

A Systematic Literature Review of a Machine Learning Model for Predicting Student Performance and Engagement in Higher Education

Nomathemba Dolores Kutyaauripo¹, Sibusisiwe Dube²

¹Student, National University of Science and Technology, Informatics and Analytics, Bulawayo, Zimbabwe

²Lecturer, National University of Science and Technology, Informatics and Analytics, Bulawayo, Zimbabwe

DOI: <https://doi.org/10.51244/IJRSI.2026.1309000150>

Received: 20 September 2026; Accepted: 25 September 2026; Published: 13 October 2026

ABSTRACT

The increasing availability of educational data has accelerated the application of machine learning (ML) to predict student outcomes and support data-driven educational decision-making. However, existing research varies considerably in algorithms, evaluation approaches, outcome variables, and geographical contexts. This study systematically analysed 17 research articles published between 2021 and 2025 that applied ML to predict student academic performance and engagement. The studies were examined in terms of prediction objectives, algorithms, reported performance, methodological approaches, and geographical contexts to identify dominant trends and research gaps. Random Forest (RF) emerged as the most frequently employed algorithm, appearing in 15 studies and frequently demonstrating strong reported performance. However, RF did not consistently outperform alternative algorithms. Some studies reported predictive accuracies approaching 100%; however, comparisons are constrained by differences in datasets, predictor variables, outcome definitions, validation procedures, and evaluation metrics. Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) were also frequently used. Academic performance was the predominant research focus (71%), while student engagement accounted for 29% of studies. The literature was geographically concentrated in Asia, with supervised learning approaches dominating. Emerging hybrid approaches combined clustering techniques, including Gaussian Mixture Models (GMM) and DBSCAN, with predictive modelling for student profiling. A significant gap was the limited integration of Explainable Artificial Intelligence (XAI), as most studies prioritised predictive accuracy over interpretability and actionable educational interventions. The findings indicate that while classical and ensemble ML models provide promising predictive capabilities, methodological heterogeneity, geographical concentration, and limited explainability constrain their practical utility. Future research should standardise evaluation practices, strengthen XAI integration, and expand research on engagement prediction and across diverse educational contexts to support systems linking prediction, explanation, and intervention.

Keywords: Machine Learning, Student Academic Performance, Student Engagement, Educational Data Mining.

INTRODUCTION

The integration of digital technologies into higher education has accelerated substantially in recent years, particularly following the COVID-19 pandemic. The increased adoption of learning management systems, online learning platforms, and other digital tools has generated large volumes of educational data that can provide valuable insights into students' learning behaviors and outcomes. At the same time, higher education typically requires greater learner autonomy and self-regulation, which may increase students' vulnerability to disengagement, poor academic performance, and eventual withdrawal. Consequently, institutions increasingly need reliable approaches to monitor and predict student engagement and academic performance to enable timely, targeted interventions.

The growing availability of educational datasets, together with advances in machine learning (ML), has created opportunities to develop predictive models that can identify patterns associated with student outcomes. ML-based predictive approaches can support the identification of students who may be at risk of poor academic performance or disengagement and facilitate the provision of personalized academic support (Daimi & Iqbal, 2023). Such approaches have the potential to strengthen student retention and completion by enabling institutions to intervene before academic difficulties become more severe (Matz et al., 2023).

Despite the growing application of ML in education, the existing literature remains diverse in terms of algorithms, prediction targets, reported performance, methodological approaches, and research contexts. Studies have examined different combinations of supervised, unsupervised, and hybrid ML techniques, while their reported predictive performance varies considerably across datasets and evaluation settings. Furthermore, academic performance appears to have received greater attention than student engagement, and the geographical distribution and contextual characteristics of existing studies remain uneven. These variations make it difficult to establish which approaches are most commonly employed, how their reported performance compares, and where important gaps remain in the literature.

A systematic synthesis of recent research is therefore necessary to consolidate existing evidence and identify emerging methodological and contextual trends. This study reviews research published between 2021 and 2025 on the application of ML to predicting student academic performance and engagement. Specifically, the study seeks to answer the following research questions:

RQ1: What machine learning models and algorithms are commonly used to predict student academic performance?

RQ2: What machine learning models and algorithms are commonly used to predict student engagement?

