# Leveraging Machine Learning to Predict Academic Success: A Literature Review and Bibliometric Trend Analysis

### Abstract

### This study examines the application of machine learning (ML) for predicting academic success through a literature review and bibliometric trend analysis. Using the Web of Science (WOS) database and the PRISMA framework, 206 articles were analyzed to identify key research themes, publication trends, and methodological advancements. Results reveal a growing focus on ML’s potential to improve educational outcomes, with an emphasis on predictive analytics and performance modeling. Despite these advances, challenges such as inconsistent metadata, limited database coverage, and overuse of ML-related terminology persist.

### Keywords

data mining; machine learning; academic success; educational analysis; systematic literature review; bibliographical analysis

## Introduction

In today’s complex world, predicting academic success has become a key focus in education research (Guanin-Fajardo et al., 2024). As the demand for skilled professionals grows, understanding what helps students succeed is more important than ever. This has led to a surge in studies aiming to predict academic outcomes, such as graduation rates and the likelihood of student dropouts, or grade point average itself.

Student success is a key indicator of educational quality, and York et al., (2015) define academic success as a sum of six core components: academic achievement, satisfaction, skill acquisition, persistence, learning objectives, and career success, though it is often reduced to grade point average alone (Alyahyan & Düştegör, 2020). This overlaps with what students perceive as academic success, i.e. a combination of outcomes including grades and more holistic outcomes of personal development and achievements (Lynam et al., 2024). Academic success varies significantly among individuals (Schillereff et al., 2023), and is shaped by a combination of personal factors (Acosta-Gonzaga, 2023) such as self-regulation, motivation, and self-esteem; environmental factors (Edgerton & McKechnie, 2023) like socioeconomic status, school environment, and peer support; and lifestyle factors (Kassaw & Demareva, 2023) including dietary habits, sleep, and stress management.

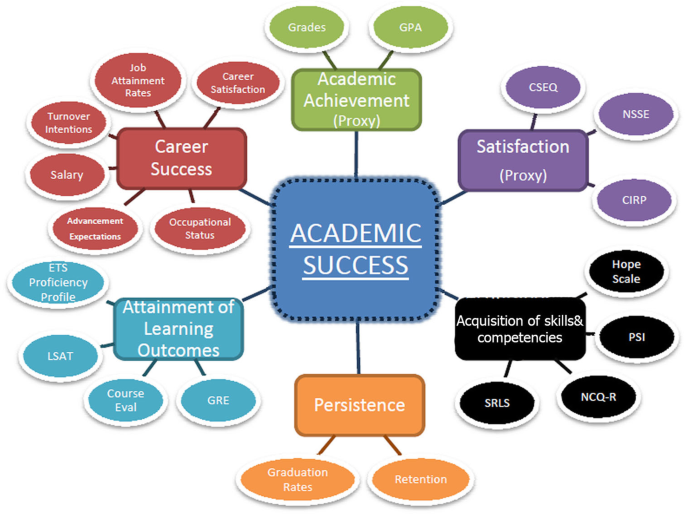


Figure 1: Definition of academic success (York et al., 2015).

Performance prediction has been studied as both a classification and regression problem, each with its own advantages and disadvantages (Alshanqiti & Namoun, 2020). A wide variety of statistical algorithms have been employed in the literature (Zhao et al., 2021). However, comparing the performance of these algorithms is challenging due to the multisourced nature of the data (Zhao et al., 2021).

One of the most promising tools in this area of prediction is machine learning (ML; Balaji et al., 2021; Jin, 2023). ML allows researchers to analyze large amounts of data to find patterns and make predictions, and in contrast to traditional statistical models, ML reduces bias, offers flexibility, and provides more robust models (Bregant et al., 2025; Hilbert et al., 2021). This technology can help identify students who might struggle, giving educators the chance to offer support early on. By using data like academic history and student behavior and, more recently, even online learning behaviors where open-access data is increasingly available (Liang et al., 2023), ML models can provide insights that help improve student outcomes. Figure 2 sketches, how Jin (2023) employed four ML algorithms for predicting academic success by accurately identifying the varying impacts of student, parent, and school efforts on exam performance.

