were limited to just 38 publications. This way, the term “Algorithm” was excluded from the search string, keeping only "Machine Learning" AND "Big Data Analytics" AND "Prediction" AND "Retail". With this method, it was possible to extract 964 publications from the Scopus database and 421 publications from the Web of Science, with publications from 2013 until May 2023, when this research was carried out.

The selected keywords were obtained from an initial analysis of the articles using the Gephi software using the “Total Link Strength” attribute, which indicates the co-occurrences of keywords that appear most frequently, as described in subsection 3.2.4.

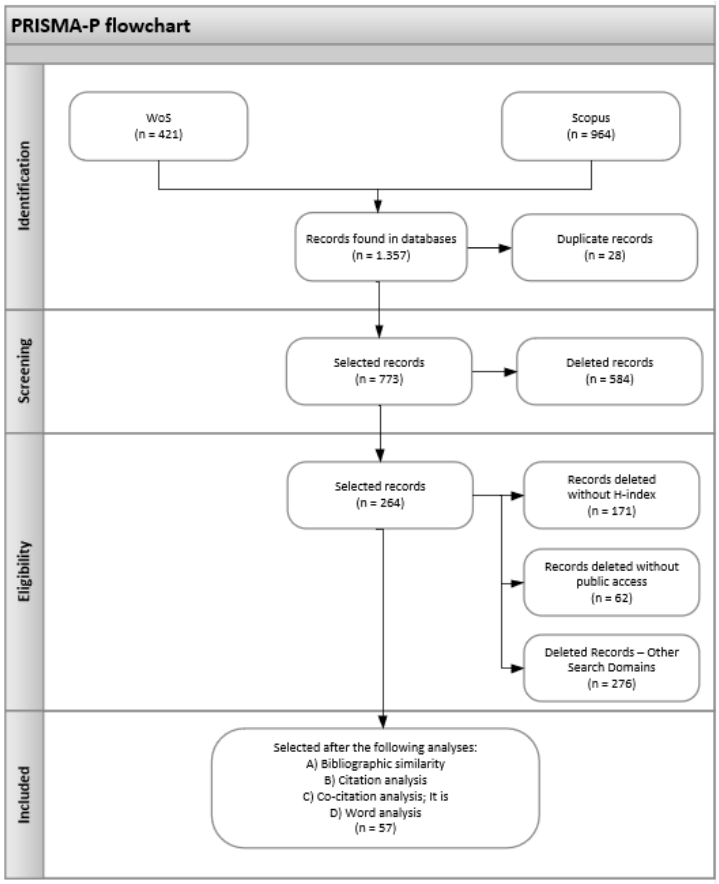
Through a descriptive analysis of the corpus of articles, as well as through a quantitative analysis of publications, citation trends for the period 2013 to May 2023 were assessed. This was followed by an analysis of the main prolific authors and the main countries with publications on Big Data and ML.

2.2 Implementation of the bibliometric protocol - Phase 2

Figure 2 presents the flowchart of the selection process of scientific publications in each of the four stages provided for by the PRISMA-P protocol: identification, screening, eligibility, and documents included for analysis.

Figure 2

PRISMA-P flowchart



Source: Results of this research.

In the screening stage, 28 duplicate publications, 53 books, 81 book chapters, 272 conference articles, 172 reviews, and 6 editorial materials were removed, totaling 584 excluded records. In this way, 773 records were selected for the next stage.

In the eligibility stage, 171 articles without H-Index were excluded, 62 articles whose access was not public, and 276 articles whose research was related to one of the following domains: Medicine, Engineering and Architecture, Education, Psychology, Agriculture and Biosciences, Arts and Humanities, Biochemistry, Geosciences, and Logistics.

264 publications were selected to proceed to the last stage, in which the following techniques were applied:

(a) Bibliographic similarity. It helps to identify a set of publications with the greatest bibliographic similarity, measured in terms of the number of shared references. This similarity reflects the degree of similarity in the research and a possible

similarity in future research directions. In this article, bibliographic similarity is performed for authors, institutions/universities, and journals, to extract insights as illustrated in subsection 3.2.

(b) Citation analysis. Citation trend analysis is performed in subsection 3.1, to evaluate the contribution in terms of how many documents are referring to the article and/or citing it. This technique identifies the main authors, institutions, and countries in terms of citations. Furthermore, the PageRank metric is also used to measure the prestige of the article in renowned journals.

(c) Co-citation analysis. Co-citation analysis is performed in subsection 3.2.2, to identify similarities between publication titles and group them into different themes/topics based on their conceptual structure. The co-citation analysis is complemented with a word analysis to identify keyword co-occurrences.

(d) Word analysis. It is conducted to visualize the frequency of occurrence of a given author and index keywords on a research topic in subsection 3.2.4. It is also necessary to analyze the evolution of thematic change over time to identify emerging topics and those that are saturated.

After the eliminatory analyzes above, 57 publications were kept for qualitative analysis.

2.3 Reporting findings – Phase 3

It is structured in terms of descriptive and bibliometric analyses. The descriptive analysis includes the total number of publications and citations. The information was extracted using the "biblioAnalysis" library, a function contained in the bibliometric package of the R software. This phase comprises the contribution and extension of research collaboration, considering several authors and countries. An analysis considering the most cited authors is also conducted to understand the main research of the most cited authors.

A bibliometric analysis of publications was carried out using the R software, through RStudio 2022.07.2 Build 576, to identify bibliographic links (between authors), co-citations and co-occurrences using the "bibliometrix" v4.0.1 package in R, which contains the function default "biblioNetwork". The Gephi v0.10 tool was used to perform prestige analysis. VOSviewer v1.6.17 was used to map keyword co-occurrences, university collaboration, and

journal co-citation analysis. Emerging themes and grouping of titles were identified using word cloud analysis using Iramuteq v0.7 alpha 2. Conceptual structure analysis was also carried out using “bibliometrix”.

3 RESULTS

The results of the descriptive and bibliometric analysis are demonstrated in this section.

3.1 Descriptive and citation analysis of publications by periodicals

A descriptive analysis of the “BibTex” exported files from 2013 to May 2023 was conducted and displayed in Table 1, using the “summary” function of the “bibliometrix” library. After consolidating publications from the Scopus database (964 documents) and WoS (421 documents), 28 duplicate publications were eliminated, resulting in 1,357 publications.

Of the 1,357 publications, 773 are research articles, 53 are books, 81 book chapters, 272 conference articles, 172 reviews, and 6 editorial materials.

The frequency distribution of keywords per author is 3,727, which implies that these keywords are frequently used by authors in ML and BDA publications. The distribution of keywords extracted from journal articles in the domain is 4,407. The number of authors was 4,205, with 5,233 appearances, including single authorship and appearances by multiple authors. Of the 4,205 authors, 89 authors published articles with a single author, while the remaining 4,116 authors published articles with multiple authors, indicating a high degree of research collaboration in the published articles. The number of unique authorships of documents is only 92, while the remaining 1,265 are documents with multiple authorships. The number of documents per author, that is, the ratio of the total number of documents (1,357) to the total number of authors (4,205) is 0.323. The reciprocal ratio of this metric, the number of authors per document (4,205/1,357), is 3.10, while the number of co-authors per document is 3.86. The collaboration index, that is, the ratio of the total number of authors in multiple-authored documents to the number of multiple-authored documents (4,116/1,265) is 3.25, thus indicating that for a multiple-authored document, there are approximately three authors. This finding corroborates a robust index in the collaboration network.

