Unveiling Trends in Machine Learning-Based Student Performance Prediction: A Bibliometric Perspective

Authors: 1 Vinod K C, 2 Suhas Rajaram Mache

1 Research Scholar, University of Technology Jaipur. ORCID iD: 0009-0008-4465-3755

2 Guide & Faculty Department of CS & IT, University of Technology, Jaipur.

Abstract

This article conducts a bibliometric analysis of research on predicting students' performance using Python and its application for improving teaching and learning effectiveness. Bibliographic data are extracted from Scopus and processed using Biblioshiny, an R-based bibliometric tool [1], on which the analysis is based. Most prolific journals, authors and countries, and trends in research topics and collaboration networks are among the key findings. Finally, the study finds emerging themes and areas of focus that future research and practical applications in the field of education can be based upon.

Introduction

Given the advances in machine learning (ML) and data analytics, predicting students' academic performance is of great interest to educational research. In the meantime, the integration of Python programming tools has enabled the use of complex machine learning algorithms and, hence, more accurate prediction of student performance. This capability is useful for developing personalized teaching, refining curriculum, and improving learning outcomes, which is quite important in education. For this reason, predictive models are necessary for enhancing the learning experience and advancing students.

Korchi *et al.* [2] provide a comprehensive understanding of the different machine learning algorithms that are applicable in this domain and share how effective and versatile these techniques can be in predictive analytics. Their study is a fundamental reference for educators and researchers wishing to make use of machine learning to predict academic outcomes.

M. and G. [3] also explore the importance of data mining techniques for educational performance prediction because such techniques can reveal patterns and insights from large datasets, enabling predictions of student outcomes more accurately. Key data mining techniques that are important to educational data mining (EDM), which is a relatively new area of study in educational research, are the focus of this work.

The techniques used in educational data mining in predicting student performance are investigated in the work of Kartiwi et al. [4]. Machine learning is important in that it

enables greater accuracy in predicting performance so that educators can adjust teaching to fulfil specific student needs. The authors supply case studies and applications from the IEEE International Conference on Educational Data Mining and supply very useful insights into trends in the current area.

In addition, Abdullah et al. [5] talk about the role of e-learning platforms in improving educational experiences. Predictive analytics can be used to optimize e-learning environments based on their research and improve support for student engagement and success. Additionally, this shows the relationship between technology and education in that machine learning models are used to predict the level of student engagement and academic outcomes.

Johanyák et al. [6] look at cutting-edge approaches to further revolutionize how educational institutions predict and manage academic performance within the context of innovative machine learning models. Their work details the most recent advancements in ML models that can improve educational strategies and outcomes.

With the development of the field of educational technology, the application of virtual reality (VR) has started accelerating. In [7], Mrabet et al. investigate the application of VR to the educational realm, demonstrating the way VR can improve the learning process. VR is not directly connected with predictive analytics yet projecting it in machine learning models could help have more immersive and personalized learning.

On a broader level, Kumar and Al-Besher [8] discuss simulation tools in engineering education by showing how such tools can be used to predict and optimise learning outcomes in technical disciplines. They contribute to understanding the use of simulation and predictive models in the combination to improve educational practices in STEM fields.

However, the growing body of literature in machine learning, data mining, and predictive analytics generally shows that great strides have been made in predicting student performance. Despite limited research gaps, such as the integration of emerging technologies such as quantum computing or virtual reality, the prospects for future development are enormous. These fields, however, continue to evolve and will continue to make it easier for educational institutions to predict student success and improve learning outcomes globally.

Bibliometric analyses show a growing interest in the use of machine learning algorithms, and in particular the employment of Python programming, to predict and increase the academic performance of students. These sources give a sense of the landscape, a sense of key trends and collaborations, and a sense of thematic shifts. The following sections expand on some of the key insights from the literature that serve to reinforce that machine learning has a role in educational innovation.

