Advancement of Data Science in Economics Studies: A bibliometric analysis of the use of machine learning

Caio Oliveira Azevedo

Fábio Junior Clemente Gama

Bruno Castro Alves

Kaylane Manuele Nunes Feitoza

ABSTRACT
This paper aims to explore the results of a mapping of academic research focused on Data Science in economics studies, with a specific focus on the use of machine learning techniques. It seeks to identify how Data Science intercepts the economy and its various axes of action. This is a descriptive research, in that bibliometric analyzes were performed using the Bibliometrix package of the R software, extracted from the Scopus database (SCP), taking the period of 2013 as a time frame to 2023. To carry out the analyzes, the descriptors “machine” and “learning” were used, 1,415 works were found. Based on the results, it was possible to confirm the strong evolution in the publication of works that deal with machine learning applied to economics or economic factors, as well as the mapping of the main axes of discussions on the theme and the respective authorship networks.

Keywords: Data science, Machine learning, Bibliometrics, Economic sciences.

1 INTRODUCTION

Over the last few decades, technological advances have been observed that have changed the way society lives, both in the professional sphere and in interpersonal communication. At the heart of this revolution is the internet, a global information system that connects people, businesses, and institutions in different parts of the world. The technological revolution, supported by information and communication technologies, provoked a new arrangement in the spheres of society, especially in the way of producing science (LEVINE, 2009).

At the end of the last century, technological advances raised the speed and volume of information transactions to an astonishing level (FINZER, 2013), added to the greater capacity for data collection and storage (*data warehousing*), provided a drastic increase in the amount of data available and also in the possibilities of data-driven decision-making. From then on, a dynamic of scientific production and professional performance was configured, based mainly on the greater magnitude of the use of computational resources.

The widespread use of information and communication technologies, coupled with more economical data storage and faster processing by computers, has given rise to the phenomenon known as the "data deluge". This phenomenon comprises a complex system of data extracted from the connections between the

computational networks that involve the contemporary world (RAUTENBERG; CARMO, 2019).

The greater volume of information coming from the abundance of data has provided scientific researchers with greater ability to deepen their research. And not infrequently, in some cases, it enables advances that were previously unfeasible. From this perspective, it is notorious that the "deluge of data" has transformed the process of scientific production and, above all, contributed to advances in the exploration of knowledge in the most diverse areas that intercept the use of data (CLEVELAND, 2014).

In this scenario, in which data and its information have become one of the most valuable resources (ECONOMIST, 2017), Data Science emerges as a *multidisciplinary approach, configuring itself as a fourth paradigm of science, mainly because it has established itself as a fundamental discipline for the analysis and interpretation of large data sets (*big data)) in several areas.

Data Science understands, therefore, a set of methodological support that combines principles and practices from the areas of mathematics, statistics, artificial intelligence and computer engineering, aimed at capturing, exploring and mining data, to be processed by software (CLEVELAND, 2014).

It is evident, therefore, that Data Science is also inherently interdisciplinary (FINZER, 2013). A simple search on the internet is enough to find new undergraduate courses or graduate centers in Data Science in several universities around the world, with master's or doctoral programs, implemented in faculties of Computer Science, Administration, Economics, Engineering, Biostatistics, etc.

In the area of economics, for example, this approach has played an increasingly frequent role, offering new perspectives and tools to understand the completeness of economic systems, mainly through *machine learning* . For Jordan (2019), what is now labeled as artificial intelligence is nothing more than what is called *machine learning*.

Tools such as *machine learning* introduce statistical modeling practices in computer systems in order to develop algorithms capable of instructing and acquiring knowledge automatically (MONARD; BARANAUSKAS, 2003). Informally, the artificial intelligence is related to efforts to automate human cognitive tasks (CHOLLET; ALLAIRE, 2017), is, therefore, viscerally related to the development of computers or programming.

From this perspective, it should be noted that the objective of *machine learning* is not the same as that considered in regression analysis, a traditional method in econometric studies. While econometrics aims to understand how each predictor variable is associated with the response variable, in *machine learning* the goal is to select the model that produces the best predictions, even if the variables selected for this purpose are not those considered in a standard analysis (MORETTIN; SINGER, 2022).

Athey and Imbens (2019) point out that the intersection between *machine learning* and traditional econometrics methods presents much more productive, secure, and sophisticated research results. The authors reinforce that the use of *machine learning* deserves to be highlighted in the area of economic

research, since it implies significant advances.

Lechner (2023), for example, relates the consequences of the transformations in the use of data science in econometric methods and how these computational mechanisms have revolutionized the robustness of the results, and consequently of the methodologies. It highlights that the use of *machine learning algorithms* is capable of obtaining much more efficient results on the impacts of a public policy and its effects on subgroups and target audiences.

It is evident, therefore, that Data Science intersects the economy and its various axes of action. From this perspective, this article aims to explore the advancement of Data Science in Economics studies, with a specific focus on the use of *machine learning techniques*. To this end, a bibliometric approach is proposed, which will allow examining and quantifying the progress and evolution of these studies over time, as well as measuring the collaborations between authors from different countries, since these metrics are indicative of the degree of expansiveness and repercussion of the analyzed theme.

Bibliometric analysis is widely recognized as a valuable tool for evaluating scientific output, identifying emerging trends, and analyzing collaboration among researchers in a given field. In this study, a series of bibliometric metrics, such as the number of publications, the frequency of citations, and the identification of co-authorship networks, will be applied to examine the scientific production related to the use of *machine learning* in Economics.

The objective of this bibliometric study is to contribute to the understanding of the advancement of Data Science in Economics studies, providing an overview of the prevailing trends and approaches. In addition, it is expected that the bibliometric metrics used will offer *valuable insights* for researchers, practitioners, and students interested in exploring the applications of *machine learning* in this field.

For this purpose, a quantitative method of statistical measurement was used, with applications in the *R* [1] software and the use of the Biblioshiny *interface of the* bibliometrix*2* [2] package. The searches were carried out in the Scopus database (SCP), taking the period from 2013 to 2023 as a time frame. The descriptors "machine" and "learning" were used to perform the analyses, *and 1,415 studies were found.*

Based on the results, it was possible to confirm the strong evolution in the publication of works that deal with *machine learning* applied to economics or economic factors, as well as the mapping of the main axes of discussions on the subject and the respective authorship networks.