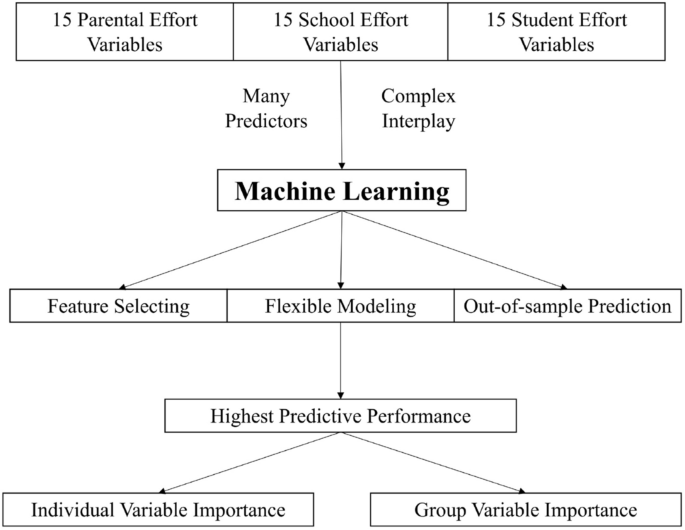


Figure 2: Example of machine learning modeling for student performance prediction by Jin (2023).

As we have seen, many studies (see Alyahyan & Düştegör (2020) section *Influential factors in predicting academic success*) have explored different factors that contribute to academic success, such as personal effort, family support, and the learning environment, and are summed in Table 1. These factors are crucial and have been widely researched, helping schools and educators develop strategies to support students.

|  |  |
| --- | --- |
| Factor Category | Factor Description |
| Prior Academic Achievement | Pre-university data: high school background (i.e., high school results), pre-admission data (e.g., admission test). University data: semester GPA or CGPA, individual course letter marks, and individual assessment grades. |
| Student Demographics | Gender, age, race/ethnicity, socioeconomic status (i.e., parents' education and occupation), place of residence/traveled distance, family size, and family income. |
| Students' Environment | Class type, semester duration, type of program. |
| Psychological | Student interest, behavior of study, stress, anxiety, time of preoccupation, self-regulation, and motivation. |
| Student E-learning Activity | Number of logins, number of tasks, number of tests, assessment activities, number of discussion board entries, number/total time material viewed. |

Table : Grouping and description of the most influential factors on the prediction of students’ academic success (Alyahyan & Düştegör, 2020).

Most cited studies in the field use various algorithms such as logistic regression, decision trees, random forests, support vector machines, XGBoost, and ensemble methods, with logistic regression and random forest often yielding the best performance, as demonstrated by Cankaya et al., (2024); Jang et al., (2022), while ensemble methods like GLMnet, as highlighted by (Bertolini et al., 2021) and wrapper-based feature selection techniques combined with classifiers like k-NN and SVM, as seen in (Abdelkader et al., 2022), are also widely employed to enhance predictive accuracy and interpretability, particularly through explainable AI (XAI; methods and techniques that make the outputs of machine learning models understandable to humans, enabling transparency and trust) techniques such as SHAP and LIME. However, there is a noticeable gap in bibliometric analyses within this field. Understanding the research trends, publication patterns, and thematic developments through a bibliometric approach can provide valuable insights.

Bibliometric analysis is a quantitative method used in systematic literature reviews to assess research productivity and impact (Lim & Kumar, 2024). Many seminal guides on this topic have been published, e.g., by Donthu et al., (2021); Lim & Kumar, (2024), focusing on performance metrics such as publication counts, citation analysis, authorship trends, keyword clustering, nomological networking, knowledge gaps, coverage of research, and so on. This technique leverages big data and software tools (e.g., *Gephi*, *VOSviewer*, or other already established statistics library tools such as *bibliometrics* package in *R*) to provide an objective overview of a research domain.

The present review study is focused on research about ML methods for predicting academic success and therefore opens the following research questions:

RQ1: What are the primary research themes and trends in the domain of academic success prediction using machine learning, based on the analysis of abstracts and keyword co-occurrence?