RQ3: What levels of predictive performance are reported for machine learning models used to predict academic performance and student engagement?

RQ4: What geographical and educational contexts have been represented in research on machine learning-based prediction of student academic performance and engagement?

METHODOLOGY

Table 2.1 shows the research protocol used in this work, with source databases Springer Nature Link and IEEE spanning the years 2021 to 2025. This SLR only accepts journal papers written in English. This survey excluded periodicals, books, SLR, and journal articles written in languages other than English.

Table2.1. Research Protocol

Protocol Element	Research Item
Digital library	Springer Nature Link, IEEE Xplore
Time Interval	2021 - 2025
Inclusion Criteria	Higher education Journal papers in English
Exclusion criteria	magazine, not higher education, books, systematic reviews, articles not written in English

This systematic literature review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and focused on studies published between 2021 and 2025. Relevant literature was identified primarily through SpringerLink and IEEE Xplore. These databases were retained because the search strategy produced relevant studies aligned with the scope of the review. Other academic databases were also explored during the search process; however, the search string used did not

produce relevant results for inclusion. This may reflect differences in database indexing, search syntax and search functionality. The search strategy was developed around four key concepts: machine learning, student engagement, academic performance, and higher education, together with related synonyms to maximize the retrieval of relevant studies. In SpringerLink, the following search string was applied: (“Machine learning models” OR “ML models” OR “Predictive Models”) AND (“Prediction” OR “forecasting” OR “projecting”) AND (“Student” OR “Learner” OR “Trainee”) AND (“performance” OR “achievement” OR “attainment”) AND (“engagement” OR “participation” OR “responsiveness”) AND (“Higher Education” OR “College”). Equivalent search terms and Boolean operators were adapted to the syntax and search functionality of IEEE Xplore. The retrieved records were subsequently screened against predefined inclusion and exclusion criteria to identify studies relevant to machine learning-based prediction of student academic performance and engagement.

The article search began with 1341 publications (IEEE Xplore=622 and Springer Nature=719). The initial step of scanning the abstracts removed 1161 records that did not directly explore the issues of e-Learning platforms or present any AI-driven ideas for overcoming the deficiencies of e-Learning platforms, leaving only 180 relevant publications. A full-text review of these 180 publications excluded 163 irrelevant articles that presented no findings related to the goal of this study, leaving only 17 relevant studies. The SLR results are discussed in the Results section.

RESULTS AND DISCUSSIONS

Geographical Distribution of the Reviewed Studies

The research is heavily concentrated in Asia—specifically China, India, Pakistan, Turkey, Iraq, Saudi Arabia and UAE (listed as Asia in the table). European representation is limited to Slovakia, Portugal and Germany. Collectively, these studies account for 13 of the 17 studies (76%), indicating that Asia represents the dominant geographical context for research on machine learning-based prediction of student outcomes.

European countries accounted for three studies (17.6%), while two studies (11.8%) were conducted in Saudi Arabia and the United Arab Emirates. One study (5.9%) did not specify its geographical context. The distribution therefore demonstrates that research in this area is concentrated in a relatively limited number of geographical regions, with comparatively less evidence from other parts of the world, particularly Africa, Latin America, and other developing-country contexts.

This geographical concentration has implications for the generalizability of existing findings. Student engagement and academic performance are influenced by institutional, cultural, socioeconomic, technological, and educational factors. Consequently, predictive models developed using data from predominantly Asian or other specific contexts may not necessarily perform similarly when applied to institutions with different educational systems, student populations, technological infrastructures, and socioeconomic conditions. The limited representation of diverse geographical contexts therefore constitutes an important research gap and highlights the need for further studies involving underrepresented regions.

Table3.1. Geographical distribution of studies included in the systematic review (2021–2025).