In addition to this introduction, this article is divided into 4 more sections. The next one presents a brief contextualization about applications of Data Science in Economics, highlighting all the current debate about the use of predictive models of *machine learning* and the resulting results. The third section, on the other hand, shows the methodological procedures, including the description of the data and the empirical strategy. The fourth

[1] R Core Team (2022)
[2] Aria and Cuccurullo (2017)

section presents the results of the research, while the fifth section lays out the final considerations.

2 *MACHINE LEARNING* EM ECONOMIA

This section presents seminal studies related to the theme of this article, in addition to contextualizing the recent history of development and contribution of empirical studies on the efficiency in the use of *machine learning* models applied to Economic Sciences.

Petrová (2022) states that the advent of the technological revolution has impacted the spheres of the neoclassical labor market, given that these tools, as the author points out, are essential factors to strengthen the economies of countries. In order to keep up with the transformations caused by technological progress, the dynamics of new economists in the market must be expanded by qualification in the use of these mechanisms. It is in this scenario that works in the area of economics have evidenced the integration between machine learning techniques and traditional methods of economic analysis.

The incorporation of intelligent systems has profoundly marked the new work dynamics of the economy, as evidenced by research with applications in econometrics, which present advances in the understanding of the behavior of causal relationships of variables that are the focus of study in this science (LECHNER, 2023). In this sense, the economist's arrangement of tools should be expanded in accordance with the use of *Machine Learning* systems, in order to preserve, simultaneously, the main traditional methods of econometrics (ATHEY; IMBENS, 2019).

The crucial point of economic science is the ability to interpret the information established in a data set, developed from the use of statistical modeling in order to present an overview of actions in social, political and private spheres, in addition to optimizing and guiding decision-making. In view of this, the study entitled *Causal Machine Learning and its use for public policy*, published by the US journal *Swiss Journal of Economics and Statistics* by Lechner (2023) points out the exponential technological development of intelligent systems and the relevance of applications of *Causal Machine Learning techniques* in the economic field.

The results allow for more robust and accurate estimates, in addition to the differential point of the use of these techniques in relation to traditional econometric methods, which is to identify the heterogeneities of the impacts of these policies, in order to investigate them methodically in order to achieve the objectives foreseen during the elaboration of these actions, improving them and correcting the identified gaps. and thus use public resources efficiently (LECHNER, 2023).

As presented by Athey and Imbens (2019), one of the main and innovative applications of ML methods is for the estimation of average treatment effects, an essential parameter for economic analyses, whether at macro or microeconomic levels. Such methods present more efficient performances in relation to traditional econometric methods, especially when it comes to analyzing large databases or more complex

mechanisms of treatment distribution. The authors of the study present another scenario for the use of machine learning techniques, this one focusing on optimal policy estimates in economic models. ML methods for optimal policy estimates are able to to enable researchers to take into account a set of options when capturing the complexity of the relationships between variables, in order to promote much more precise policy suggestions (ATHEY; IMBENS, 2019).

In addition to the use of intelligent system gears in econometric studies, in the field of macroeconomics these mechanisms have also been incorporated in the analysis of estimates and causalities of macroeconomic aggregates. de Jesus and Besarria (2023) developed from the use of *K-means* techniques, an unsupervised ML method that processes *clusters* of the sample of banking institutions, the assembly of new metrics for the classification of bank insolvency risks for financial institutions that trade on Brazilian stock exchanges. The use of these techniques presents differentials in the results because they have the ability to extract unstructured data and treat their non-linearity effectively (de Jesus; BESARRIA, 2023).

Accordingly, Casabianca et al. (2022) by using the traditional *logit* model and a supervised machine learning algorithm, (*AdaBoots*), to identify which are the determinants of banking crises among the samples of countries under study and concluded that the performance of the *Machine Learning* engine stands out from the traditional model. This aspect was observed by the authors based on the analysis carried out outside the sample in which the *AdaBoots predictive model* presented a more efficient performance in the face of the sample, since this tool stands out in terms of "AUROC, sensitivity, specificity, and relative utility" (CASABIANCAet al., 2022).

Bitetto, Cerchiello, and Mertzanis (2023) also developed a study on the estimation of causality in economic growth and expansionary policies with a technical approach to uplift *modeling*, an area still under exploration by machine learning. The ML framework mentioned allows us to work with causal effects, estimating causality in each individual on real GDP growth and expansionary economic policy changes. This analysis focuses on the use of causal algorithms that are branched in order to identify lagged causal effects on expansionary policies and, in view of this identification, to estimate the impact on economic growth (BITETTO; CERCHIELLO; MERTZANIS, 2023). The authors point out that *the Uplift* modeling presents results that, in addition to meeting the literature, also express macroeconomic behaviors of theory and reality.

In the scenario of corporations and finance, the use of *Machine Learning*, as pointed out by the authors Khan et al. (2023), promotes a more concise prediction about the level of corporate vulnerability of the sample of countries studied by identifying relationships between the effects of the COVID-19 pandemic and this crisis scenario in financial institutions. Despite this defense, the study argues that ML modeling may present a risk to the accuracy of the results due to the assumptions processed by the algorithm, given that the analysis of the performance of this tool in predicting credit default risk presents risks to the supervisory

validation method (KHAN et al., 2023).

The effects of changes in the socio-political scenario in the macro and microeconomic field, estimates and predictions about the implications of economic policies, are examples of economic analyses that can be made and improved from the use of intelligent systems such as *Machine Learning* in conjunction with traditional methods of data analysis. It is evident, therefore, in this conjuncture of advances in Data Science, that ML algorithms help and enhance decision-making based on the results obtained by them, especially those directed to predictions, as pointed out by the studies surveyed.