RQ2: How do publication patterns and cited reference counts vary ~~across different publishers and~~ over time in the research on academic success prediction?

## Methods

### Literature search

A preliminary analysis was conducted across *SCOPUS, Web of Science (WOS), Taylor & Francis* databases, and register *Digital Portal of the author(s)’ university*. A comparison between the databases and registries (using the tool specified in *data analysis* section) yielded a high degree of overlap in studies related to the prediction of academic success using machine learning techniques. Despite WOS having the fewest search results, it was chosen for the final bibliometric study due to it being the most widely used database of research publications and citations (Birkle et al., 2020). Articles from all selected databases and registries were accounted for in the review based on their relevance.

For the final report presented in this article, the final search in the WOS core collection was conducted in January 2025. The search string applied was: *("machine learning" OR "ML") AND ("academic success" OR "academic performance" OR "student success") AND ("educational sciences" OR "education") AND (“prediction”)*. As the PRISMA (PRISMA-P Group et al., 2015) flowchart in Figure 3 indicates, the search results in 307 matches, 101 of which were excluded from the bibliometric study.

Records identified from\*:

Databases:

WOS (*n* = 307)

TAYLOR & FRANCIS (*n*=1158)  
SCOPUS (*n* = 504)

University registers (*n* = 785)

Records removed *before screening*:

Records not in English (*n* = 3)

Records not articles (*n* = 86)

Records not in WOS (*n* =2447) – containing duplicates from other databases

Records published elsewhere than journals (*n* = 0).

**Identification**

Records excluded\*\*

Duplicate records (*n* = 1)

Records not deemed suitable (*n*=5) – not appropriate from statistical standpoint

Records screened based on title, abstract and key words

(*n* = 212)

**Screening**

Articles included in bibliographical analysis

(*n* = 206)

**Included**

Figure 3: PRISMA diagram. PORAVNAVA!

Inclusion and exclusion criteria were applied to ensure the relevance and quality of the data. Articles were included if they were written in English, published between 2015 and 2025, and classified as peer-reviewed articles (excluding book chapters, conference papers, and non-peer-reviewed works). Both qualitative and quantitative studies were considered. Articles with expressions of concern or retractions were excluded to maintain the validity of the data. The years 2015 and 2025 were omitted from the analysis for specific reasons: The results from 2015 were not included as the thematic focus during that year was still fairly new and in the early stages of progression (only 1 article), making the data less relevant for current trends. Additionally, 2025 was excluded as the data from that year is still new and scarce, limiting its impact on the overall analysis (only 4 articles).

### Data analysis

The final selection of 212 articles was analysed using *Python* programming language to construct and visualize the bibliometric network. Despite *Python* already having a library *pyBibX* designated for bibliometric and scientometric analysis we chose to develop our own solution, which is openly available (Bregant, 2025), and can help future studies. The researchers analyzed eligible studies following a predefined coding protocol. The descriptive coding covered various study attributes, such as the publication year, title, keywords, abstract, and citation count, among others.

## Results

Firstly, we made an extensive qualitative literature review of the 10 most cited articles from systematic literature search. The 10 articles, published between 2020 and 2024 and cited between 93 and 143 times, collectively explore the application of ML and educational data mining (EDM) across diverse educational settings, including higher education, K-12, and online learning, with a strong focus on early performance prediction, student retention, and online learning satisfaction. Predominantly categorized under "Education & Educational Research" and "Computer Science, Information Systems" in WOS, these studies, mainly published by Springer, MDPI, and others, employ a variety of ML algorithms, such as logistic regression, random forest, support vector machines, XGBoost, and ensemble methods, to address challenges like identifying at-risk students, improving degree completion rates, and assessing satisfaction in online learning environments. XAI techniques, including SHAP and LIME, are frequently utilized to enhance model interpretability, enabling educators and policymakers to derive actionable insights. Key findings highlight the effectiveness of ensemble methods and the importance of feature selection in improving predictive accuracy, with logistic regression and random forest often yielding the best performance. Conducted in diverse regions, such as the USA, Spain, and global online settings, these studies underscore the potential of ML to personalize interventions, reduce attrition, and optimize educational strategies, particularly in higher education and online learning contexts.