Geographic Context	Number of Studies
Asia	13
Europe	3
Unspecified	1

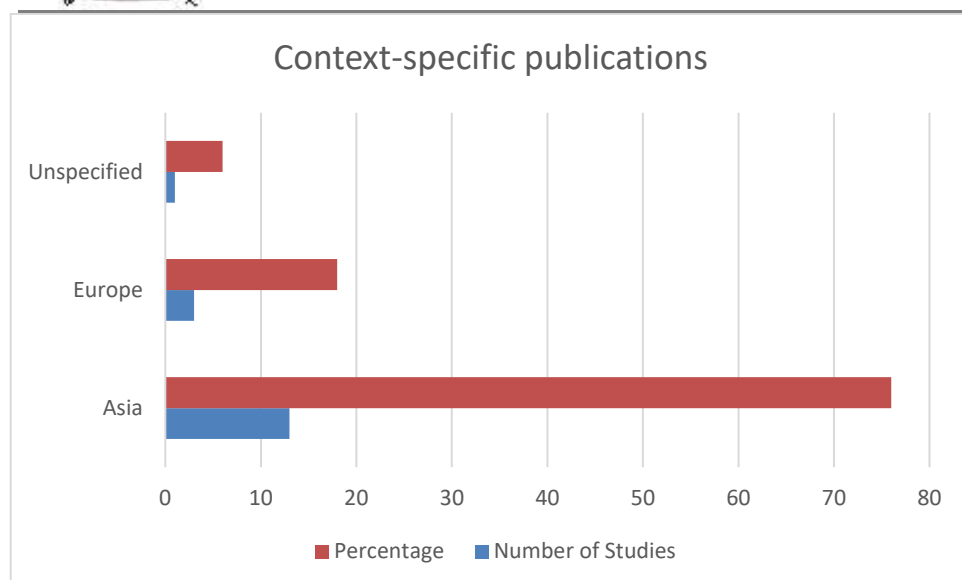


Figure 3.1 Geographical distribution of studies included in the systematic review (2021–2025).

Figure 3.1 depicts the distribution of the studies included in this SLR across regions and shows the highest representation in Asia (76%). This concentration may reflect several factors. First, Asian countries have increasingly adopted and implemented emerging technologies to support educational activities. Countries such as China and India have also invested in technology for education and research, creating opportunities for researchers to explore innovative solutions using machine learning. There is also growing interest in using data-driven insights to improve student outcomes. These developments may have contributed to increased research activity in this field.

However, the geographical concentration of the reviewed studies is important when interpreting the findings. Educational systems differ in assessment practices, student populations, institutional resources and access to technology. Models developed and evaluated within one geographical or institutional context may therefore not achieve similar performance in other settings. The limited representation of regions such as Africa and Latin America means that there is less evidence about how these models perform in these contexts. Further research across more diverse geographical and educational settings is therefore needed.

Models for Academic Performance

Supervised learning dominates. Random Forest (RF) is the most frequently used algorithm (appearing in ~80% of academic performance studies). Other common models include Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Trees (DT), Logistic Regression (LR), and Naive Bayes (NB). Ensemble methods (XGBoost, Extra Trees, AdaBoost, LightGBM) are also widely applied. Unsupervised methods (clustering like BIRCH, DBSCAN, GMM) are rarely used for direct prediction.

This is sensible for the following reasons. First, educational systems naturally produce labeled academic performance and engagement data that yield clearly defined targets such as final grades (GPA, pass/fail), exam scores, or binary engagement indicators (active/inactive). This makes supervised regression or classification directly applicable, and such an observation is consistent with existing literature. For example, Gul, Abbasi, & Wani (2025) used GPA as the regression target, achieving 99.93% accuracy with RF; Alturki, Cohausz, & Stuckenschmidt (2022) predicted academic performance using labeled course grades; and Albreiki, Habuza, & Zaki (2022) relied on assignment submission data and final marks as ground truth. This proves that researchers are taking advantage of institutions' archived student records with these labels and simply apply supervised models without costly annotation.

Secondly, supervised models (especially RF, SVM, ensembles) consistently achieve high accuracy (0.80-100)..A study by Yağcı (2022) found RF and Neural Networks achieved the highest F1-scores (0.746) among six

supervised algorithms. In addition, Gupta, Singhal, & Khattri (2024) reported RF at 100% accuracy for performance based on attendance, while Zhao et al. (2021) demonstrated SVM (0.889) and GBRT (0.877) outperforming other methods. Such literature confirms that high accuracy supports institutional decision-making (early alert systems, intervention targeting), justifying the dominance of supervised approaches

Thirdly, supervised learning enables direct benchmarks and comparability, which is why most studies compare multiple supervised algorithms (as seen in your table), to allow standardized performance evaluation using metrics like accuracy, precision, recall, F1, and RMSE. This comparability is a key driver. This was the case of a study by Brahim (2022), who compared several models, including random forest, to identify the best classifier. Similarly, Borges et al. (2025) compared 8 supervised/unsupervised models but only reported performance for the supervised ones; unsupervised (BIRCH, DBSCAN, GMM) were used for profiling, not predicting. Furthermore, Zou, Zhong, Du, & Yuan (2025) benchmarked 9 supervised regression models against each other while Ali, Ayazuddin, & Sanjaya (2025) explicitly compared DT, RF, GBooster, NB, and KNN to identify the best classifier.