3 METHODOLOGICAL PROCEDURES

3.1 EMPIRICAL STRATEGY

Bibliometrics is a quantitative method (statistical measurement) for mapping and evaluating a research area based on the bibliographic data of its scientific production, enabling the researcher to identify, from the most cited articles in the area to the most relevant terms and concepts (SILVA; HAYASHI; HAYASHI, 2011; MAN; VITORIANO, 2015).

In this sense, in the field of bibliometrics, the empirical strategy is fundamental to collect and analyze data on scientific production, including citations, authors, journals and keywords. This approach aims to identify trends, patterns, and relevant research areas that contribute to the development of a solid body of knowledge (BROADUS, 1987).

The empirical strategy in bibliometrics goes through several stages. Initially, researchers collect bibliographic data from reliable sources, such as academic databases, scientific journals, and relevant conferences. This data is then organized and structured to facilitate quantitative and qualitative analysis.

During the empirical analysis, statistical and data visualization techniques are applied to identify patterns in scientific production, such as the most cited sources, main authors, journals with the highest impact, and emerging themes. This information provides a comprehensive understanding of the field of study, allows for the identification of promising research areas, knowledge gaps, and opportunities for future investigations.

To this end, in the next two subsections, the process of data collection will be applied, as well as their treatment, in order to explore the scientific panorama in a systematic and data-based way, contributing to the advancement of knowledge and the promotion of relevant discussions in the field of study in question.

3.2 DATA

The bibliometric mapping based on the use of statistical and mathematical metrics allows the elaboration and analysis of the systematic literature review in order to identify and evaluate the advances in the use of Machine Learning computational gears in the economy in the last ten years. Therefore, data

were collected from the Scopus Elsevier database (ELSEVIER, 2023).

The choice of the Scopus database is based on the following reasons: (i) comprehensiveness and diversity of content; (ii) data quality and accuracy: The platform is known for its rigorous selection of sources and indexing processes, which results in high data quality and accuracy. (iii) citation and impact metrics, which are essential for conducting cocitation analysis.

In addition to the observations highlighted, the Scopus Elsevier database includes a significant part of the main scientific productions in the area of economics and data science, which enables a bibliometric mapping in a broader and more concise way between the interdisciplinarity of the axes addressed.

A total of 1608 articles were registered using the Scopus advanced search command given the following parameters: TITLE-ABS-KEY (machine AND learning) AND PUBYEAR> 2012 AND (LIMIT-TO (SRCTYPE, "j")) AND (LIMIT-TO (OA, "all")) AND (LIMIT-TO (PUBSTAGE, "final")) AND (LIMIT-TO (SUBJAREA, "ECON")) AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (LANGUAGE, "English")).

The criterion of delimitation of the parameters in the ten-year time interval is based on the purpose of ensuring that the researched articles are updated in accordance with the evolutions of Data Science, especially to identify the trend of potential advances in the use of *Machine Learning* in economic works in the face of the expansion of the *Big Data* phenomenon in the last decade. The restrictions followed by the "and" command aim at the selection of articles from academic journals that are related to the specific area of Economics and that are published.

Table 1 – Description of the pre-selected variables

Variables Analyzed	Nomenclature	Description
Annual Scientific Production	ASP	It refers to the monitoring of the production of
		scientific studies in a specific field of study over the course of the year.
		of the analysis period.
Most Global Cited Docu-	MGCD	It refers to the most cited scientific documents globally.
ments		mind within a particular field of study or area
		of research. These documents are the ones that have received
		the highest number of citations from other scientific papers
		at the international level, demonstrating its wide-ranging influence.
		and recognition in the academic community.
WordCloud	TOILET	Known as a word cloud, it is a representation of the word cloud.
		visual tion of the most frequent words in a text or
		set of texts. In this representation, the words are
		arranged in sizes proportional to their frequency, or
		That is, words that appear more often have a
		larger in the cloud.
Co-occurrence Network	CON	Represents the relationships between elements that co-occur

		in
		a set of data or documents. In this type of network,
		The elements are represented by nodes (or vertices) and the
		co-occurrences between them are represented by edges (or
		connections).
Collaboration Network	COLN	It represents collaborations between researchers, countries
		or institutions in a particular field of research, or
		area of expertise.

Source: Prepared by the authors from Elsevier (2023).

Therefore, as shown in Table 1, the variables analyzed in this bibliometric study play a fundamental role in the understanding and evaluation of scientific research. These variables combined offer a broad and valuable approach to exploring scientific research, thus allowing for in-depth analysis and a more complete view of the trends, interactions, and impact of scholarly production in this ever-evolving field.

4 TREATMENT

The treatment of the data collected for the bibliometric analysis was carried out using the *R*3 software, in which bibliometrix *4*[2] was used, a package that enables the interface between *biblioshiny* application. This is a tool that enables the generation and visualization of various bibliometric analyses in bibliographic data files (ARIA; CUCCURULLO, 2017).

According to Hjorland (2008), bibliographic coupling was introduced by Kessler (1963), and co-citation analysis was independently suggested by Marshakova and Small, both in 1973. In his experiments, Kessler (1963) found a high degree of semantic correlation between the grouped documents when using the criterion of bibliographic coupling.

Clustering packages use bibliographic coupling as the basis for their process. bibliometrix employs the "walk trap" clustering algorithm, an effective method for grouping bibliographic documents based on their citations. The algorithm selects a random document and then expands to group others that are cited by the initial. This process is repeated until all documents have been grouped. The result is a set of clusters of documents that are semantically related (ARIA; CUCCURULLO, 2017).

A coupling network can be obtained from the general formulation:

$$B = The \times AT \qquad (1)$$

[2] 3 R Core Team (2022)
4 Cobo, López-Cózar and Herrera-Viedma (2019)

where A is a two-way network. The *bij* element indicates how many bibliographic couplings there are between manuscripts i and j. In other words, *bij* gives the number of paths of length 2, which moves from i along the arrow and then to j in the opposite direction. Matrix B *is* symmetric (B = BT). The strength of the coupling of two articles, *i* and *j*, is defined simply by the number of references that the articles have in common, as provided by the *bij element* of matrix *B*.

The degree of bibliographic coupling between A and B (Bipartite Network) is determined by the frequency with which their documents are cited simultaneously, as is the case with C.