For word cloud analysis, the *CountVectorizer* module was used in collaboration with *Latent Dirichlet Allocation* (LDA) to extract the most frequently used words from keywords and abstracts. Common English stopwords (e.g., “the”, “and”, “is”) were removed to minimize noise, along with words appearing in more than 95% of the documents and those appearing in fewer than two documents.

As depicted in Figure 4, the word clouds for each year reveal key themes and shifts in research focus over time. In 2017, terms such as "prediction," "classification," and "attrition" were prominent, reflecting an early emphasis on predictive analytics and student outcomes. By 2018, there was a noticeable focus on "machine learning" and "performance," indicating a growing interest in applying machine learning techniques to educational data. This trend continued through subsequent years, with "student," "learning," and "data mining" consistently appearing as central topics. By 2024, terms such as "educational," "prediction," and "performance" highlight ongoing research into predictive models for student success and academic performance. Overall, the progression of themes suggests a sustained and evolving interest in leveraging data analytics and machine learning to enhance educational outcomes. Interestingly, 2018 yielded no significant results, which may indicate a gap in research publications or data availability for that year.

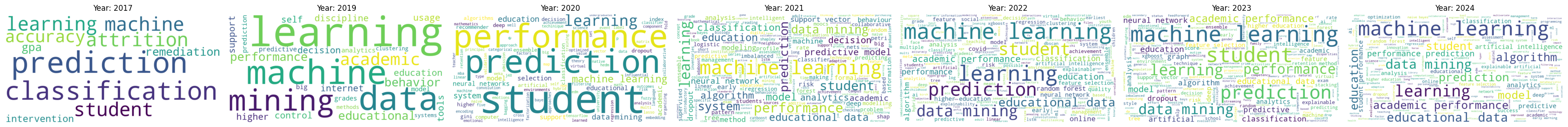


Figure : Word cloud analysis through the years.

A collage of words

Description automatically generated

As we can observe in Figure 5, the distribution of cited reference counts among the articles presents a notable variation, illustrating a substantial spread within the interquartile range. The majority of articles receive a moderate number of citations, but several outliers represent publications with significantly higher citation counts. This indicates that a few studies attract disproportionate scholarly attention. The observations corroborate the methodology used in WoS Citation Reports (*Web of Science Core Collection: Finding the Average Number of Citations per Article in a Journal*, 2022), where the average citations per item provide an essential measure of scholarly impact.

A diagram of a box plot

Description automatically generated

Figure : Cited reference counts. Averaged citated reference count was 47.8, with standard deviation 21.0.

As illustrated in Figure 6, the number of publications in this field has been steadily increasing over the years, reflecting growing interest within the academic community. However, it is important to note that not all publications explicitly use "machine learning" as a keyword, as depicted in Figure 7. This discrepancy suggests that a significant portion of the research may be related to ML concepts without directly referencing the term, which highlights the diverse terminology used in the field and the broad scope of research topics that intersect with ML methodologies.

A graph of blue bars

Description automatically generated with medium confidence

Figure : Number of publications by year.

A graph of a number of publications

Description automatically generated

Figure : Number of publications per year with highlighted publications using “machine learning” as a keyword.

The growing popularity of the subject is also reflected in the increasing total number of citations over the years, indicating a rising impact and recognition within the academic community. As more research is conducted and published in this field, the cumulative citation count continues to climb, as seen on Figure 8. Interestingly, while the total number of citations has seen significant growth, the average number of citations per year has remained relatively consistent, as pictured in Figure 9. This suggests that while the volume of research output is expanding, the influence and reach of individual publications are maintaining a steady rate of scholarly engagement and reference over time.