Table 1

Summary of descriptive analysis of records

Description	Results
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

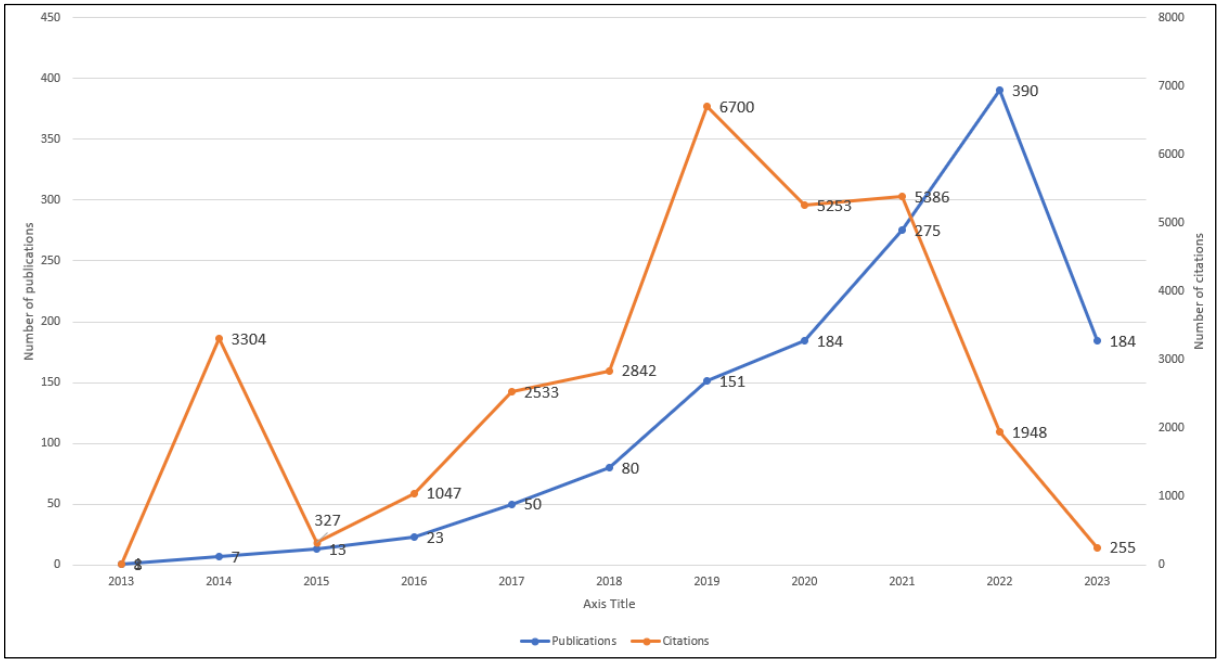
Source: Results of this research.

The average number of citations per document is 21.58, which implies that journal articles are cited an average of almost 22 times.

Through Figure 3, a perspective was obtained on the number of articles in the ML and BDA domain published over 10 years of study, identifying the publication trend (measured in terms of number of articles) and the citation trend for the period of 2013 to May 2023.

Figure 3

Trend of publications and citations

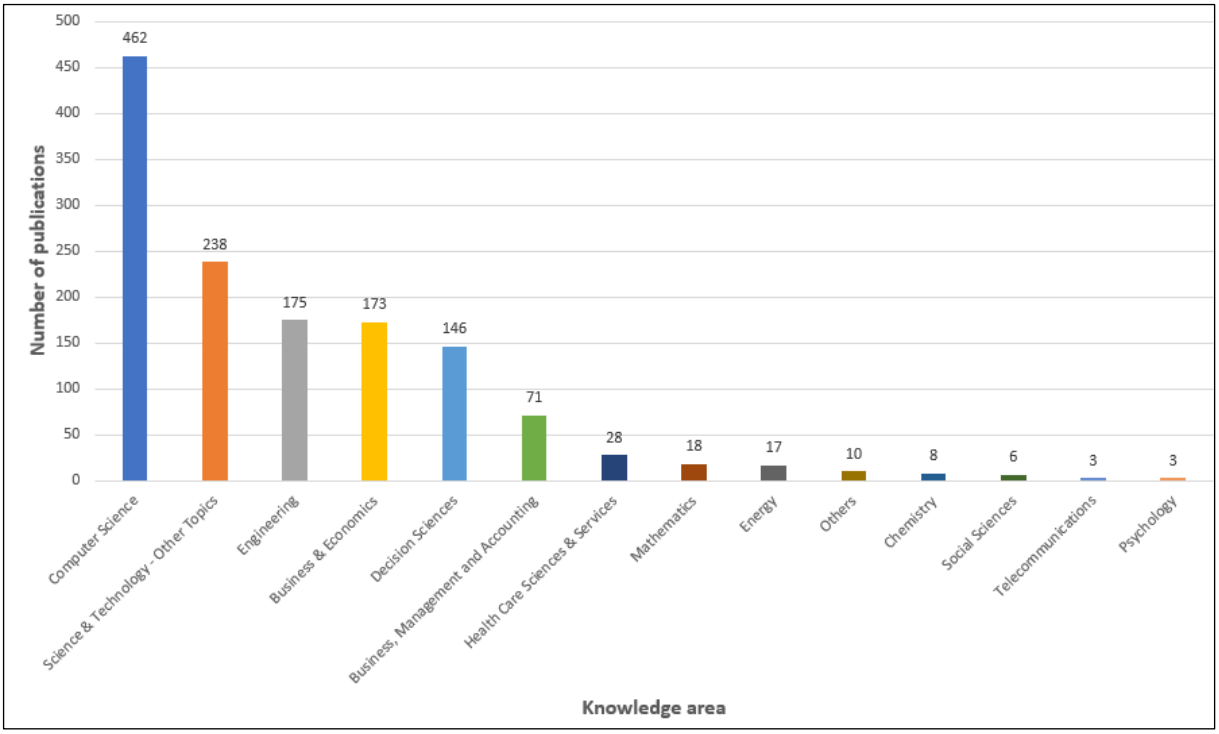


Source: Results of this research.

From Figure 4, Computer Science is considered the predominant domain with 34% of articles in this category, followed by Science and Technology (18%), Engineering and Business (13%), Decision Sciences (11%), Business Management (5%), and Health (2%). Other areas with a marginal contribution to ML and BDA include Chemistry and Psychology. These areas have a more interdisciplinary research scope.

Figure 4

Publications by area of knowledge



Source: Results of this research.

3.1.2 Analysis by author

Considering the high degree of collaboration, we identified the most cited authors in terms of the number of total publications (NP), as well as the number of citations (TC), and the number of citations per publication (C/P). As shown in Table 2, Yunhao Liu (abbreviated as Liu Y) from China was found to be a frequent household name with a high C/P value of 294.7, followed by Yogesh Kumar Dwivedi (Dwivedi Y.) from the United Kingdom, with a C/P value of 137.6. The next most productive authors in terms of C/P were Kar A. from Poland and Wang Y. from China. These authors are renowned scholars in the field of ML, contributing to the state of the art in research articles, and providing researchers and professionals with a great framework of knowledge. To analyze authors' productivity in terms of total citations (TC) over the analyzed period, the graph shown in Figure 5 was extracted from the 'Bibliometrix' library. The AuthorProdOverTime function calculates and plots authors' productivity in terms of the number of publications and the total number of citations per year.