And to build the collaboration network by clusters of authors, the following general formulation is used:

$$AC = AT \times The \quad (2)$$

where: *A* is a bipartite network of Manuscript Authors. Like matrix *B*, matrix *C* is also symmetric. The main diagonal of *C* contains the number of cases in which a reference is cited in our data frame. In other words, the diagonal element *ci* is the number of local citations of reference *i*.

At the end of this process, in which the data matrices already formed by bibliometrix are indexed in the interface, the execution of the scientific mapping begins, through the use of the main bibliometric indicators to identify the metrics of the topic studied.

By analyzing the selected bibliographic corpus, we will seek to identify publication patterns, main authors and institutions, collaborations, most relevant journals, and most recurrent keywords. These analyses will allow us to assess the growth of the scientific literature in this specific area, as well as identify the main topics and research approaches adopted by scholars of the intersection between Data Science and Economics.

5 RESULTS

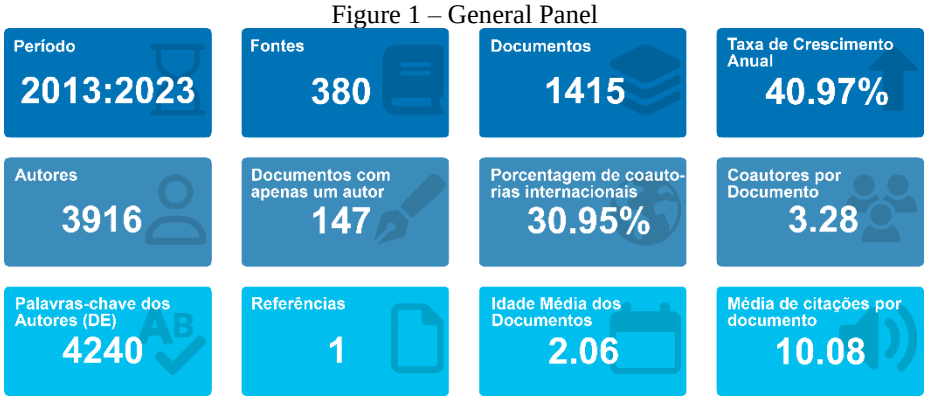
The systematic review of the literature was based on 1,415 research documents obtained from 380 different sources, published between 2013 and 2023. The data reveal a significant annual growth rate of 40.97%, which demonstrates the advances in the application of these tools in the economic sphere over the last ten years. In addition, the documents have an average age of 2.06 years, a parameter that shows the relevance of this theme and its growth in accordance with advances in data science.

The average number of citations per document of 10.08 indicates that the content of these documents has received notoriety and scientific recognition in the academic community. Another factor analyzed is the dataset, which involves a total of 3,916 authors who contributed to the 1,415 documents, demonstrating a trend of collaboration between authors, with an average of 3.28 co-authors per document.

On the other hand, 147 of the documents were authored by a single individual, which represents

10% of the total analyzed. This suggests that joint research, in the Big Data scenario, *is a relevant practice in this new dynamic of scientific production marked* by the intersection of Data Science and Economics by allowing broader perspectives and intellectual influences in the face of the application of intelligent systems for the elaboration of research. In addition, approximately 30.95% of the collaborations involve international partnerships, indicating the scope of the theme in academic research beyond the borders of the institutions of the countries identified.

Figure 1 – General Panel



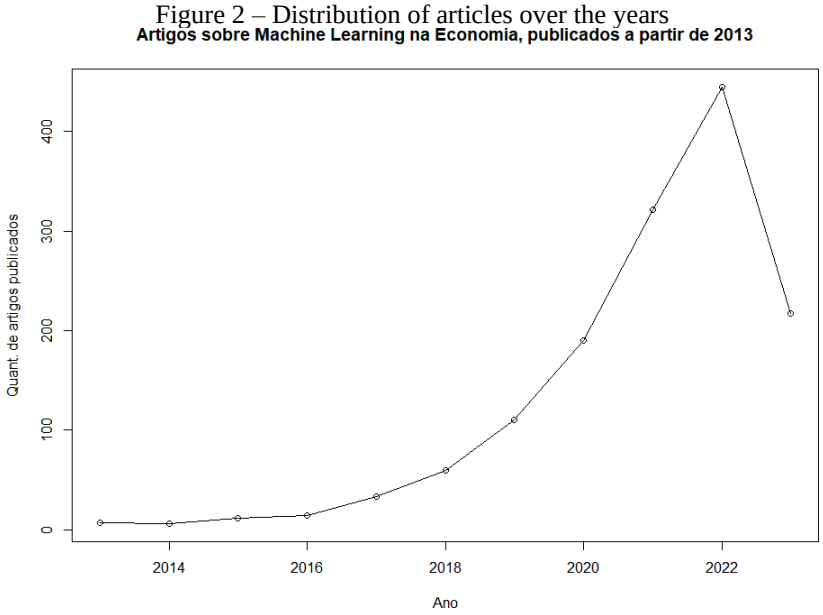
Source: Prepared by the authors from Elsevier (2023).

The growth trajectory in the publication of articles on *machine learning* in economics over the years has been exponential (Figure 2), confirming the trend of the intersection between Data Science and Economic Theory. Initially, there was a steady and moderate increase until 2016, followed by an accelerated increase. This is mainly related to technological advances, added to the greater capacity for data collection and storage (data *warehousing*), in the face of the era of data deluge, providing greater possibilities in the application of artificial intelligence methods.

This scenario of continuous technological advances provides researchers with greater ease of retaining, obtaining and storing varied data that, when processed through the application of ML algorithms, can be used as raw material for the development of increasingly in-depth research. These innovations make it possible to solve problems in studies that were previously little explored due to the absence of information contained in these cybernetic records and the difficulty of extracting and storing them. In this sense, the increasingly efficient results of studies focused on this theme boost the production of articles in this area, which explains the significant advance from 2018 to the following years, thus confirming the optimistic changes in the new dynamics of scientific production in the face of advances in technologies in spaces of economic science.