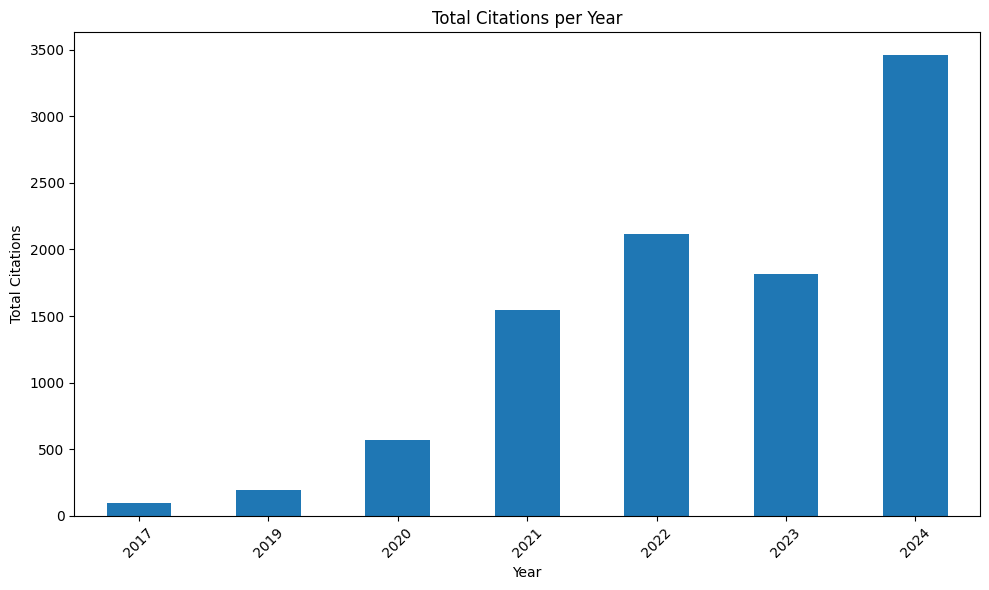


Figure : Total citations.

A graph of blue bars

Description automatically generated with medium confidence

Figure : Average citations per year.

~~DODAMO ŠE WOS KATEGORIJE, IN REGRESIJA~~

## Discussion and conclusion

The application of ML in predicting academic success has emerged as a transformative area of research, bridging the fields of education, data science, and psychology. This study highlights the increasing role of ML in predicting academic success, as reflected in the growing volume of research and citations in recent years. The bibliometric trend analysis reveals a sustained interest in this field, with a noticeable shift toward more sophisticated data-driven approaches to educational analysis. The findings underscore the potential of ML models to provide actionable insights for improving student outcomes, identifying at-risk students, and enhancing the overall quality of education.

One of the key observations is the diversification of research themes over time. Early studies predominantly focused on basic predictive analytics and classification techniques. However, recent work has expanded to incorporate more complex models and interdisciplinary approaches, integrating insights from psychology (Burman et al., 2021), sociology (Davies et al., 2021), and others. This trend suggests a maturing field that is increasingly leveraging the power of ML to address multifaceted challenges in education. This reflects a growing consensus on the importance of predictive models in understanding and enhancing academic performance. Additionally, the rising number of publications and citations indicates a broader acceptance and recognition of ML's value in educational research. With rapid technological advancements, AI will undoubtedly play a crucial role in educational sciences (Istenič Starčič, 2019).

However, it is crucial to recognize that the terms "ML" and "AI" are increasingly overused and misused (Chen & Lin, 2024), often serving as buzzwords that lead to misconceptions about their capabilities and limitations. This loose application dilutes the precision of academic discourse, making it difficult to categorize and synthesize relevant research effectively. Many studies claiming to use ML or AI lack rigorous methodology, while others employing related concepts may not explicitly reference these terms due to varying terminologies. This inconsistency hampers research discoverability (Zhao et al., 2023) and risks overshadowing genuinely innovative work, highlighting the need for more standardized and precise terminology in academic publications.

Furthermore, the choice of databases significantly influences the scope and comprehensiveness of bibliometric analyses. Different databases, such as SCOPUS, WOS, and specialized educational repositories, offer varying levels of coverage, indexing, and citation data (Mongeon & Paul-Hus, 2016), which can affect the identification of key studies and trends (Delgado‐Quirós et al., 2024). The discrepancies between databases highlight the importance of selecting appropriate sources to ensure a balanced and accurate representation of the research landscape.