The h-index is measured in terms of the number ‘h’ of publications with minimum ‘h’

times citations. The g-index indicates the number 'g' of articles with at least 'g^2' citations. The article's m-index is calculated with the ratio between the h-index and the number of years since the author's first publication was made. The 'h', 'g', and 'm' indices are presented as measures of citation and productivity, with Wang Y., Wang X., and Chen G. having the highest h-index values (6, 5, and 5) respectively, g-index (7, 5, and 5) and m-index (1.2; 0.625; and 0.8333333). Furthermore, the results show that for Yunhao Liu, 4 articles were cited 4 times, and the top 6 articles were cited at least 6^2 (36 times). Since Yunhao Liu has been active since 2015 (8 active years of publishing), his m-index is (4/8) = 0.5.

Table 2

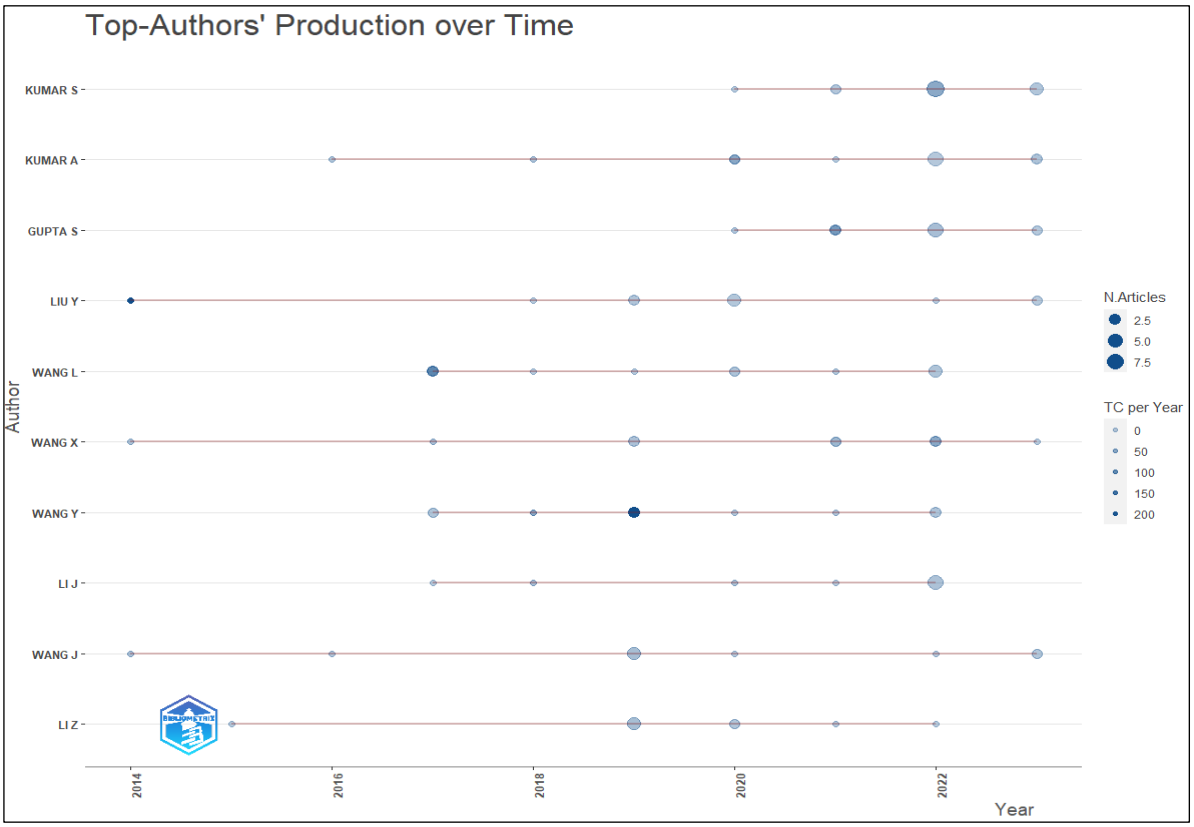
The ten most productive authors in publications and citations

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

Source: Results of this research.

Figure 5

AuthorProdOverTime function from the *Bibliometrix* library



Source: Results of this research.

3.1.3 Analysis by country

To identify the countries with the highest number of citations, the citations per publication (C/P) metric is used, which indicates that the United Kingdom, Canada, and China are the three countries that have the most citations with a C/P of 55.9, 42.2, and 32.7 respectively, as shown in Table 3. In terms of the number of publications (NP), the three main countries are India, China, and the United States, while Singapore and Korea are the least cited.

Table 3

The ten most productive countries in publications and citations

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

Source: Results of this research.

3.1.4 Analysis by universities

An analysis was performed by universities, shown in Table 4. The most productive universities are the University of Michigan, Pennsylvania State University, and King Saud University, in terms of numbers of publications (NP) and citations (TC). Contributions from the University of Hong Kong and the University of South Carolina are scarcer.

Table 4

Publications by universities

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

Source: Results of this research.

3.1.5 Analysis by journal

The top 10 journals in terms of number of publications (NP) and total citations (TC) are summarized in Table 5. IEEE Access and International Journal of Information Management appear with 34 and 12 publications respectively and with 1662 and 827 citations, demonstrating that such journals have great relevance in the domain of ML and BDA. Through a descriptive analysis in conjunction with a bibliometric and network analysis, the next section will address the identification of bibliographic coupling, co-citations, and topic co-occurrences.

Table 5

Publications by journal

Journal	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

Source: Results of this research.

3.2 Bibliometric and network analysis

A coupling analysis by author and a co-citation network was performed to analyze collaboration and mutual dependence of research citations between authors, universities, and journals. Emerging themes and relevant topics were identified using keyword co-occurrence analysis and bibliometric grouping of documents. Additionally, a cluster time series analysis is presented to recognize emerging themes with scope for future research. The results were validated by a graph of keyword co-occurrences, along with a word cloud of titles, abstracts, and themes. A co-citation network analysis was performed to identify the direction of new research based on previous articles with high numbers of citations.

3.2.1 Bibliometric author coupling

Bibliometric coupling between authors was developed to identify collaboration between them to analyze ML and BDA information, as shown in Table 6 and Figure 6. In terms of bibliographic coupling, Wang Y. and Dwivedi are in different clusters, however, it is through these two authors that the other groups can relate in terms of collaboration. At the same time, Rana N. P. and Tamilmani K. are in the same cluster and are active contributors.