It is important to note that the 2023 analysis took into account exclusively the first six months of the year, a parameter that influenced the totals presented. Thus, it is expected that the trend of expansion of published works, which continued on an increasing path until

Figure 2 – Distribution of articles over the years



Source: Prepared by the authors from Elsevier (2023).

2022, reaching a peak with 445 publications this year, should be maintained for the following months, when considering the magnitude of the numbers already recorded in the first six months, which not only manage to surpass the twelve months of 2021 but also approach the results of the full year 2022.

From this perspective, Table 2 presents the most cited works by articles that deal with *machine learning*. Citation analysis provides the main theoretical references on the field in question and the contributions of journals to the evolution of the topic. The conceptual associations with the highest incidence show how relevant these productions are to the field of study of economics with the insertion of ML methods, since they indicate the argumentative support for the development of works in this area.

The first among the most cited, *Double/debiased machine learning for treatment and structural parameters*, addresses the challenge of inference in a low-dimensional parameter ($\theta 0$) in the face of high-dimensional parameters ($\eta 0$). The authors introduce the concept of "Double or Debiased Machine Learning" (DML) and propose a method that uses machine learning techniques and Neyman's orthogonal moments to estimate $\theta 0$ consistently, so as to overcome the bias introduced by regularization and *overfitting* in the estimation of $\eta 0$ (CHERNOZHUKOV et al., 2018).

Next, the second most cited work is *Empirical Asset Pricing via Machine Learning*, which consists of a comparative analysis of machine learning methods for the classic problem of empirical asset pricing. The authors demonstrate that the use of machine learning predictions, specifically decision trees and neural networks, can generate large economic gains for investors. The predictive increase obtained, compared to regression models, is attributed to the ability of *machine learning* methods to capture nonlinear interactions between predictors, which is not explored by linear regressions (GU; KELLY; XIU, 2020).

The article *Machine learning and deep learning* is the third most cited, with 342 citations and being the most recent among the top three, published in 2021. The work explores the use of

Table 2 – Table of documents most cited by articles on Machine Reading

Pos.	Documents	Quotes	Newspaper
1	Chernozhukov et al. (2018)	520	*The Econometrics Journal*
2	Gu, Kelly e Xiu (2020)	369	*The Review of Financial Studies*
3	Janiesch, Zschech e Heinrich	342	*Electronic Markets*
	(2021)		
4	Dubey et al. (2020)	216	*International Journal of Production Economics*
5	Baek, Mohanty e Glambosky	211	*Finance Research Letters*
	(2020)		
6	Rust Huang (2021)	206	*Journal of the Academy of Marketing Science*
7	Athey and Imbens (2019)	202	*Annual Review of Economics*
8	Deryugina et al. (2019)	171	*American Economic Review*
9	Zhu et al. (2019)	156	*International Journal of Production Economics*
10	Belloni et al. (2017)	149	*Econometric*

Source: Prepared by the authors from Elsevier (2023).

Of intelligent systems with artificial intelligence capabilities, often based on machine learning and deep learning. The study provides a conceptual overview of these methods, with emphasis on both the challenges faced when implementing such systems in electronic markets and networked businesses, as well as the importance of human-machine interaction and the provision of artificial intelligence services (JANIESCH; ZSCHECH; HEINRICH, 2021).Table 3 shows the diversification of the areas of journals that lead the publication of articles in the field in question, with a clear trend towards the area of finance. This result reinforces the use of intelligent systems for better job performance in risk and vulnerability analysis in the financial sector. Of particular note are the journals *Quantitative Finance* and *Financial Innovation*, both of great international relevance, with a strong influence in research and academic teaching centers.

Figure 3 shows the most frequent terms in scientific articles about the application of *machine learning methods* in the economic axis. The main objective of this analysis is to identify the trends and areas of greatest interest in this field of research. The word cloud, extracted from *kayswords*, points out the main ML methods applied in these studies. From this analysis, it was found that the term *forecasting* appears with greater predominance, demonstrating the tendency to apply ML algorithm techniques for forecasting purposes.

The occurrence, also in evidence, of *learning algorithms* suggests that studies are focused on investigating and comparing different machine learning algorithms. This approach highlights the importance of developing efficient algorithmic approaches to solve specific problems of the various axes of the economy, since the efficient performance of *machine learning methods* is closely related not only to the structures of the data worked, but also to the objectives expected by the application, which confirms the

most appropriate algorithms according to each particularity of results that are sought in the development of the data. Research.

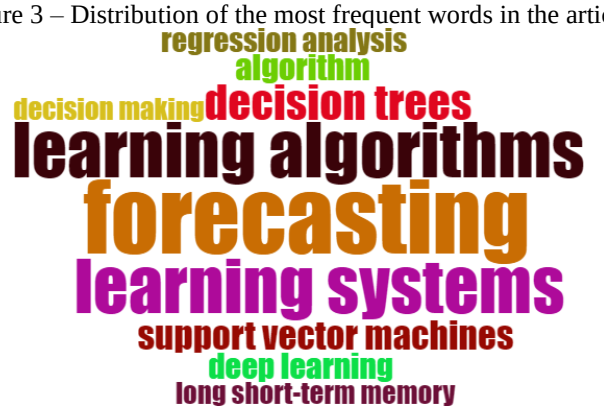
Following this, the term learning systems *reflects the importance attributed to the practical implementation* of machine learning algorithms in real systems and opens up the possibility of variations in the nomenclature mentioned in the articles. In relation to the specific machine learning techniques, the *elements decision trees* and *support vector machines* are recurrent, which

Table 3 – Journal that has published the most articles on Machine learning

Sources	Articles	Qualis	Area
Frontiers in Energy Research	107	A3	Biotechnology
Journal of Advanced Transportation	77	A2	Public and Business Administration, Accounting Sciences and Tourism
Risks	76	B2	Economy
Forecasting	34	A2	Economy
Humanities and Social Sciences Communications	29	B4	Social sciences
Journal of Risk and Financial Management	29	B2	Economy
Journal of Finance and Data Science	28	-	Finance & Data Science
Quantitative Finance	27	A1	Economy
Financial Innovation	19	A1	Public and Business Administration, Accounting Sciences and Tourism
Periodicals of Engineering and Natural Sciences	19	-	Engineering

Source: Prepared by the authors from Elsevier (2023).