In conclusion, this review underscores the significant impact of machine learning in predicting academic success and its growing prominence in educational research. The findings highlight the potential of ML to revolutionize educational practices, but also call for greater methodological rigor, standardized terminology, and interdisciplinary collaboration to fully realize its benefits.

## Limitations and future directions

While this study provides valuable insights into the application of ML for predicting academic success, several limitations must be acknowledged. Firstly, the reliance on bibliometric data from the WOS database may introduce selection bias, as WOS might not capture the entirety of relevant publications, especially those published in regional or non-indexed journals. Similarly, the focus on specific keywords in the search strategy could have excluded studies employing alternative terminologies, thereby limiting the comprehensiveness of the analysis.

Another limitation pertains to the variability in the quality and consistency of metadata across different publications. Inconsistencies in keyword usage, authorship details, and citation practices can affect the reliability of bibliometric indicators. Furthermore, the dynamic nature of the field means that recent publications, particularly from 2025, may not yet have accumulated sufficient citations to accurately reflect their impact, leading to potential underestimation of their significance.

Future research should address these gaps by incorporating broader databases and more inclusive search strategies. Advanced text-mining techniques could uncover latent trends and themes, while longitudinal studies could track the evolution of ML applications in education over time.

Integrating bibliometric analysis with qualitative assessments would provide deeper insights into the relevance of key studies. Collaboration between researchers, educators, and policymakers is essential to ensure ML models are effectively applied to enhance educational outcomes and support student success.

## References

Abdelkader, H. E., Gad, A. G., Abohany, A. A., & Sorour, S. E. (2022). An Efficient Data Mining Technique for Assessing Satisfaction Level With Online Learning for Higher Education Students During the COVID-19. *IEEE Access*, *10*, 6286–6303. https://doi.org/10.1109/ACCESS.2022.3143035

Acosta-Gonzaga, E. (2023). The Effects of Self-Esteem and Academic Engagement on University Students’ Performance. *Behavioral Sciences*, *13*(4), 348. https://doi.org/10.3390/bs13040348

Alshanqiti, A., & Namoun, A. (2020). Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification. *IEEE Access*, *8*, 203827–203844. https://doi.org/10.1109/ACCESS.2020.3036572

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1), 3. https://doi.org/10.1186/s41239-020-0177-7

Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of Machine Learning Models towards Student Academic Performance Prediction: A Systematic Review. *Applied Sciences*, *11*(21), Article 21. https://doi.org/10.3390/app112110007

Bertolini, R., Finch, S. J., & Nehm, R. H. (2021). Testing the Impact of Novel Assessment Sources and Machine Learning Methods on Predictive Outcome Modeling in Undergraduate Biology. *Journal of Science Education and Technology*, *30*(2), 193–209. https://doi.org/10.1007/s10956-020-09888-8

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, *1*(1), 363–376. https://doi.org/10.1162/qss\_a\_00018

Bregant, B. (2025). *Statistics code of the article* [Repository]. Github. https://github.com/borbregant/Doktorat\_all/blob/main/raziskovalni\_seminar/obdelava.ipynb

Bregant, B., Doz, D., & Hudovernik, S. (2025). Factors influencing tandem learning in mathematics. *International Journal of Instruction*, *18*(1), 437–463.

Burman, I., Som, S., & Hossain, S. A. (2021). Model for Analyzing Psychological Parameters Recommending Student Learning Behavior using Machine Learning. *International Journal of Computing and Digital Systems*, *10*(1), 973–990. https://doi.org/10.12785/ijcds/100188

Cankaya, B., Roberts, R., Douglas, S., Vigness, R., & Oztekin, A. (2024). What postpones degree completion? Discovering key predictors of undergraduate degree completion through explainable artificial intelligence (XAI). *Journal of Marketing Analytics*. https://doi.org/10.1057/s41270-024-00290-6

Chen, J. J., & Lin, J. C. (2024). Artificial intelligence as a double-edged sword: Wielding the POWER principles to maximize its positive effects and minimize its negative effects. *Contemporary Issues in Early Childhood*, *25*(1), 146–153. https://doi.org/10.1177/14639491231169813