Table 6

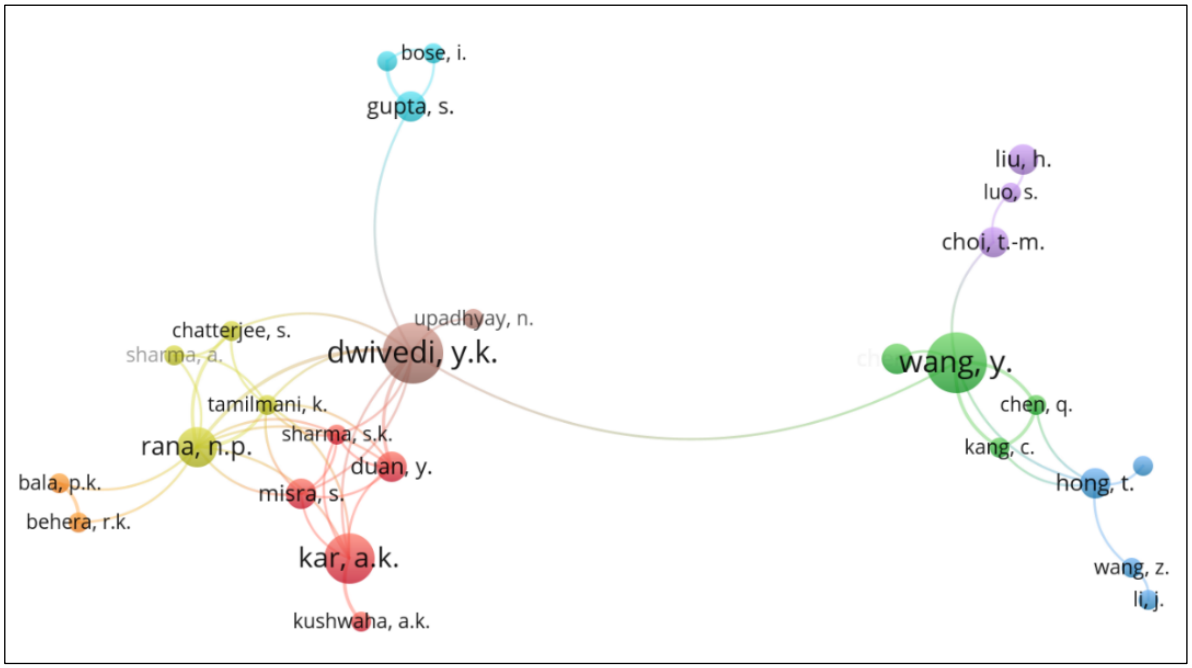
The main authors in terms of bibliographic coupling

Author	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

Source: Results of this research.

Figure 6

Coupling by author



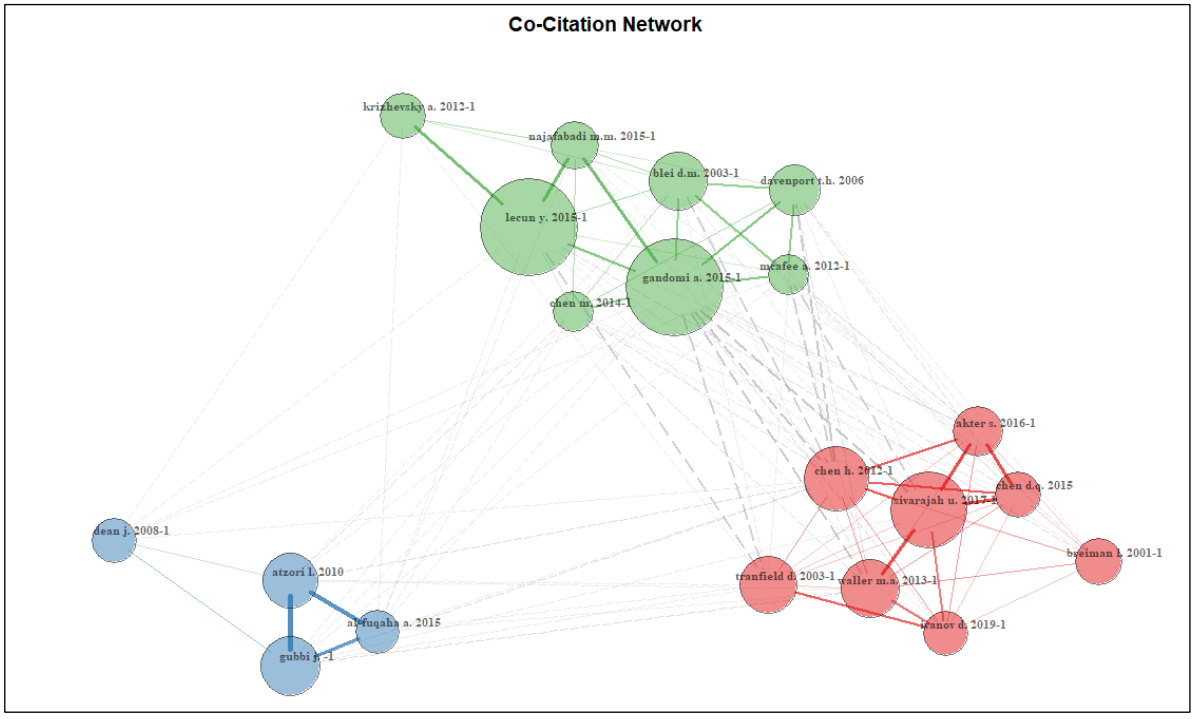
Source: Results of this research.

3.2.2 Co-citation network analysis

A co-citation network was created using the R software to analyze the citation strength among the main authors. In Figure 7, the number of citations per author is shown using circles, that is, the larger the size of the circle, the greater the citation strength. Additionally, the number next to the year indicates the strength of the quote. For example, Chen M. (2014) is denoted as “Chen M. 2014-1” due to its citation strength being equal to “Bleid D.M. 2003-1”. Regarding local citations, that is, the frequency of two authors being cited in the same article, they are indicated by solid lines. In contrast, global mutual citations between two different articles, when both are cited in a third article, are indicated by dashed lines. It appears that Gandomi (2015) and Sivarajah (2017) are frequently cited and are indicated by large green and red circles respectively. The solid lines indicate the authors' local citations, for example, Waller (2013) has been frequently cited by Sivarajah (2017) in the same article, as evidenced by the red solid line. Dotted lines indicate global citations, for example, Tranfield (2003) and Chen (2012) are often cited by the third Gandomi (2015), and therefore a dashed line is drawn from Tranfield and Chen to Gandomi, indicating global mutual citation.

Figure 7

Co-citation network



Source: Results of this research.

3.2.4 Keyword co-occurrence analysis

In Figure 8, keyword co-occurrences measured in terms of total link strength are considered high between “Machine Learning” and “Supply Chains” (violet circles), while the term “Big Data” frequently occurs with “Data Analytics ”, “Advanced Analytics”, and “Data Handling” (red circles), indicating that “Advanced Analytics” and “Data Handling” are emerging in “Big Data” research areas. The terms “Artificial Intelligence”, “Learning Systems”, and “Deep Learning” are themes with frequent co-occurrences (yellow circles). Furthermore, the terms “Big Data Analytics”, “Decision Making”, and “Supply Chain Management” are correlated (blue circles), while the terms “Data Mining”, “Forecasting”, “Sales”, “Commerce”, and “Predictive Analytics” are correlated (green circles). Keywords that occur frequently in a given topic are identified by circles that have the same color.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.