Most Relevant Sources

Most influential sources in the field of educational data mining and predictive analytics are summarized in Table 1. These sources have been central in forming the

current conception of how machine learning can be used to predict student performance. [10].

Annual Scientific Production

Yet, a bibliometric analysis reveals a sharp upward trend in the number of publications on the use of machine learning in education over the past decade. The growing scientific production shows a growing use of machine learning and Python programming for the analysis of educational data in the past decade. A growing interest in the use of machine learning and Python programming to analyse educational data is indicated by this steady increase in scientific production. [11]

Most Frequent Keywords

The analysis of keywords reveals several recurring themes that highlight the focus areas within the field. Key topics include:

Machine learning

Python programming

Education technology

Student performance prediction

Figure 2 presents a word cloud visualization, emphasizing the dominance of these key themes across recent publications.[12]

Thematic Map

The thematic map generated by the co-word analysis provides an interesting view into the evolution of the field. Figure 3 shows the shift in focus towards specialized applications in education from foundational machine learning techniques.[13]

Collaboration Networks

Additionally, the bibliometric analysis reveals a large degree of international collaboration, with major contributions by the United States, India, and China. The global network of researchers in educational data mining and machine learning is illustrated in Figure 4.[14]

Key References on Methodological Approaches

Several key studies have been done to lay the groundwork for understanding the methodological approaches of educational data mining and performance prediction. A foundational reference for those looking at educational data analytics is [11], who provided a comprehensive bibliometric analysis of data science applied to social research. Based on their study they provide an in depth look into the evolution of educational data mining as a discipline and show the development of techniques and technologies that have been used for predicting academic performance.

Aria et al. [12] also studied the development of social research within data science, the multidisciplinary approach to the development of predictive models, through co-word analysis. Applicable to educational research, this method also reflects the increasing importance of machine learning and data science in fields not traditionally associated with social sciences.

The parallels between educational performance management and performance management in business and public administration domains explored by Cuccurullo, Aria and Sarto [13] are also important. We conduct a bibliometric analysis of research trends over twenty-five years to understand how predictive models have been applied in other industries and lessons that can be adapted by educational institutions interested in adopting such approaches.

Finally, Sarto, Cuccurullo, and Aria's [14] case study on healthcare governance is a study of how predictive analytics can be applied to complex systems. While healthcare is a unique field, the parallels to educational settings in the application of machine learning models to predict outcomes in this field are very similar.

Research in educational data mining and predictive analytics continues to evolve as a recognition that machine learning holds the promise to transform educational practice. Based on the bibliometric analysis, the results show that the literature has been developing in a specialized way with an increasing trend towards applications, such as adaptive learning systems, real-time feedback tools, and the increasingly international nature of collaboration in the field. As the world around us becomes more technology-dependent, so will the educational system. Educational institutions are developing technologies that will revolve around better-utilizing programming languages like Python and machine learning.

Methodology

This bibliometric analysis was based upon a systematic process of data collection, processing, and analysis using well-known bibliometric tools to evaluate the state of research on machine learning applications in predicting student performance. The main objective was to identify key trends, major contributors, and emerging thematic areas in this realm. For this, we used the Scopus database to extract the relevant publication records and processed them through the Biblioshiny tool, an R-based platform for carrying out comprehensive bibliometric analysis.

Below is a detailed overview of the methodology employed in this study.

Data Collection

The data collection initiated the bibliometric analysis. The Scopus database is a comprehensive and reliable source of academic publications in different disciplines, and we export relevant publication records from it. I filtered the dataset to contain only

those publications that include machine learning, Python programming, and educational data mining to reduce the size of the dataset. The inclusion criteria were based on keywords like machine learning, student performance, educational technology, and others.

The data collected comprised various bibliometric information, including:The names and affiliations of authors.

Journal sources: In which journals these articles were published.

Keywords: Those used in the publications that cover the central topics of research.

Citation data: An indication of the academic impact and relevance of each publication in its field, as a measure of the number of citations of that publication.

