Machine Learning Methods for Academic Achievement Prediction: A Bibliometric Review

Fajar Nugraha[a,b,*], Widowati Widowati[c], Aris Sugiharto[d]

[a] Doctoral Program of Information System, School of Post Graduate Studies, Diponegoro University,
Jl. Imam Bardjo S.H., No. 5, Pleburan, Semarang, Indonesia 50241
[b] Department of Information System, Faculty of Engineering, Universitas Muria Kudus, Gondangmanis, Bae,
Kudus, Indonesia 59327
[c] Department of Mathematics, Faculty of Science and Mathematics, Diponegoro University, Jl. Prof. Soedarto, S.H.,
Tembalang, Semarang, Indonesia 50275
[d] Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Jl. Prof. Soedarto, S.H.,
Tembalang, Semarang, Indonesia 50275

Abstract

This study examines research trends regarding the prediction of academic achievement using machine learning. Research in the field of academic achievement is currently continuing to develop, but has not been explored comprehensively in a bibliometric context. The visualization provided includes a map of publication development using machine learning methods based on country, analysis of bibliographic pairs and keywords used. To find out the visualization results, bibliographic analysis was used using VOSviewer. The data used in this analysis were 76 articles collected from the Scopus database from 2018-2023. From the results of the analysis, it is known that research related to academic achievement still shows a growing trend in publications in the field of discussion of factors or predictors that influence academic achievement as well as research that proposes or evaluates models for predicting academic achievement. The research results show that although machine learning techniques such as Random Forest and Support Vector Machine are often used in academic achievement prediction research. Future research could consider developing a more adaptive and comprehensive approach regarding the contribution of specific factors that influence the accuracy of more in-depth prediction models in this field.

Keywords: prediction; academic achievement; machine learning; bibliometric review.

1. Introduction

Academic performance refers to a student's achievement in learning activities, including the final exam results attained during their educational journey. Academic achievement also serves as a measure of the performance of students, teachers, or schools in achieving the educational goals set within a specific period (Bayar & Kurt, 2021). Academic achievement heavily depends on classroom results, which indicate how well students have met the learning objectives (Yan, 2020). The low academic achievement can be caused by various factors such as the suitability of interests, talents, and previous educational history with the study program being taken at this time. The low academic achievement can be anticipated if students are able to identify various factors that cause weaknesses so that support in better academic achievement can be achieved (Hoffait & Schyns, 2017). For higher education institutions, academic achievement is a crucial factor that influences the achievement of educational goals.

Machine learning is a method used to analyze data by automatically building a model. This is possible because, in machine learning, a modeling system analyzes and learns from data by identifying patterns within the available data. To obtain these patterns, machine learning algorithms generate a mathematical model based on sample data (training data).

Machine learning is also a computer program that learns from experience with specific tasks, where performance improvement can be measured (Tom & Mitchell, 1997). In this study, bibliographic analysis is used to determine the relationship, application and trends of machine learning in predicting academic achievement. The data used in this bibliographic analysis were 76 articles collected from the Scopus database from 2018-2023.

Although various studies have been conducted regarding the application of machine learning in academic achievement prediction, there are still gaps that have not been explored in depth. Previous studies tend to focus on the use of specific algorithms in predicting academic achievement, but provide less explanation regarding the contribution of specific factors that influence the accuracy of prediction models. This shows the need for a more

*) Corresponding author: fajar.nugraha@umk.ac.id

comprehensive study in mapping relationships, applications and research trends regarding research topics in the field of using machine learning to predict academic achievement.

Through this mapping, it is hoped that this research can provide more comprehensive insight into developments in the use of machine learning in predicting academic achievement and identifying future research opportunities. This study contributes to the development of more adaptive academic prediction approaches and enriches the literature in the context of research trends in the field of utilizing machine learning to predict academic achievement.

2. Research Method

This research uses bibliometric analysis methods to examine research trends and developments related to the use of machine learning in predicting academic achievement. Bibliometric analysis was chosen because this method allows in-depth mapping of scientific literature, including analysis of author networks, keyword trends, and geographic distribution of research. With this approach, the research is expected to provide a comprehensive understanding of publication patterns and the evolution of research topics related to machine learning in the context of academic achievement prediction.

2.1. Research Data

The data used in this bibliometric analysis study uses data collected in international publications sourced from the Scopus database (www.scopus.com) on the topic of using machine learning to predict academic achievement. The selection of the Scopus database as a data source was based on its reliability in providing complete and varied bibliometric information.