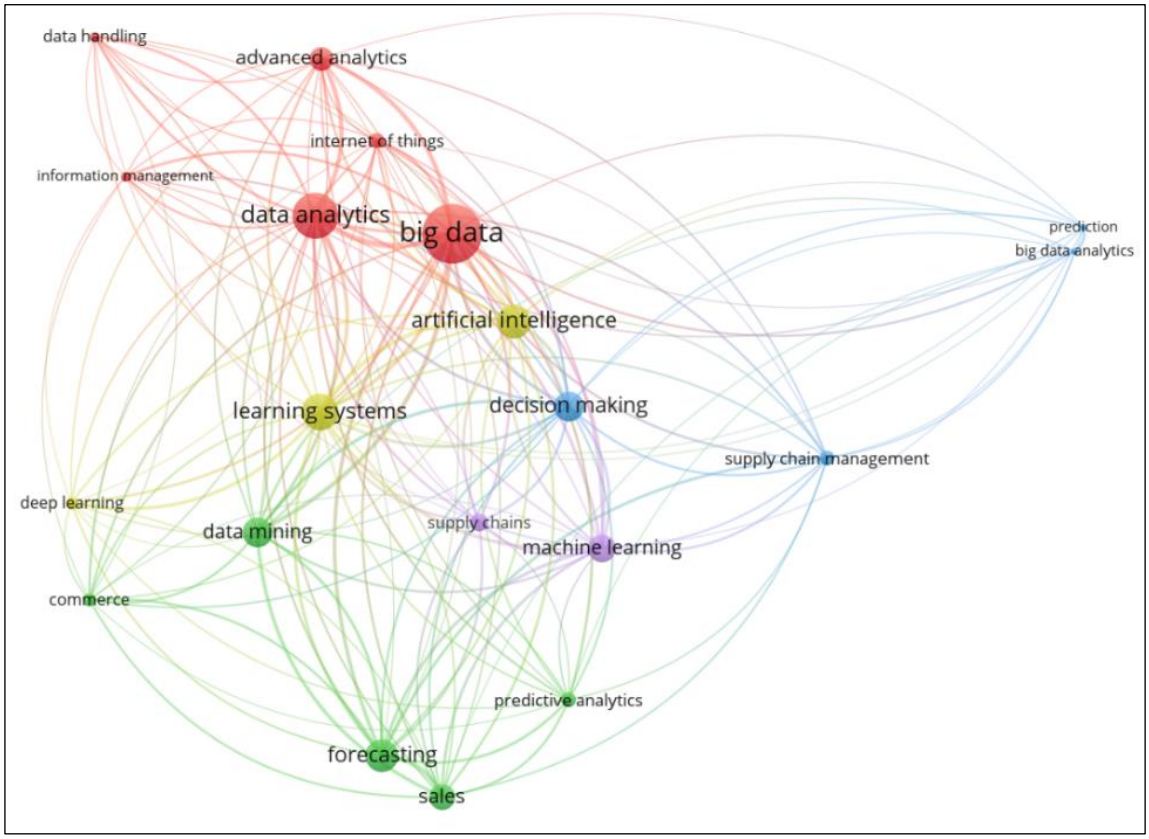
Top keywords by author, measured by “*Total Link Strength*”, are: *Machine Learning, Big Data, Big Data Analytics, Artificial Intelligence, Data Mining, Deep Learning, Data Analytics, Analytics, Internet of Things,* and *Prediction.*

Top index keywords, measured by “*Total Link Strength*”, are: *Big Data, Data Analytics, Learning Systems, Artificial Intellingence, Forecasting, Decision Making, Data Mining, Machine Learning, Sales,* and *Advanced Analytics.*

The keyword co-occurrence graph, presented in Figure 8, can also be interpreted to highlight specific topics that appear frequently, as well as topics based on general keywords that have a greater scope of coverage. The size of the circles indicates how frequently a given keyword occurs. It is observed that topics about Big Data, Data Analytics, Artificial Intelligence, and Machine Learning are recurrent. At the same time, Deep Learning and Data Handling are less frequent topics, that is, with a smaller number of articles, but with high interest for future research. Furthermore, bibliographic grouping is carried out to understand the different areas of emerging research, providing insights into the direction of future research.

Figure 8

Keyword co-occurrence



Source: Results of this research.

3.2.5 Bibliographic grouping through conceptual structure map

Bibliographic clustering is adopted to group research in the domain of ML and Big Data into different subject areas. Similar themes are grouped in the same cluster. Furthermore, the density of clusters can serve as a measure of the extent of research carried out in a particular subject area. Clusters considered dense are considered saturated areas for research, while clusters with sparse data are considered to have room for future research. The analysis of the bibliographic cluster is carried out through the conceptual structure map using the multidimensional scale (MCA), whose dendrogram graphs are represented in Figure 9. The themes are grouped into 5 clusters.

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

Studies that deal with ML are in the densest cluster, highlighted in green in Figure 9.

As reported by L´Heureux *et al.* (2017), the Big Data revolution promises to transform the way we live, work, and think, enabling the optimization of processes, enabling the discovery of insights, and improving decision-making. Realizing this great potential depends on the ability to extract value from this massive data through data analysis; Machine learning is critical because of its ability to learn from data and provide data-driven insights, decisions, and predictions. However, traditional machine learning approaches were developed in a different era and are therefore based on several assumptions, such as the dataset fitting entirely in memory, which is unfortunately no longer true in this new context.

Artificial intelligence (AI) has existed for more than six decades and has matured over time. The rise of computing superpowers and Big Data technologies appears to have boosted AI in recent years. The new generation of AI is expanding rapidly and has once again become an attractive topic for research. Duan *et al.* (2019) investigate the challenges associated with the use and impact of revitalized AI-based systems for decision-making and offer a set of research proposals for information systems (IS) researchers.

Even with more than two decades of continuous development, learning from imbalanced data is still an intense focus of research. With the expansion of machine learning and data mining, combined with the arrival of the Big Data era, it has been possible to gain deeper insight into the nature of imbalanced learning, while also addressing new emerging challenges. Data-level and algorithmic methods are constantly being improved and hybrid approaches are gaining increasing popularity. Recent trends focus on analyzing not only the disproportion between learnings but also other difficulties embedded in the data. New real-life problems motivate researchers to focus on computational efficiency, adaptive, and real-time methods. Krawczyk (2016) discusses open questions and challenges that need to be resolved to further develop the field of imbalanced learning. Some vital areas of research on this topic have been identified, covering the entire spectrum of learning from imbalanced data: classification, regression, clustering, data flows, Big Data analytics, and applications, for example, in social media and computer vision.

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

Studies examining BI and BDA are in the red cluster shown in Figure 9. The results indicate that much-emerging research accepted and published in journals falls into this category. This cluster is therefore highly dominant. The studies in this thematic group

examine the need to adopt BI and BDA techniques. Most notably, Chen, Mao, and Liu (2014) review the background and state of the art of Big Data, focusing on the four phases of the BD value chain: data generation, data acquisition, data storage, and data analysis. data. Articles in this domain were analyzed in terms of emerging research topics, as well as in terms of top researchers and most important contributions. Furthermore, BD presents a unique characteristic when compared to traditional data: it is commonly unstructured, requiring more analysis in real-time, as reported by HU *et al.* (2014).

The importance of Big Data in improving a company's performance is corroborated in the study by Choi *et al.* (2018), who explored big data analysis techniques, identifying their strengths and weaknesses, as well as main functionalities. In this way, BD analysis strategies were discussed to overcome the respective computational and data challenges.

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

Studies in IoT, Data Handling, and Cloud Computing have been extensively researched, marked in violet in Figure 9. Gill, Tuli, and Xu (2019) explore how the three emerging paradigms (Blockchain, IoT, and AI) will influence future computing systems. cloud computing and proposed a conceptual model for cloud futurology to explore the influence of emerging paradigms and technologies on the evolution of cloud computing.

Calatayud, Mangan, and Christopher (2019) explore how the supply chain of the future will be autonomous and have predictive capabilities, bringing efficiency gains in an increasingly complex and uncertain environment. The study is carried out through a systematic and multidisciplinary review of the literature, reviewing 126 articles published in the period 1950-2018.

Ahmadi *et al.* (2019) report that IoT is an ecosystem that integrates physical objects, software, and hardware to interact with other objects. An aging population, scarcity of healthcare resources, and rising medical costs make IoT-based technologies necessary, which can be adapted to address these healthcare challenges. Therefore, this systematic literature review was carried out to determine the main application area of IoT in healthcare.