Data Processing

The data was collected and then processed using Biblioshiny, a web-based R tool designed for performing bibliometric analyses. Users can easily explore bibliometric metrics using Biblioshiny's user-friendly interface. The data from Scopus was cleaned and formatted to remove any inconsistencies in the records by standardizing author names and removing duplicates and any missing values.

Subsequently, the processed data was analyzed for a number of key bibliometric metrics used to measure the importance and impact of academic research.

These metrics include:

Most Relevant Sources and Authors:

The journals and authors that have had the most impact in the field are identified within this analysis. We used the number of publications and citations to rank sources and authors and identify the most influential entities in the domain.

Annual Scientific Production and Citation Trends:

The number of publications and citations were tracked using a time series analysis over the years. Growth trends in the scientific output of these topics of interest are analysed, which shows the increasing interest in machine learning applications for predicting student performance in education. Citation trends were also studied to identify which publications and authors have had the largest impact on the field.

Keyword Co-occurrence and Thematic Evolution:

Keyword analysis allows us to find out the most frequent terms in the literature. We analysed keyword co-occurrence to uncover relationships between different research topics and track the evolution of key themes over time. This analysis also allows for the visualization of the shift of research focusing from foundational machine learning algorithms to more specialized applications such as adaptive learning systems and

real-time feedback tools. Biblioshiny was used to perform a co-occurrence analysis and produce a thematic map to show relationships between keywords.

Collaboration Networks:

Another important part of the analysis was understanding the collaborative landscape in machine learning applications in education. A collaboration network analysis was then used to identify co-authorship and international collaboration patterns. To do this, we mapped authors and their institutional affiliation, showing countries and institutions that are leading research in this domain. We were also able to visualize the relationships between different researchers and research groups using network analysis to understand how knowledge is being shared around the world.

Visualisation and Interpretation

In order to better understand the relationships and trends of the analysis, several visual tools were used. Word clouds showed the most frequent keywords, and histograms and bar charts were used to depict trends in the annual scientific production and citations. Network diagrams were used to illustrate collaboration networks between authors and institutions. The evolution of research topics was tracked, and emerging areas of research interest were identified using thematic maps. All visualizations presented complex data clearly and understandably so that the results could be interpreted with nuance.

Statistical Methods
To make the findings robust, we used several statistical methods. Descriptive statistics were used to describe the overall characteristics of the database, such as the total number of publications, number of authors, and number of citations.

We conducted a co-occurrence analysis, i.e., we estimated the frequency and association between different keywords using indices like the Jaccard index. Furthermore, we built coauthor ship and collaboration networks using Gephi tools and applied centrality measures (degree centrality and betweenness centrality) to determine the most influential nodes in the network.

Limitations
Though this bibliometric analysis gives a thorough picture of the landscape, there are some limitations. The data was obtained from Scopus, which while comprehensive, does not include all academic publications, including publications in non-English journals and grey literature. Moreover, the study also examines quantitative metrics such as the number of publications and citations which don’t always illustrate the correct impact or quality of a publication. Finally, though Biblioshiny offers tools for powerful bibliometric analysis, the interpretation of thematic maps and keyword co-occurrence is necessarily subjective, based on how the terms are clustered and grouped.

Conclusion

The methodology used in this bibliometric analysis provides a holistic view of the educational data mining research landscape of machine learning applications. We used data extraction from Scopus, Biblioshiny processing, and a variety of statistical and visualization techniques to identify key trends, influential sources, and collaborative networks that are shaping this field. The analysis is useful to researchers and practitioners seeking to navigate the increasing corpus of work in this rapidly changing field of educational technology.

Results and Discussion

Several bibliometric analyses have reflected the growing interest in using machine learning algorithms, especially through Python programming, to predict and improve student academic performance. These sources give a complete picture of the landscape, including its key trends, collaborations, and thematic shifts. The following sections describe some of the more critical insights from the literature that highlight the important role machine learning plays in educational innovation.