Fourth, unsupervised methods are confined to exploratory roles; when they appear (clustering), they serve to discover patterns or label data before prediction, not as the final predictor. Onwuzurike and Kpogli (2025) demonstrated this view by applying unsupervised learning to classify and supervised learning to predict students' engagement and behavioral outcomes. Borges et al. (2025) similarly used unsupervised learning only for feature grouping, while prediction relied on supervised RF, XGB, DT, etc.

Fifth and lastly, the research goal of predicting specific student outcomes is a supervised task; that is why it is possible through supervised learning to provide educational institutions with actionable outputs such as "Which student will fail?" or "Who is disengaged?". This view is consistent with literature where supervised learning has been used to predict students' outcomes (Albahli, 2025) and academic performance (Marín et al., 2025; Alhazmi & Sheneamar, 2023). For example, Fu, Zhong, Qin, & Luo (2025) predicted engagement via SVM/KNN/NB/MLP (AHP 90.72%), Wang (2025) predicted student engagement as a binary class using DT, SVM, and RF (RF at 0.968) and Zhang et al. (2025) predicted engagement levels using 9 supervised algorithms, with RF reaching 0.8 accuracy.

Models for Student Engagement

RF and SVM are again the most common. This is consistent with existing literature, such as Wang (2025) directly compared DT, SVM, and RF for student engagement prediction, with RF achieving 0.968 accuracy. Similarly, Zhang et al. (2025) used RF, SVM, GBDT, AdaBoost, MLP, and KNN, reporting RF at 0.80 and SVM at 0.79. Furthermore, Ayouni, Hajjej, Maddeh, & Al-Otaibi (2021) employed DT, SVM, and ANN specifically for student engagement, and Fu, Zhong, Qin, & Luo (2025) used SVM, KNN, NB, and MLP for engagement, achieving 90.72% with AHP integration.

SVM is followed by DT, KNN, NB, and MLP. Gradient boosting variants (GBDT, AdaBoost, GBoost) also feature prominently. The focus shifts slightly toward classification of engagement levels rather than regression. This is evident in studies such as Ali, Ayazuddin, & Sanjaya (2025) compared DT (0.9606), RF (0.9611), GBooster (0.9578), NB (0.82), and KNN (0.71) for student engagement, and Fu et al. (2025) included KNN, NB, and MLP alongside SVM for engagement classification. Gradient boosting variants (GBDT, AdaBoost, GBoost): Zhang et al. (2025) specifically evaluated GBDT (0.79) and AdaBoost (0.79) for engagement, which is confirmed by Ali et al. (2025) tested GBooster, achieving 0.9578 accuracy for engagement prediction.

There is now a shift toward classification rather than regression. Table 3.2 depicts some engagement-focused studies in the table treat engagement as a categorical outcome (e.g., engaged vs. disengaged) rather than a continuous variable. Wang (2025), Zhang et al. (2025), and Ali et al. (2025) all frame engagement prediction as a classification task. This contrasts with academic performance studies (e.g., Gul et al., 2025; Zou et al., 2025) that frequently use regression to predict numerical GPA or scores.

Table3.2. Algorithms used in studies included in the systematic review (2021–2025).