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

Articles that investigate new technologies related to healthcare are marked in blue in Figure 9.

With the growth of Big Data in the biomedical and health areas, it is becoming possible to carry out precise analyses of the benefits of medical data in advance for detecting diseases. However, analysis accuracy is reduced when the quality of medical data is incomplete. Furthermore, different regions exhibit characteristics of certain regional diseases, which can weaken the prediction of disease outbreaks. In this way, Chen *et al.* (2017) simplify machine learning algorithms for the effective prediction of chronic diseases and frequent disease outbreaks in communities.

Razavian *et al.* (2015) presented a new approach to population health in which data-driven predictive models are learned based on type 2 diabetes outcomes. The approach enables risk assessment from readily available electronic claims data in large populations. The proposed model reveals early and late-stage risk factors. We collected: complaints, pharmacy records, use of health services, and laboratory results from 4.1 million individuals between 2005 and 2009, in an initial set of 42,000 variables that, together, describe the complete health status and history of each individual. Machine learning was then used to methodically refine the set of predictive variables and fit models that predict the onset of type 2 diabetes.

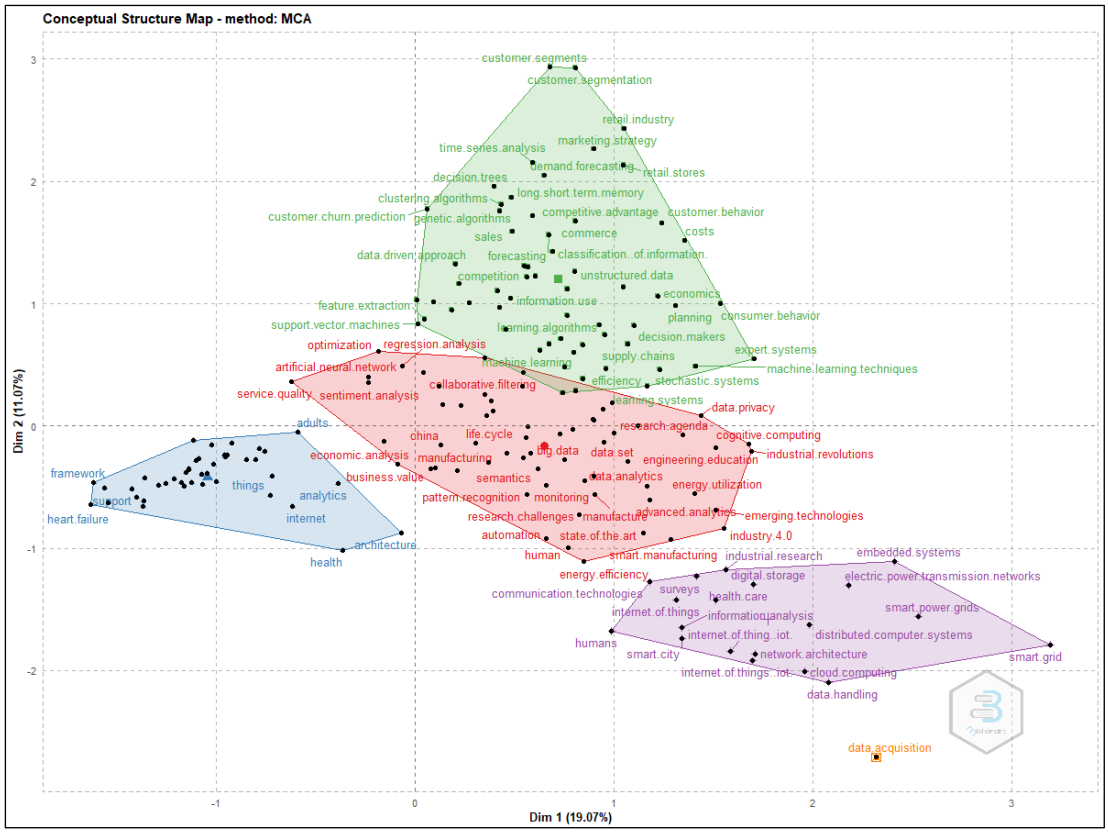
Cluster 5: *Data Acquisition* and *Predictive Analytics*

Studies aiming to analyze the impact of data acquisition and predictive analysis (marked in orange in Figure 9) are scarce, that is, they demonstrate that studying the impact of classification and predictive analysis represents a broad scope for future research.

Making appropriate decisions is, in fact, a key factor in helping companies facing supply chain challenges. Nguyen *et al.* (2021) report two data-driven approaches that allow you to make better decisions in supply chain management. The method *Long Short Term Memory* (LSTM), based on multivariate time series data forecasting networks, is suggested, and a method LSTM *Autoencoder* combined with support vector machine (SVM) algorithm class for sales anomaly detection.

Figure 9

Conceptual structure map (initial topic clusters k = 5)



Source: Results of this research.

Therefore, clusters 1, 2, and 3 are dense, that is, they will continue to emerge in the future. Clusters 4 and 5 have little research, but are emerging research areas. An overview of the 5 clusters is presented in Table 7.

Table 7

Overview of the 5 clusters

Cluster	Central Focus	Main themes explored	Total Articles	Most cited article in each cluster		
				Article Title	Main author	Total citations
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

Source: Results of this research.

The evolution of the theme groups in the form of a timeline was plotted for 10 years, in 5 intervals (2013–2014, 2015–2016, 2017-2018, 2019-2020, 2021-2023), as shown in Figure 10. Cluster 1 has a highly dense cluster with continuous and growing interest, i.e., indicating an increase in articles published in the ML domain, confirming its dominance. Cluster 2 is also considered significant due to the growth in the use of BI and BDA.

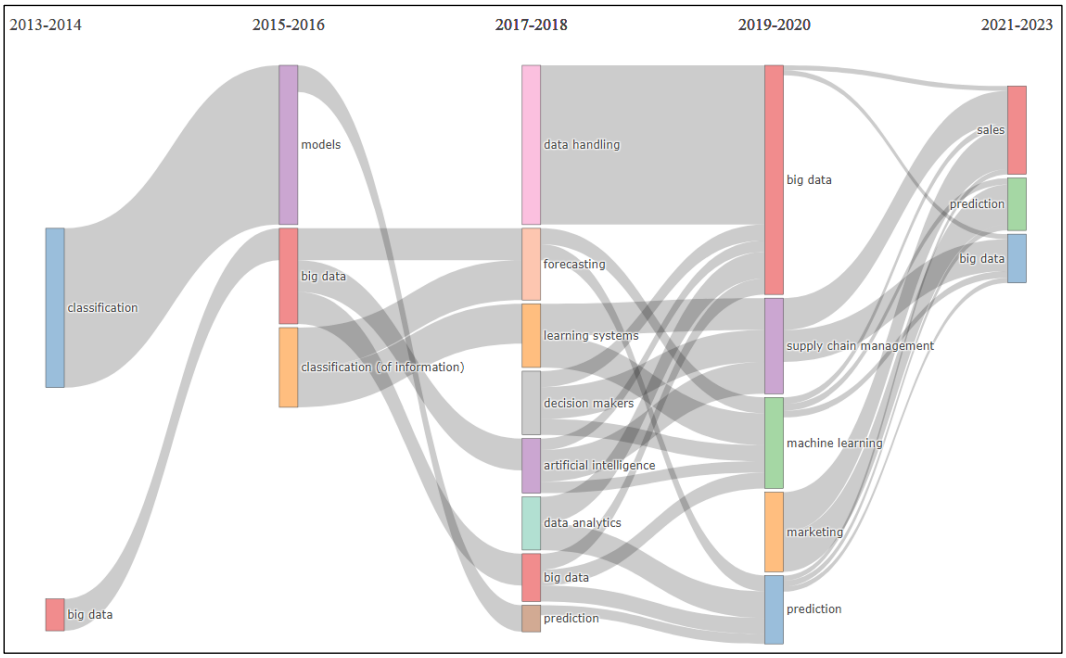
Clusters 3 and 4 continue to gain traction due to the use of IoT devices, Cloud Computing, as well as early disease detection. However, in cluster 5 there are topics with little coverage, indicating low interest in the themes of the articles, that is, they demonstrate that studying the impact of classification and predictive analysis represents an area with low academic exploration.