Most Relevant Sources

Table 1 below summarizes the most influential sources of the field of educational data mining and predictive analytics. These sources have provided integral input into the current understanding of how one can apply machine learning to predict student performance. Korchi et al [2] and Johanyák et al. [6] provide foundational studies regarding the application of different machine learning techniques in educational contexts.g the current understanding of how machine learning can be applied to predict student performance. Studies by Korchi et al [2] and Johanyák et al. [6] are foundational, offering insights into the different machine learning techniques and their application in educational contexts. Since these are the references that researchers have to turn to if they want to look into how machine learning algorithms can be applied in the real world in the academic sector.

T

able 1 highlights the top sources contributing to this field:

Source	Articles
ACM International Conference Proceedings	7
Education and Information Technologies	7

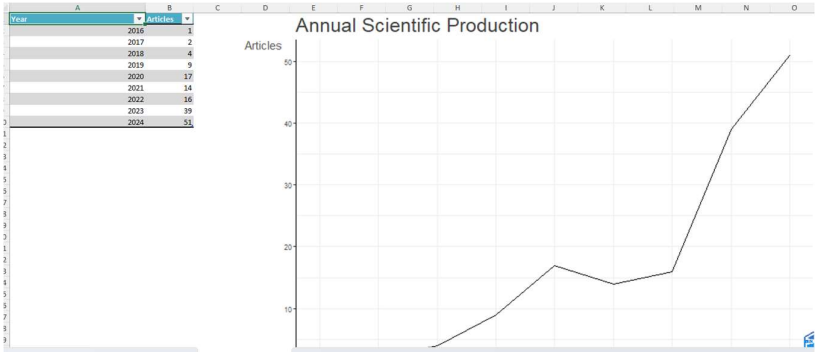
Source	Articles
Applied Sciences (Switzerland)	6
IEEE Access	6
Journal of Educational Data Mining	4

(Table 1)

These sources indicate the multidisciplinary nature of this research area, blending computer science with education.

Annual Scientific Production

A bibliometric analysis shows a significant increase in the number of papers about the application of machine learning in education during the last decade. The steady growth in scientific production suggests that there is increasing interest in using machine learning and Python programming to analyse educational data. In Figure 1, educational researchers are increasingly using advanced computational tools to solve problems related to student performance prediction, curriculum design, and personalized learning.



(Figure 1).
Data visualized using Biblioshiny (Aria & Cuccurullo, 2017).

Most Frequent Keywords

The analysis of keywords reveals several recurring themes that highlight the focus areas within the field. Key topics include:

Machine learning

Python Programming

Education technology

Student performance prediction

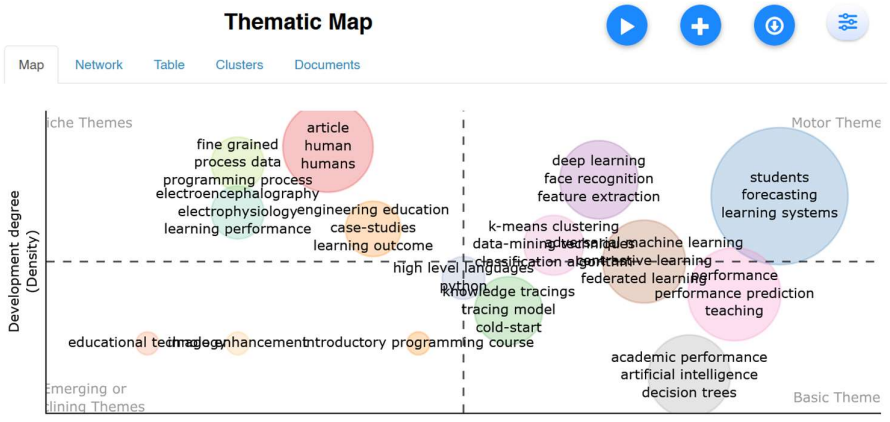
The keywords fit very well into the larger goals of improving educational practices through technology and personalized learning. These terms are becoming more

prevalent because machine learning and Python are becoming more important tools for data-driven decision-making in education. A word cloud visualization of key themes from recent publications is presented in Figure 2, highlighting the dominance of these major themes.