Davies, H. C., Eynon, R., & Salveson, C. (2021). The Mobilisation of AI in Education: A Bourdieusean Field Analysis. *Sociology*, *55*(3), 539–560. https://doi.org/10.1177/0038038520967888

Delgado‐Quirós, L., Aguillo, I. F., Martín‐Martín, A., López‐Cózar, E. D., Orduña‐Malea, E., & Ortega, J. L. (2024). Why are these publications missing? Uncovering the reasons behind the exclusion of documents in free‐access scholarly databases. *Journal of the Association for Information Science and Technology*, *75*(1), 43–58. https://doi.org/10.1002/asi.24839

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Edgerton, E., & McKechnie, J. (2023). The relationship between student’s perceptions of their school environment and academic achievement. *Frontiers in Psychology*, *13*, 959259. https://doi.org/10.3389/fpsyg.2022.959259

Guanin-Fajardo, J. H., Guaña-Moya, J., & Casillas, J. (2024). Predicting Academic Success of College Students Using Machine Learning Techniques. *Data*, *9*(4), 60. https://doi.org/10.3390/data9040060

Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D., & Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*, *9*(3), Article 3. https://doi.org/10.1002/rev3.3310

Istenič Starčič, A. (2019). Human learning and learning analytics in the age of artificial intelligence. *British Journal of Educational Technology*, *50*(6), 2974–2976. https://doi.org/10.1111/bjet.12879

Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students’ performance using machine learning and eXplainable AI. *Education and Information Technologies*, *27*(9), 12855–12889. https://doi.org/10.1007/s10639-022-11120-6

Jin, X. (2023). Predicting academic success: Machine learning analysis of student, parental, and school efforts. *Asia Pacific Education Review*. https://doi.org/10.1007/s12564-023-09915-4

Kassaw, C., & Demareva, V. (2023). Determinants of academic achievement among higher education student found in low resource setting, A systematic review. *PLOS ONE*, *18*(11), e0294585. https://doi.org/10.1371/journal.pone.0294585

Liang, G., Jiang, C., Ping, Q., & Jiang, X. (2023). Academic performance prediction associated with synchronous online interactive learning behaviors based on the machine learning approach. *Interactive Learning Environments*, 1–16. https://doi.org/10.1080/10494820.2023.2167836

Lim, W. M., & Kumar, S. (2024). Guidelines for interpreting the results of bibliometric analysis: A sensemaking approach. *Global Business and Organizational Excellence*, *43*(2), 17–26. https://doi.org/10.1002/joe.22229

Lynam, S., Cachia, M., & Stock, R. (2024). An evaluation of the factors that influence academic success as defined by engaged students. *Educational Review*, *76*(3), 586–604. https://doi.org/10.1080/00131911.2022.2052808

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

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

Schillereff, D., Clarke, L., Shuttleworth, E., & Alderson, D. (2023). Evaluating success in a changing academic landscape. *Earth Surface Processes and Landforms*, *48*(12), 2387–2394. https://doi.org/10.1002/esp.5634

*Web of Science Core Collection: Finding the average number of citations per article in a journal*. (2022). Clarivate. https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Core-Collection-Finding-the-average-number-of-citations-per-article-in-a-journal?language=en\_US

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and Measuring Academic Success. *Practical Assessment, Research, and Evaluation*, *20*(1), 5. https://doi.org/10.7275/HZ5X-TX03

Zhao, L., Chen, K., Song, J., Zhu, X., Sun, J., Caulfield, B., & Namee, B. M. (2021). Academic Performance Prediction Based on Multisource, Multifeature Behavioral Data. *IEEE Access*, *9*, 5453–5465. https://doi.org/10.1109/ACCESS.2020.3002791

Zhao, L., Shen, C., Liu, M., Zhang, J., Cheng, L., Li, Y., Yuan, L., Zhang, J., & Tian, J. (2023). Comparison of Reporting and Transparency in Published Protocols and Publications in Umbrella Reviews: Scoping Review. *Journal of Medical Internet Research*, *25*, e43299. https://doi.org/10.2196/43299