Figure 10 illustrates the thematic evolution map for the years 2013-2023. The period

from 2013 to 2014 revolves around the theme of "Big Data" and classification methods to deal with the large amount of data. Gradually, other themes evolved in the period 2015 to 2016, e.g. "*Learning Systems*" and "*Classification (of information)*", indicating the need to use learning systems to classify information. In the period from 2017 to 2018, the terms *"Artificial Intelligence", "Decision Making", "Deep Learning"*, and *"Prediction"* stand out, indicating the search for mechanisms to support the industry in decision making and event prediction. From 2019 to 2020, the terms "*Prediction*", "*Machine Learning*", "*Big Data*", and "*Decision Making*" gain traction, thus indicating the demand for using machine learning methods for decision-making. In the period from 2021 to 2023, it is clear that the terms IoT, *Big Data*, *Decision Making*, and *Prediction* continue to dominate the themes of academic articles.

Figure 10

Thematic evolution map for the years 2013-2023



Source: Results of this research.

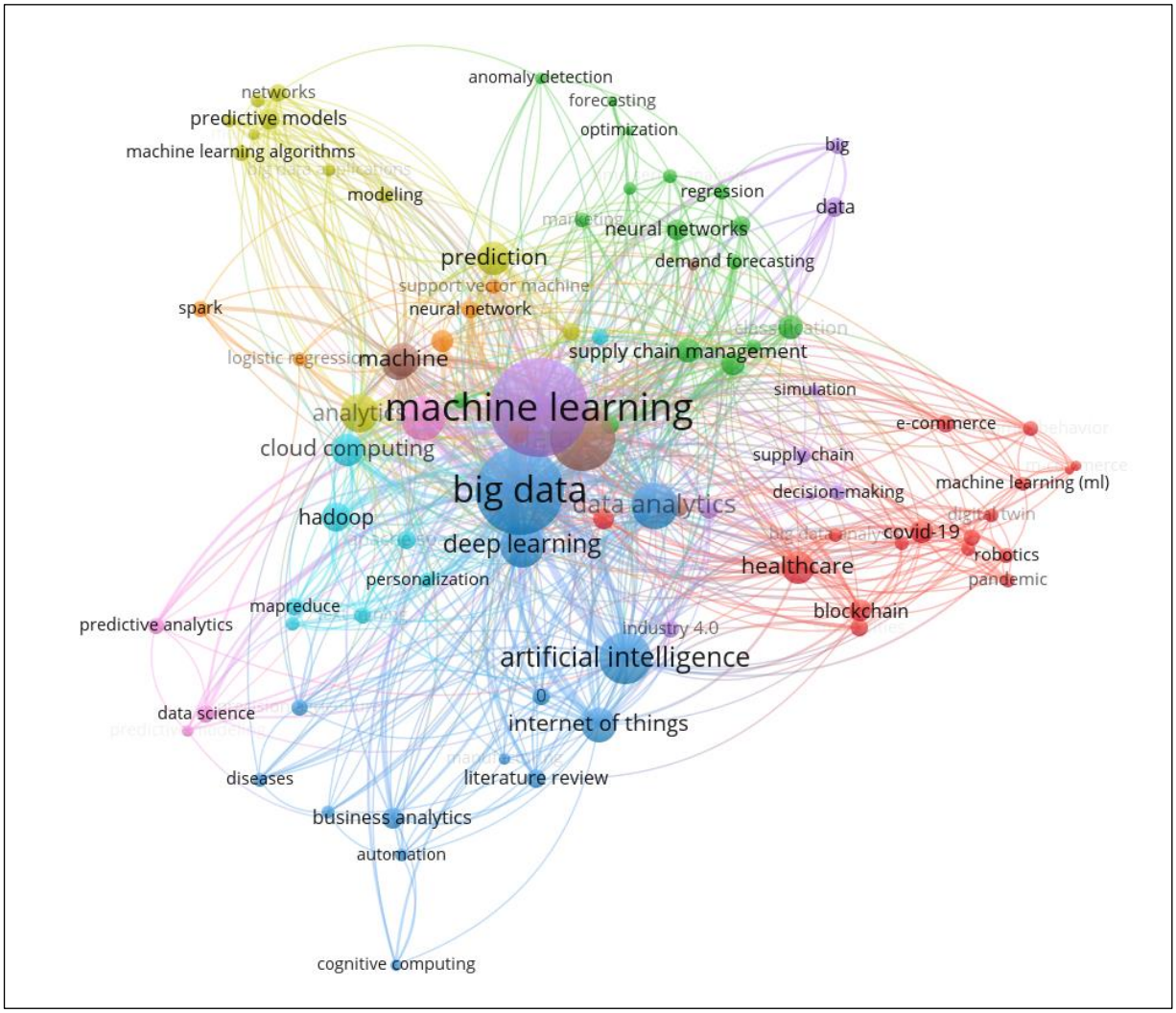
3.2.6 Title co-occurrences

In Figure 11, words that occur frequently and concomitantly in articles are analyzed, forming a network. The most frequently occurring words include *Machine Learning*, *Big Data*, *Data Analytics*, *Deep Learning*, and *Artificial Intelligence*. Other research domains are also identified as *Healthcare*, *Blockchain*, *Covid-19*, and *Supply Chain Management*. In truth,

Machine Learning has often been used with *Big Data* and *Data Analytics*, since ML depends on a "*Data Set"* to be trained and applied.

Figure 11

Title co-occurrence



Source: Results of this research.

4 DISCUSSION

To address the first research question, i.e. the focus of the research on ML, a descriptive analysis was conducted on the bibliometric corpus. The results are summarized in the 7 items below:

(a) Main authors: In terms of the number of citations and bibliographic coupling results, illustrated in Tables 2 and 6, it was found that Yunhao Liu from China is a frequently cited name, followed by Yogesh K. Dwivedi from the United Kingdom. In terms of the number of publications (NP), it appears that Zhenhua Li from China, J. W. Wang. from Japan, and Yi Wang from China are the most productive authors. Nripendra P. Rana, Yogesh K. Dwivedi, Kuttimani Tamilmani, Arpan Kumar Kar, Yi Wang, Yanqing Duan, Santosh Misra, Sujeet Kumar Sharma, Qixin Chen, and Chongqing Kang are the top 10 authors in terms of bibliographic coupling.