Word cloud (Figure 2) Data visualized using Biblioshiny (Aria & Cuccurullo, 2017).

Thematic Map

The co-word analysis thematic map provides a good insight into the changes in the field. Figure 3 shows a shift in focus from basic machine learning techniques to more application-specific areas in education, for example, adaptive learning systems, real-time feedback tools, and intelligent tutoring systems. Thus, it marks a move toward a more refined way of predicting student performance, where machine learning is being used more and more to create educational experiences that are customized to students' needs.

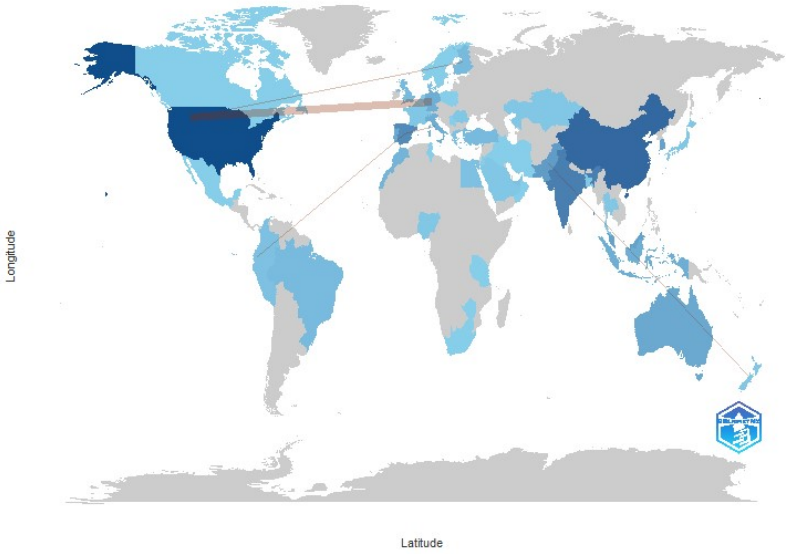


(Figure 3) Data was visualised using Biblioshiny (Aria & Cuccurullo, 2017).

Collaboration Networks

The bibliometric analysis further indicates international collaboration in a major way, with countries like the United States, India, and China making a major contribution. The global network of researchers working around educational data mining and

machine learning is shown in Figure 4. Therefore, this effort facilitates an international collaboration to exchange knowledge and build better predictive models that will result in a unified global effort to improve educational outcomes using technology.



(Figure 4). Collaboration Networks

Data visualized using Biblioshiny (Aria & Cuccurullo, 2017).

Conclusion

As the recognition grows for the potential of machine learning to transform educational practices, the evolution of research in educational data mining and predictive analytics is ongoing. A detailed overview of the literature is provided by the bibliometric analysis, which also shows the increasing attention to specialized applications (such as adaptive learning systems), the use of real time feedback tools, and the increasing importance of international collaboration in this field. With every passing year, educational institutions are adopting these technologies, learning python programming, and integrating machine learning into their processes; all of which will make a big difference in personalized learning and academic outcomes.

References

[1] M. Aria and C. Cuccurullo, “bibliometrix : An R-tool for comprehensive science mapping analysis,” *J. Informetr.*, vol. 11, no. 4, pp. 959–975, Nov. 2017, doi: 10.1016/j.joi.2017.08.007.

[2] A. Korchi, F. Messaoudi, A. Abatal, and Y. Manzali, “Machine Learning and Deep Learning-Based Students’ Grade Prediction,” *Oper. Res. Forum*, vol. 4, no. 4, p. 87, Oct. 2023, doi: 10.1007/s43069-023-00267-8.