Figure 3 – Distribution of the most frequent words in the articles



Source: Prepared by the authors from Elsevier (2023).

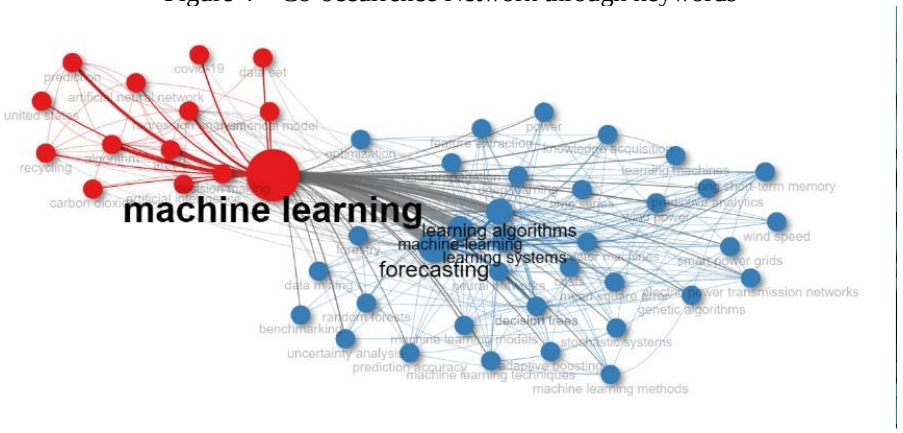
Indicates the interest in exploring these models for predictive classification and performance maximization that assist in assertive decision-making based on more robust results.

In addition, the terms *algorithm and deep learning* are also found with some frequency, indicating a comprehensive and in-depth analysis of the development and application of various algorithms, such as

deep neural networks to learn complex patterns. Finally, the term *regression analysis* is also highlighted, which reveals the interest in analyzing the relationship between variables through statistical modeling, applied in particular ML contexts.

In addition, it is worth noting that among the most used words in the articles, they are basically divided into two subgroups, as pointed out by the word co-correspondence analysis. The analysis of the Co-occurrence Network of keywords, as shown in Figure 4, points to the existence of distinct thematic clusters. Clusters are sets of objects or data that share similar characteristics, constitute an important technique in the analysis and analysis of

Figure 4 – Co-occurrence Network through keywords



Source: Prepared by the authors from Elsevier (2023).

Organization of information. The process of cluster formation involves the use of clustering techniques, which analyze the similarities between the elements and group them based on these characteristics, creating cohesive and distinct groups. The red cluster is made up of machine learning research focused on regression analysis and its association with decision-making. This grouping allows us to explore how regression techniques have been applied in the context of decision-making, whether in autonomous systems or in complex business environments. On the other hand, the Blue Cluster covers articles that address various machine learning techniques, with a focus on prediction and Machine Learning Models. In this cluster, one can find works that explore predictive algorithms, such as decision trees, neural networks, vector support machines, among others, as well as studies that deal with the creation and evaluation of different machine learning models to solve specific problems in various areas.

As highlighted, about 90% of the publications related to the theme are composed by more than one author, therefore, we can conclude that there are cooperation networks among the authors. Bibliometrix makes it possible to evaluate these networks through graphs of cooperation between countries and authors. Figure 5 presents an analysis of international partnerships between countries in the field of economic research through the World Map of Authors' Collaboration on articles on *machine learning*. The results

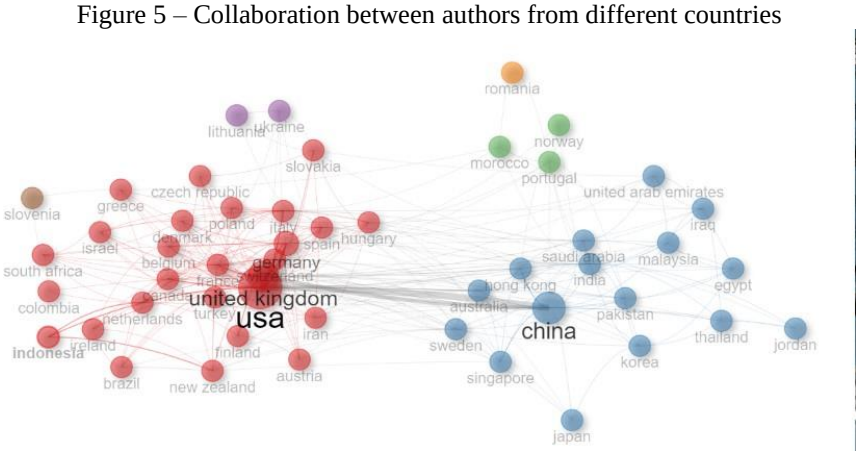
indicate that the United States (USA) and the United Kingdom lead the collaborations, with 37 collaborations, followed by cooperation between the United States and China (28 collaborations), and China and the United Kingdom (23 collaborations).

Collaboration analysis is an important tool for bibliometric studies. It allows you to identify networks of researchers who work together, as well as the areas of research that are most active and collaborative. This is useful for identifying research trends, tracking the development of new areas, and identifying researchers who are making important contributions to a field.

Through these partnerships, it is possible to tackle complex challenges, develop new approaches, and enhance machine learning techniques. These results highlight the importance of international cooperation for scientific and technological progress, allowing significant advances in the field of *machine learning*. Partnerships between countries contribute to the development of innovative solutions and can positively impact society, paving the way for advances in areas such as medicine, industry, finance, among others.