This search focused on publications published in the 2018–2023 time frame, as this period reflects the rapid development and application of machine learning in educational contexts.

Data collection was carried out through searches in the Scopus database using a search method using a combination of keywords TITLE-ABS-KEY ("academic achievement") AND TITLE-ABS-KEY (prediction) AND TITLE-ABS-KEY ("machine learning")) AND PUBYEAR > 2017 AND PUBYEAR < 2024 AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (SRCTYPE , "j") OR LIMIT-TO (SRCTYPE , "p")) AND (LIMIT-TO (PUBSTAGE , "final")).

The obtained data was then analyzed using VOSviewer to build and visualize bibliometric networks for mapping research trends (Pan et al., 2018).

2.2 Data Analysis Process

Bibliometric data analysis in this study was carried out using VOSviewer software, which is known to have the ability to build and visualize bibliometric networks (Jan van Eck & Waltman, 2018). Networks may include journal, author, or publication source information that can be built on citations, co-citations, or co-authorship relationships. VOSviewer can identify relevant search combinations with an integrated mapping and clustering approach to inform co-citation data and co-occurrence networks as well as to provide visualization of linkages between articles, bibliometric data networks, such as the number of citations or co-occurrence relationships between keywords and several functions other bibliometric visualizations.

The use of bibliometric analysis in the context of academic achievement prediction offers a new perspective on the trends and developments of this topic. Bibliometric analysis makes it possible to provide comprehensive mapping by combining co-author network analysis, co-citation, and keyword co-occurrence to provide a deeper understanding of collaboration patterns, research themes, and topic developments in a relatively recent time span (2018–2023).

3. Results and Discussion

3.1 Overview of Data Sources

In this study, data was collected from the Scopus database in June 2024, with publication years ranging from 2018 to 2023. Based on the search keywords, 76 articles related to the use of machine learning to predict academic achievement were obtained. Detailed information regarding the growth of article publications is presented in Table 1.

Table 1. The growth of research article publication in Scopus database.

Publication Year	Number of Articles	Percentage
2018	2	3 %
2019	5	7 %
2020	7	9 %
2021	11	14 %
2022	24	32 %
2023	27	36 %

Based on the development of the growth of international publications with the theme of discussing the use of machine learning to predict academic achievement, it shows that between 2018 - 2023 there will be relatively consistent growth. In this research, filters are used to display only final articles that will appear in search data. Based on tracking the growth in the number of articles, it proves that research on machine learning to predict academic achievement is

still a research theme that continues to develop until now.

3.2 Co-author Citation Network Analysis

Based on the co-author citation network, a total of 453 authors were identified and grouped into 3 clusters using a combined connection strength of 10 authors. **Cluster 1:** Baashar Y.; Alkawsi G.; Ali N.; Alhussian H.; Bahbouh H.T.; Bosch N.; Chen C.-H.; Yang S.J.H.; Weng J.-X.; Ogata H.; Su C.-Y.; Razaque A.; Alajlan A.M. **Cluster 2:** Alhazmi E.; Sheneamer A.; Kavitha R.K.; Rajan Krupa C.; Menezes J.I.; Yağcı M.; **Cluster 3:** Pallathadka H.; Wenda A.; Ramirez-Asís E.; Asís-López M.; Flores-Albornoz J.; Phasinam K.; Rodríguez-Hernández C.F.; Musso M.; Kyndt E.; Cascallar E. The formation of these clusters is visualized in Figure 1.
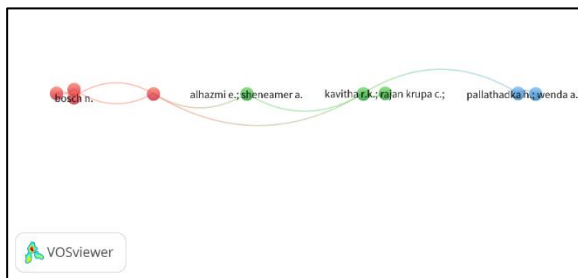

Figure 1. The co-author's citation network visualization.

The author with the highest number of citations is Yağcı M., with 66 citations and has a link strength of 5 in the co-author citation network, presented in table 2.

Table 2. Co-author's citation network data.