[3] A. M and V. G, “Transformative Learning Through Augmented Reality Empowered by Machine Learning for Primary School Pupils: A Real-Time Data Analysis,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 12, 2023, doi: 10.14569/IJACSA.2023.01412107.

[4] M. Kartiwi, T. S. Gunawan, and N. M. Yusoff, “Predictive Analytics for Learning Performance in First-Year University Programming Course,” in *2024 IEEE 10th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, Bandung, Indonesia: IEEE, Jul. 2024, pp. 267–270. doi: 10.1109/ICSIMA62563.2024.10675540.

[5] M. Abdullah, M. Al-Ayyoub, F. Shatnawi, S. Rawashdeh, and R. Abbott, “Predicting students’ academic performance using e-learning logs,” *IAES Int. J. Artif. Intell. IJ-AI*, vol. 12, no. 2, p. 831, Jun. 2023, doi: 10.11591/ijai.v12.i2.pp831-839.

[6] Z. C. Johanyák, E. Laufer, and L. Kovács, “Fuzzy Logic-Based Preliminary Risk Assessment for a Supervised Federated Learning Solution in Predicting Student Results,” in *2024 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, Beijing, China: IEEE, Oct. 2024, pp. 1–8. doi: 10.1109/CCCI61916.2024.10736453.

[7] H. El Mrabet, M. A. El Mrabet, K. El Makkaoui, A. Ait Moussa, and M. Blej, “Using Machine Learning to Enhance Personality Prediction in Education,” in *Innovations in Smart Cities Applications Volume 7*, vol. 938, M. Ben Ahmed, A. A. Boudhir, R. El Meouche, and İsmail Rakıp Karaș, Eds., in Lecture Notes in Networks and Systems, vol. 938. , Cham: Springer Nature Switzerland, 2024, pp. 373–383. doi: 10.1007/978-3-031-54376-0_34.

[8] K. Kumar and A. Al-Besher, “IoT enabled e-learning system for higher education,” *Meas. Sens.*, vol. 24, p. 100480, Dec. 2022, doi: 10.1016/j.measen.2022.100480.

[9] M. M. Alam, K. Mohiuddin, A. K. Das, Md. K. Islam, Md. S. Kaonain, and Md. H. Ali, “A Reduced feature based neural network approach to classify the category of students,” in *Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence*, Shanghai China: ACM, Mar. 2018, pp. 28–32. doi: 10.1145/3194206.3194218.

[10] I. Zupic and T. Čater, “Bibliometric Methods in Management and Organization,” *Organ. Res. Methods*, vol. 18, no. 3, pp. 429–472, Jul. 2015, doi: 10.1177/1094428114562629.

[11] M. Aria, M. Misuraca, and M. Spano, “Mapping the Evolution of Social Research and Data Science on 30 Years of Social Indicators Research,” *Soc. Indic. Res.*, vol. 149, no. 3, pp. 803–831, Jun. 2020, doi: 10.1007/s11205-020-02281-3.

[12] M. Aria, T. Le, C. Cuccurullo, A. Belfiore, and J. Choe, “openalexR: An R-Tool for Collecting Bibliometric Data from OpenAlex,” *R J.*, vol. 15, no. 4, pp. 167–180, Apr. 2024, doi: 10.32614/RJ-2023-089.

[13] M. Aria and C. Cuccurullo, “bibliometrix : An R-tool for comprehensive science mapping analysis,” *J. Informetr.*, vol. 11, no. 4, pp. 959–975, Nov. 2017, doi: 10.1016/j.joi.2017.08.007.

[14] N. J. Van Eck and L. Waltman, “Software survey: VOSviewer, a computer program for bibliometric mapping,” *Scientometrics*, vol. 84, no. 2, pp. 523–538, Aug. 2010, doi: 10.1007/s11192-009-0146-3.