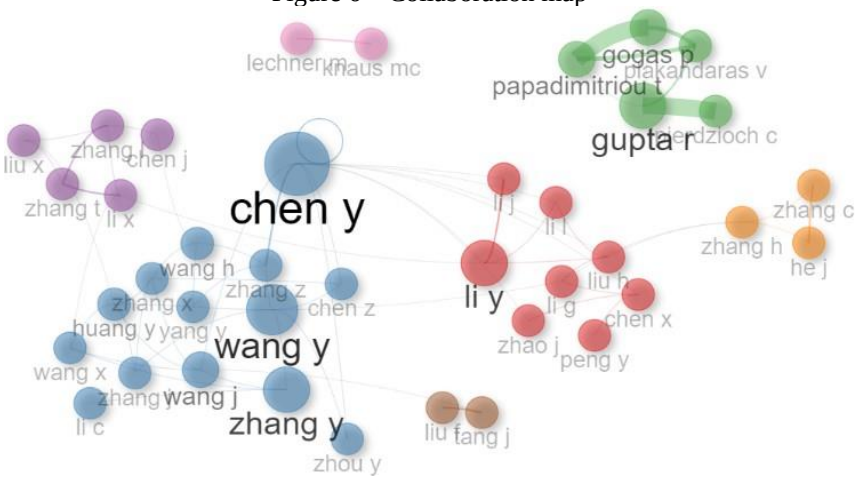
The collaboration of different authors in scientific works is essential to promote the

Figure 5 – Collaboration between authors from different countries



Source: Prepared by the authors from Elsevier (2023).

Figure 6 – Collaboration map



Source: Prepared by the authors from Elsevier (2023).

The complex and interdisciplinary nature of contemporary problems requires the combination of expertise and knowledge of multiple specialists. Through collaboration, authors can aggregate and combine their individual perspectives, skills, and academic experiences to enrich the quality and breadth of scientific work. When analyzing graph 6, the Blue Cluster stands out, which is composed of authors such as Chen Y 5, Wang Y [6], Zhang Y[7, Wang H 8, Wang J[9] and others. Among them, the author Chen Y stands out with a very high value in "Between the Lines", suggesting a significant contribution to articles in machine learning. In these clusters, we observed a dense network of collaboration, evidenced by the relatively high values of "Proximity" for all authors, indicates that they collaborate closely with each other in the production of articles. ¨

The exponential growth evidenced in this analysis reflects the growing recognition of the relevance of *machine learning* in deepening the understanding and improvement of economic processes. Through advanced analytics and data-driven decisions, ML emerges as a powerful tool that drives business efficiency and effectiveness, catalyzing innovations and discoveries that will meaningfully shape the economic landscape of the future. Its continuous development is of paramount importance to face the challenges and seize the opportunities of a global scenario in constant technological transformation.

6 FINAL THOUGHTS

The advancement of Data Science in the field of Economic Theory over the last decade is recognized due to the evidence of the multidisciplinary approach of this science in accordance with the increased relevance of data as decisive resources for decision-making (ECONOMIST, 2017), especially from the use of machine learning algorithms.

The bibliometric analysis developed in the present work proves the high incidence of interest and engagement of the economic field in the use of *machine learning* tools, showing a constant growth until

2016, as shown in figure 2, followed by a rapid increase in the number of publications. This result suggests that the researcher of the Economic Sciences, in the prominent scenario of *Big Data*, is led to apply algorithms that help him in the analysis of large databases. Through the use of advanced analytics and informed decision-making, *machine learning* has the potential to drive business efficiency and effectiveness and drive innovations in the economic future.

Another result provided by this bibliometrics can be analyzed from the three most cited articles, highlighted in table 2, which demonstrate the diversity of relevant topics in *machine learning* that are being applied in economic studies. The surveys of these studies point to the development of new technical arrangements of machine learning and its practical applications in finance, confirming, from the analysis of the journals that most use ML algorithms, the prominence of this area in the use of these tools to optimize the predictive performance of risk analysis found in the face of the volatility of financial markets. The extraction and mining of information from data allows these researchers to broaden their analysis of financial movements around the world as a result of the magnitude of increasingly comprehensive data.

The analyses also provide insight into focus areas and prevailing trends, punctuating the use of techniques such as "Forecasting", "Decision Trees" and "Support Vector Machines". These findings indicate the application of ML methods for predictive analytics, modeling, and mapping of outcomes and decisions. In practice, this application demonstrates the interest and responsibility of researchers for more sophisticated and robust results regarding forecasts of economic impacts and/or variations in the various axes of the economy, among other conjunctures in this field that require efficient decision-making.

Another aspect that the study reveals is the significant advance in the trajectory of this theme in relation to theoretical exchange, as observed by the wide network of collaboration between researchers from different locations around the world. This analysis allows us to understand that the leadership of international partnerships is based on the prominent academic framework of countries that essentially present a certain consolidation in technological developments, such as the United States, the United Kingdom and China, a scenario of new dynamics of scientific production Enhanced by the search for innovations capable of improving the approach to data in the economy. Thus, it is evident that the application of computational gears of ML advances in accordance with the trajectory of ascendancy of technologies.

In summary, the contributions from this research provide theoretical and practical directions for the exploration of the main trends in the use of ML in notoriety in Economic Theory. From bibliometrics, researchers and professionals interested in the axis addressed by the study can visualize the main purposes of the applications of algorithms of intelligent systems for approaches similar to those identified during the research.

In addition, the study presents theoretical contributions for future research in the area of economics and innovation, according to the results that confirm the intersection of Data Science and Economics. The

parameters evaluated during the study prove the upward trends of the application of ML, thus providing pertinent insights for the academic community to develop and boost work with the efficient support of *Machine Learning algorithms*, since the development of scientific fields is a gradual and continuous trajectory.

JEL CLASSIFICATION

A12, B00, B49, C55

REFERENCES

ARIA, M.; CUCCURULLO, C. bibliometrix: An r-tool for comprehensive science mapping analysis. *Journal of Informetrics*, Elsevier, v. 11, n. 4, p. 959–975, 2017. Disponível em: <https://doi.org/10.1016/j.joi.2017.08.007>.

ATHEY, S.; IMBENS, G. W. Machine learning methods that economists should know about. *Annual Review of Economics*, Annual Reviews, v. 11, p. 685–725, 2019.

BAEK, S.; MOHANTY, S. K.; GLAMBOSKY, M. Covid-19 and stock market volatility: An industry level analysis. *Finance Research Letters*, v. 37, 2020.