Author	Citations	Total Link Strength
Yağcı M.	66	5
Baashar Y. et al.	14	4
Alhazmi E.; Sheneamer A.	1	2
Bosch N.	9	2
Chen C.-H. et al.	15	2
Kavitha R.K. et al.	0	2
Razaque A.; Alajlan A.M.	6	2
Rodríguez-Hernández C.F. et al.	37	2
Pallathadka H. et al.	12	1

3.3 Co-author Country Network Analysis

In this section, 39 countries are identified through the bibliographic linkage of the co-author network with other countries. The first rank is Malaysia based on the strength of the link from a total of 39 countries followed by 5 other countries after the first rank, namely Saudi Arabia, Thailand, India, the United States and China. Of the 6 countries with the lowest link strength ranking, China is the lowest. Malaysia, Thailand, the United States and China have the same number of documents, each with 6 documents

published. Data on the number of articles, citations and link strength are presented in Table 3.

Table 3. Co-author's country network data.

Country	Number of Documents	Citations	Total Link Strength
Malaysia	6	58	52
Saudi Arabia	5	31	51
Thailand	6	19	20
India	8	21	19
United State	6	55	2
Cina	6	61	0

From the results of this analysis, it appears that Malaysia plays an important role in research on predicting academic achievement using machine learning which is known from the number of publications and the strength of the network between authors. This marks a shift in research center from Western countries to Asia, especially on the topic of applying machine learning in education, especially in the field of predicting academic achievement. Figure 2 is a visualization of the co-author's country network.
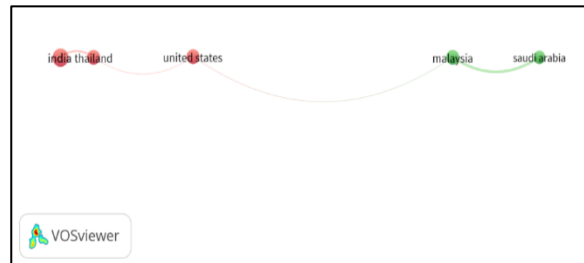

Figure 2. The co-author's country network visualization.

3.4 Most Frequent Keyword Analysis

Based on the keyword search used, the keywords that often appear used by the author are "machine learning" with a total of 42 times, the keyword "academic achievement" is used 33 times and prediction is used 20 times. These three keywords are the main domains used in this study. Other keywords found in this search are also very relevant to the search for the field of research on the use of machine learning to predict academic achievement. A list of 15 keywords displayed in this search is shown in Table 4.

Visualization of the relationships between these keywords shows that the main topic that dominates research is the use of machine learning algorithms. It can be seen that several machine learning algorithms are used to predict academic achievement, including the Support Vector Machine algorithm, Random Forest and Decision Trees. Visualization of the relationship between frequently used keywords is shown in Figure 3.

Table 4. The list of frequently used keywords.

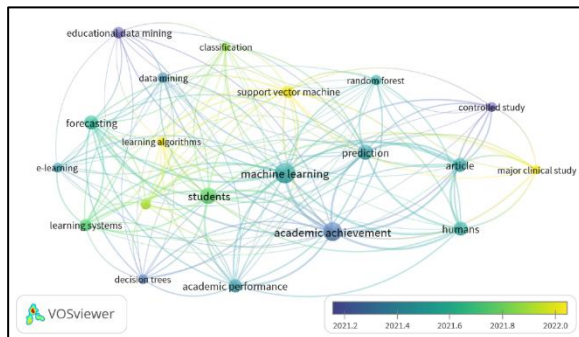| Keywords	Occurrence	Total Link Strength
Machine Learning	42	194
Academic Achievement	33	174
Student	24	132
Prediction	20	107
Forecasting	17	99
Human	17	96
Article	16	92
Academic Performance	14	81
Support Vector Machine	11	78
Learning Systems	11	65
Student Performance	8	50
Data Mining	8	48
Random Forest	8	45
Decision Trees	7	44
Controlled Study	7	42

Figure 3. Visualization of Frequently Used Keywords.

3.5 Identification of Academic Achievement Prediction using Machine Learning

Research that specifically uses the title or by using the keyword academic achievement, based on article searches there are research groups that review predictors that influence academic achievement. In this case, predicting students who have the potential to fail the final exam and are used to find students who are at high risk of low academic achievement early on. Student academic achievement is usually measured by the grade point average with a continuous numeric value. In this study, predicting academic achievement is defined as a classification task, specifically predicting whether a student's performance is excellent, good, or poor. Table 5 is a summary of the list of articles with discussions on factors or predictors that influence academic achievement and research that proposes or evaluates models for predicting academic achievement.