(b) Key Countries: The most productive countries for the ML domain in terms of total number of publications (NP) are the USA, India, China, and the United Kingdom, followed by Germany and Canada, as illustrated in Table 3. For citations by publication, Poland, Egypt, and Singapore are the top three countries with the highest C/P metric, while China, the US, and the UK appear with the highest number of total citations (TC).

(c) Top universities: The 3 most productive universities are *the University of Michigan*, *Pennsylvania State University*, and *King Saud University*, as shown in Table 4. In terms of research collaboration, the *Indian Institute of Technology*, *Swansea University*, and *Copenhagen Business School* are engaged in a broad research collaboration in the ML domain.

(d) Main journals: The 10 journals with the highest number of publications and citations are: *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*, and *Industrial Management and Data Systems.* Based on bibliographic coupling, the top 10 journals are: *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*, and *International Journal of Production Economics.*

(e) Areas of knowledge: In Figure 4, it is evident that the areas of knowledge with the highest number of publications are: *"Computer Science", "Engineering", "Decision Sciences", "Social Sciences", "Business, Management and Accounting"*, and *"Mathematics"*.

(f) Co-citation network analysis by Author: It is inferred from Figure 6, in which Dwivedi *et al.* (2021) and Wang *et al.* (2019) are highly cited, these authors are responsible for the link between the two co-citation networks. On the left side, it is evident that the authors Sharma S.K., Misra S., Duan Y., and Kar A.K. frequently cite Dwivedi Y.K., and on the right side, Chen Q., Kang C., and Choi T.M. frequently quote Wang Y.

(g) (g) Prestige analysis: The three most prestigious articles are Liu Y. *et al.* (2020), Li Z. *et al.* (2020) and Wang Y. *et al.* (2019). However, the article by Chen M. *et al.* (2014) received the highest number of citations.

For the second research question, that is, the thematic analysis of the evolution of ML, a grouping of keyword co-occurrences was carried out, with the following results:

(a) Top keyword co-occurrences: "*Artificial Intelligence"* (AI), "*Learning Systems"*, and "*Deep Learning"* are often correlated, while the term "*Big Data*" often occurs in conjunction with "*Data Analytics*", "*Internet of Things*", and "*Advanced Analytics*". The terms "*Data Mining*", "*Predictive Analytics*", "*Forecasting*", "*Commerce*", and "*Sales*" are often correlated. The terms "*Decision Making*", "*Big Data Analytics*", "*Prediction*", and "*Supply Chain Management*" are directly related. Despite the term "*Machine Learning*" present a direct relation with "*Supply Chains*", it is observed in VOSViewer that ML has a strong co-occurrence with all other groups: "*Big Data*", "IA", "*Decision Making*", "*Data Mining*", and "*Learning Systems*".

(b) Informações dos *Clusters*: From Figure 15 and Table 7, the emerging themes were grouped into five bibliographic clusters, with Cluster 1 (*Machine Learning Techniques, Decision Makes, Sales*, and *Forecasting*) presenting greater density, followed by Cluster 2 (BI and BDA), Cluster 3 (IoT, *Data Handling*, and *Cloud Computing*), Cluster 4 (*Activity Recognition, Health, Medicine, Service Quality,*

Therapy, *Heart Failure*, *Classification*, and *Risk*) and Cluster 5 (*Data Acquisition* and *Predictive Analytics*), which presents diverse articles and has low exploration in predictive analysis.

(c) Word cloud analysis by title and abstract: *Big*, *Learn*, *Machine*, *Prediction*, *Analysis*, *Intelligence*, and *Retail* are frequent words in publication titles, while *Big*, *Learn*, *Prediction*, *Algorithm*, *Model*, *Method*, *Analysis*, and *Technique* are frequent words in abstracts, which corroborates the themes covered by the journals.

For the third research question, that is, the identification of future theoretical and/or practical research areas, the results of the cluster analysis and co-citation analysis showed that DBA, ML, and Deep Learning are essentially linked to solving problems business forecasting and decision making in application areas such as stock markets, marketing, and supply chain management. The role of cloud computing and IoT are also cited to serve as infrastructure and generate a large amount of data from sensors and actuators.

CONCLUSION

The present study presented a bibliometric analysis of ML, considering articles from *Scopus* and *Web Of Science* journals from 2013 to May 2023.

In terms of theoretical contributions, the results achieved can help future researchers identify emerging themes for research and potential collaborations. First, the study examined the focus of current ML usage. The focus illustrated the main contributions in terms of authors, universities, journals, and countries to the ML domain. Secondly, it highlighted the main thematic areas, grouping them bibliographically into five clusters. Regarding the scope for future research, it was observed that in previous studies, such as Batistic and Van (2019), the same bibliometric protocol was adopted to study the impact of BDA techniques on companies. This study extends existing research to narrow the focus to predictive applications of large databases in an ML context. Existing literature on ML analytics has also been detailed, with a thematic evolution map indicating emerging themes. Thus, these themes provide direction for future research in the ML domain.

Regarding practical contributions, the study provided insight into the different topics published in the last ten years in the ML domain. This thematic evolution shows that ML is becoming a domain sought after by researchers and information system professionals. Today,

data is a new commodity and this research can help companies that wish to invest in the adoption of ML techniques to obtain a competitive advantage through more assertive diagnosis. Research and development teams can adopt this bibliometric protocol with small adjustments to the search string, being able to deepen future searches, and retrieve relevant documents as reference checkpoints for other approaches related to Big Data and ML.

While the study reveals some interesting findings and provides useful insights, there are also some limitations. First, the data sample was limited to the Scopus and Web of Science databases due to the availability of access to extract relevant articles from the last ten years. Second, a specific combination of keywords was used for bibliometric analysis, which can be adjusted to derive different insights. Furthermore, the period for extraction can be varied to reveal different trends in publications and citations.

The need for predictive analytics is found not only in the corporate sector for diagnostics but also as an emerging research area. The main motivator for research in this field is the need to develop highly accurate tools with a high ability to predict resources obtained in different segments, such as banking and financial services, marketing, supply chain, people management, and sales prediction.

AUTHORS' CONTRIBUTIONS

Contribution	Martins, E.	Galegale, N. V.
Contextualization	60%	40%
Methodology	40%	60%
Software	60%	40%
Validation	60%	40%
Formal analysis	60%	40%
Investigation	60%	40%
Resources	60%	40%
Data curation	60%	40%
Original	60%	40%
Revision and editing	60%	40%
Viewing	60%	40%
Supervision	40%	60%
Project management	40%	60%
Obtaining funding	---	---

REFERENCES

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

Antonopoulos, I., Robu, V., Couraud, B., Kirli, D., Norbu, S., Kiprakis, A., Flynn, D. *et al.* (2020). Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renewable and Sustainable Energy Reviews* 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

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

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

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

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

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., Aarts, G., Coombs, C., Crick, T., Duan, Y. *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

Galegale, N. V., Fontes, E. L. G., & Galegale, B. P. (2017). Uma contribuição para a segurança da informação: um estudo de casos múltiplos com organizações brasileiras. Perspectivas em Ciência da Informação, 22(3), 75–97. http://dx.doi.org/10.1590/1981-5344/2866

Gill, S. S., Tuli, Shreshth, Xu M., Singh, I., Singh, K. V., Lindsay, D., Tuli, Shikhar *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

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

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

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

Martins, E., & Galegale, N. V. (2023). Sales forecasting using machine learning algorithms. Revista de Gestão e Secretariado (Management and Administrative Professional Review), 14(7), 11294–11308. https://doi.org/10.7769/gesec.v14i7.1670

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

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

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

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