BELLONI, A. et al. Program evaluation and causal inference with high-dimensional data. *Econometrica*,v. 85, n. 1, p. 233 – 298, 2017.

BITETTO, A.; CERCHIELLO, P.; MERTZANIS, C. Measuring financial soundness around the world: A machine learning approach. *International Review of Financial Analysis*, Elsevier Inc., v. 85, 1 2023. ISSN 10575219.

BROADUS, R. N. Toward a definition of "bibliometrics". *Scientometrics*, Springer, v. 12, p. 373–379,1987.

CASABIANCA, E. J. et al. A machine learning approach to rank the determinants of banking crises overtime and across countries. *Journal of International Money and Finance*, Elsevier Ltd, v. 129, 12 2022. ISSN 02615606.

CHEN, Y. Comparing content marketing strategies of digital brands using machine learning. *Humanities and Social Sciences Communications*, v. 10, n. 1, 2023.

CHERNOZHUKOV, V. et al. Double/debiased machine learning for treatment and structural parameters.

The Econometrics Journal, v. 21, n. 1, p. C1–C68, 01 2018.

CHOLLET, F.; ALLAIRE, J. J. *Deep learning with R*. [S.l.]: Manning Publications, 2017.

CLEVELAND, W. S. Data science: An action plan for expanding the technical areas of the field of statistics. *Statistical Analysis and Data Mining*, John Wiley and Sons Inc, v. 7, p. 414–417, 12 2014. ISSN 19321872.

COBO, M. J.; LóPEZ-CóZAR, E. D.; HERRERA-VIEDMA, E. *Bibliometrix: A Comprehensive R Package for Bibliometric Analysis*. [S.l.], 2019. Disponível em: <https://cran.r-project.org/web/packages/bibliometrix/bibliometrix.pdf>.

de Jesus, D. P.; BESARRIA, C. da N. Machine learning and sentiment analysis: Projecting bank insolvency risk. *Research in Economics*, v. 77, n. 2, p. 226–238, 2023. ISSN 1090-9443.

DERYUGINA, T. et al. The mortality and medical costs of air pollution: Evidence from changes in wind direction. *American Economic Review*, v. 109, n. 12, p. 4178 – 4219, 2019.

DUBEY, R. et al. Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations. *International Journal of Production Economics*, v. 226, 2020.

ECONOMIST, T. *The world´s most valuable reource is no longer oil, but data (2017)*. 2017. Disponível em: <https://www.economist.com/leaders/2017/05/06/ the-worlds-most-valuable-resource-is-no-longer-oil-but-data>.

ELSEVIER. *Scopus: A Database of Academic Journals and Articles*. [S.l.], 2023. Acesso em 10 de maiode 2023. Disponível em: <https://www.scopus.com/>.

FINZER, W. The data science education dilemma. *Technology Innovations in Statistics Education*, California Digital Library (CDL), v. 7, 2013.

GAO, G.; WANG, H.; GAO, P. Establishing a credit risk evaluation system for smes using the soft voting fusion model. *Risks*, v. 9, n. 11, 2021.

GU, S.; KELLY, B.; XIU, D. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, v. 33, n. 5, p. 2223–2273, 02 2020.

HJORLAND, B. What is knowledge organization (ko)? *Knowledge Organization*, v. 35, n. 2/3, p. 86–101, 2008.

HUANG, M.-H.; RUST, R. T. A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science*, v. 49, n. 1, p. 30 – 50, 2021.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. *Electronic Markets*, Springer, v. 31, n. 3, p. 685–695, 2021.

JORDAN, M. I. Artificial intelligence—the revolution hasn't happened yet. *Harvard Data Science Review*, PubPub, v. 1, n. 1, p. 1–9, 2019.

KESSLER, M. M. Bibliographic coupling between scientific papers. *American Documentation*, v. 14,n. 1, p. 10–25, 1963.

KHAN, M. A. et al. Corporate vulnerability in the us and china during covid-19: A machine learning approach. *Journal of Econ omic Asymmetries*, Elsevier B.V., v. 27, 6 2023. ISSN 17034949.

LECHNER, M. Causal machine learning and its use for public policy. *Swiss Journal of Economics and Statistics*, Springer Science and Business Media Deutschland GmbH, v. 159, 12 2023. ISSN 22356282.

LEVINE, D. K. Economics: Eyes on the prize? *Science*, v. 323, p. 1296–1297, 2009. ISSN 00368075.

LIANG, D. et al. A novel fault monitoring method based on impedance estimation of power line communication equipment. *Frontiers in Energy Research*, v. 11, 2023.

MEDEIROS, J. M. G. de; VITORIANO, M. A. V. A evolução da bibliometria e sua interdisciplinaridade na produção científica brasileira. *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*,v. 13, n. 3, p. 491–503, 2015.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003.

MORETTIN, P. A.; SINGER, J. d. M. *Estatística e ciência de dados*. 1ª. ed. Rio de Janeiro: LTC, 2022.

PETROVá, K. The impact of digital technologies on neoclassical labour market. *DANUBE*, v. 13, n. 4, p.318–330, 2022. Disponível em: <https://doi.org/10.2478/danb-2022-0020>.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponí-vel em: <https://www.R-project.org/>.

RAUTENBERG, S.; CARMO, P. R. V. do. Big data e ciência de dados. *Brazilian Journal of Information Science: research trends*, Faculdade de Filosofia e Ciências, v. 13, p. 56–67, 3 2019.

SASAKI, Y.; URA, T.; ZHANG, Y. Unconditional quantile regression with high-dimensional data.

Quantitative Economics, v. 13, n. 3, p. 955 – 978, 2022.

SILVA, M. R. da; HAYASHI, C. R. M.; HAYASHI, M. C. P. I. Análise bibliométrica e cientométrica: desafios para especialistas que atuam no campo. *InCID: revista de ciência da informação e documentação*, v. 2, n. 1, p. 110–129, 2011.

WANG, J.; TANG, J.; GUO, K. Green bond index prediction based on ceemdan-lstm. *Frontiers in Energy Research*, v. 9, 2022.

ZHU, Y. et al. Forecasting smes' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics*, v. 211, p. 22 – 33, 2019.