Table 5. The list of articles with discussion topic of academic achievement.

Discussion Topic	Article Title	Years
Factors or predictors that influence academic achievement	Prediction of academic achievement based on digital campus (Wang et al., 2018)	2018
	Predicting dropout using high school and first-semester academic achievement measures (Kiss et al., 2019)	2019
	Substance use initiation and the prediction of subsequent academic achievement (Meruelo et al., 2020)	2020
	Predicting at-risk university students based on their e-book reading behaviours by using machine learning classifiers (Chen et al., 2021)	2021
	Factors influencing youtube as a learning tool and its influence on academic achievement in a bilingual environment using extended information adoption model (IAM) with ML prediction (Abu-Taieh et al., 2022)	2022
	Student performance prediction model for predicting academic achievement of high school students (Nuankaew & Nuankaew, 2022)	2022
	Predictive models for dropout rates affected by covid-19 using classification and feature selection techniques (Nasa-Ngium et al., 2023)	2023
	Predicting academic success of autistic students in higher education (Bakker et al., 2023)	2023
Propose or evaluate models	Prediction of academic achievement based on digital campus (Wang et al., 2018)	2018
	Predicting at-risk university students based on their e-book reading behaviours by using machine learning classifiers (Chen et al., 2021)	2021
	Machine learning bias in predicting high school grades: a knowledge perspective (Costa-Mendes et al., 2021)	2021
	Interpretability analysis of academic achievement prediction based on machine learning (Yang & Wang, 2021)	2021
	Predictive models for dropout rates affected by Covid 19 using classification and feature selection techniques (Nasa-Ngium et al., 2023)	2023

Identification of factors that cause school dropout	Predicting dropout using high school and first-semester academic achievement measures (Kiss et al., 2019)	2019
	Educational data mining: prediction of students' academic performance using machine learning algorithms (Yağcı, 2022)	2022
	Student achievement prediction using deep neural network from multi-source campus data (Li et al., 2022)	2022
	Predictive models for dropout rates affected by covid-19 using classification and feature selection techniques (Nasa-Ngium et al., 2023)	2023

Regarding the use of machine learning methods, Random Forest and Support Vector Machine remain the algorithms most commonly used in predicting academic achievement, however there is an increase in the application of other alternative algorithms such as Decision Trees and Deep Neural Networks which indicates a diversification of the application of algorithms in research approaches in this topic.

4. Authors's Opinion

4.1 Research Trends

Based on the results of the analysis of the number of article publications, there has been an increase in publications on academic achievement predictions using machine learning in the last five years. This reflects the importance of research on predictive analysis in education. The focus of this research is not only on predicting the final results of students' academic achievement, but also on how to prevent and intervene early on for at-risk students so that they can anticipate failure during their studies.

The Random Forest and Support Vector Machine algorithms are known to be the most frequently used algorithms. Both of these methods are quite reliable in handling large data with complex features and are able to provide high predictive performance in classification tasks. Several studies have also begun to explore deep learning, but its application is still limited and has not dominated this research theme.

Based on geographical data, Asian countries such as Malaysia, Thailand, and India occupy strategic positions in research contributions, both in terms of the number of publications, so that there are still opportunities for more inclusive research collaborations in the future.

4.2 Study Limitations

In this bibliometric analysis study, the data used uses the Scopus database. Future research can increase the amount of data analyzed by utilizing various publication data databases and increasing the search period, currently using a publication period of five years from 2018-2023. While reflecting recent developments, it may not fully capture long-term trends in the use of machine learning in the prediction of academic achievement. Future studies could extend the analysis period to understand the evolution of these trends more comprehensively.

5. Conclusion

A bibliometric analysis study using the research field of using machine learning to predict academic achievement shows that the number of publications continues to increase and it can be seen that Malaysia has the highest number of articles cited in the 2018-2023 period. There is a research group that examines predictors that influence academic achievement which are used to predict students who have the potential to fail the final exam and to anticipate students who are at high risk of experiencing low academic achievement. Meanwhile, from searching keywords related to the application of machine learning algorithms, it can be seen that Random Forest and Support Vector Machine are algorithms that are often used to predict academic achievement using machine learning. Overall, this research has succeeded in providing a comprehensive picture of developments and research trends in the field of using machine learning for the prediction of academic achievement. Although there are still several limitations, this research opens up new opportunities for the development of more adaptive and contextual prediction models, and provides clear directions for future research in an effort to optimize the use of machine learning in education.

References

Abu-Taieh, E., Alhadid, I., Masa'deh, R., Alkhawaldeh, R. S., Khwaldeh, S., & Alrowwad, A. (2022). Factors Influencing YouTube as a Learning Tool and Its Influence on Academic Achievement in a Bilingual Environment Using Extended Information Adoption Model (IAM) with ML Prediction—Jordan Case Study. *Applied Sciences (Switzerland)*, *12*(12). https://doi.org/10.3390/app12125856

Bakker, T., Krabbendam, L., Bhulai, S., Meeter, M., & Begeer, S. (2023). Predicting academic success of autistic students in higher education. *Autism*, *27*(6), 1803–1816. https://doi.org/10.1177/13623613221146439

Bayar, M. F., & Kurt, U. (2021). Effects of Mobile Learning Science Course on Students' Academic Achievement and Their Opinions about the Course. *Science Education*

International, *32*(3), 254–263. https://doi.org/10.33828/sei.v32.i3.9

Chen, C.-H., Yang, S. J. H., Weng, J.-X., Ogata, H., & Su, C.-Y. (2021). Predicting at-risk university students based on their e-book reading behaviours by using machine learning classifiers. In *Australasian Journal of Educational Technology* (Vol. 2021, Issue 4). https://doi.org/10.14742/ajet.6116

Costa-Mendes, R., Cruz-Jesus, F., Oliveira, T., & Castelli, M. (2021). Machine learning bias in predicting high school grades: A knowledge perspective. *Emerging Science Journal*, *5*(5), 576–597. https://doi.org/10.28991/esj-2021-01298

Hoffait, A. S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*, 1–11. https://doi.org/10.1016/j.dss.2017.05.003

Jan van Eck, N., & Waltman, L. (2018). *VOSviewer Manual*. Retrieved from https://www.vosviewer.com

Kiss, B., Nagy, M., Molontay, R., & Csabay, B. (2019). Predicting Dropout Using High School and First-semester Academic Achievement Measures. *The 17th IEEE International Conference on Emerging ELearning Technologies and Applications*, 383–389. https://doi.org/10.1109/ICETA48886.2019.9040158

Li, X., Zhang, Y., Cheng, H., Li, M., & Yin, B. (2022). Student achievement prediction using deep neural network from multi-source campus data. *Complex and Intelligent Systems*, *8*(6), 5143–5156. https://doi.org/10.1007/s40747-022-00731-8

Meruelo, A. D., Castro, N., Nguyen-Louie, T., & Tapert, S. F. (2020). Substance use initiation and the prediction of subsequent academic achievement. *Brain Imaging and Behavior*, *14*(6), 2679–2691. https://doi.org/10.1007/s11682-019-00219-z

Nasa-Ngium, P., Nuankaew, W. S., Phanniphong, K., Jeefoo, P., & Nuankaew, P. (2023). Predictive Models for Dropout Rates Affected by COVID-19 Using Classification and Feature Selection Techniques. *International Journal of Engineering Trends and Technology*, *71*(7), 349–356. https://doi.org/10.14445/22315381/IJETT-V71I7P233

Nuankaew, P., & Nuankaew, W. S. (2022). Student Performance Prediction Model for Predicting Academic Achievement of High School Students. *European Journal of Educational Research*, *11*(2), 949–963. https://doi.org/10.12973/EU-JER.11.2.949

Pan, X., Yan, E., Cui, M., & Hua, W. (2018). Examining the usage, citation, and diffusion patterns of bibliometric mapping software: A comparative study of three tools. *Journal of Informetrics*, *12*(2), 481–493. https://doi.org/10.1016/j.joi.2018.03.005

Tom, M., & Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Wang, Z., Zhu, X., Huang, J., Li, X., & Ji, Y. (2018). Prediction of Academic Achievement Based on Digital Campus. *The 11th International Conference on Educational Data Mining*, 266–272. Retrieved from https://educationaldatamining.org/

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z

Yan, Z. (2020). Self-assessment in the process of self-regulated learning and its relationship with academic achievement. *Assessment & Evaluation in Higher Education*, *45*(2), 224–238. https://doi.org/10.1080/02602938.2019.1629390

Yang, J., & Wang, H. (2021). Interpretability Analysis of Academic Achievement Prediction Based on Machine Learning. *Proceedings - 11th International Conference on Information Technology in Medicine and Education, ITME 2021*, 475–479. https://doi.org/10.1109/ITME53901.2021.00101