Study	Algorithms Used	Focus
Wang (2025)	DT, SVM, RF	Classification
Zhang et al. (2025)	RF, SVM, GBDT, AdaBoost, MLP, KNN	Classification
Ali et al. (2025)	DT, RF, GBooster, NB, KNN	Classification
Fu et al. (2025)	SVM, KNN, NB, MLP	Classification
Ayouni et al. (2021)	DT, SVM, ANN	Classification

Reported Performance Levels

Performance Reported predictive performance varied across the reviewed studies, with accuracy ranging from approximately 70% to 100%. Random Forest was frequently among the higher-performing models, with reported values reaching 99.93% and 100%. However, it did not consistently outperform all alternative algorithms. SVM, Extra Trees and MLP also achieved strong performance in several studies, with reported values generally ranging from 0.88 to 0.98. For engagement prediction, Random Forest recorded performance between approximately 0.80 and 0.97, while SVM and Decision Tree generally showed moderate to high performance. These findings suggest that ensemble tree-based methods performed strongly across many of the reviewed studies, although performance varied according to the model and dataset used.

The strong performance of Random Forest may partly be explained by its use of multiple decision trees, which can reduce variance and improve performance when modelling complex and nonlinear relationships (Breiman, 2001). Extra Trees similarly introduces additional randomisation during tree construction, which can reduce correlation among individual trees (Geurts et al., 2006). SVM performs differently by identifying decision boundaries that maximise the margin between classes and can model nonlinear relationships through kernel functions (Cortes & Vapnik, 1995). MLP can also capture complex nonlinear relationships when sufficient data and appropriate model configuration are available (Hornik et al., 1989). Decision Trees offer a simpler and more interpretable structure but can be sensitive to overfitting when tree growth is not adequately controlled (Quinlan, 1993). Overall, the findings show strong reported performance for Random Forest and other nonlinear and ensemble approaches. However, no single algorithm was consistently the highest-performing model across all the reviewed studies.

Educational Contexts

The educational context is predominantly higher education, focusing on predicting outcomes in university courses, online learning platforms, and attendance-based performance. There is minimal representation of K-12 contexts or non-Asian/European regions. The literature's emphasis on higher education and the underrepresentation of K-12 and non-Asian/European regions stem from systemic research priorities and data accessibility, not inherent methodological limitations. Key reasons supported by literature:

- **Data availability and administrative infrastructure:** Universities systematically collect granular digital trace data (LMS logs, attendance, grades) at scale, which is essential for machine learning models. K-12 settings often lack such unified digital infrastructure, particularly in low-resource regions (Long & Siemens, 2011; Gašević et al., 2015).
- **Institutional incentives and research funding:** Higher education institutions prioritize learning analytics for accreditation, retention, and quality assurance, driving funding and publication. K-12 research faces stricter ethical constraints (parental consent) and fragmented local governance, reducing large-scale studies (Slade & Prinsloo, 2013).
- **Geographic publication bias:** Asian (e.g., China, India) and European institutions dominate because of strong national education technology initiatives (e.g., China's "Smart Education" program) and EU-funded digital

transformation projects. Conversely, African, Latin American, and Southeast Asian contexts are underrepresented due to limited research funding, weak data systems, and lower journal submission rates—a known global disparity in educational data mining (Fischer et al., Educational, 2020).

- Ethical and logistical barriers in K-12: Minors require multi-layered consent, and schools are less likely to share sensitive behavioral data. Higher education students are adults, enabling broader data usage with minimal consent friction (Zawacki-Richter & Latchem, 2018).

These structural factors explain why the literature's scope reflects research feasibility rather than the actual importance of K-12 or non-Asian/European contexts.

Limitations

Several methodological limitations should be considered when interpreting the reported performance. First, validation procedures were not consistent across the studies. Some studies applied cross-validation, while others relied on a single training and testing split. Albreiki et al. (2022), for example, used five-fold cross-validation. Alsumaidaie et al. (2024), however, evaluated models using separate training and testing samples. These differences make direct comparison difficult. Similar accuracy values may not provide the same evidence of model generalisability.

Second, exceptionally high performance requires cautious interpretation. Some studies reported values approaching or reaching 100%, particularly Gul et al. (2025) and Gupta et al. (2024). Such results do not, on their own, demonstrate data leakage or overfitting. However, they highlight the need to examine whether predictors were available at the intended prediction point. The relationship between the predictors and the target should also be considered. In addition, model development and testing data should be adequately separated. Data leakage can produce overly optimistic estimates of predictive performance (Kapoor & Narayanan, 2023; Rosenblatt et al., 2024). Variables such as accumulated academic performance, attendance, assessment results and course-completion information can provide strong predictive signals. However, their value for early prediction may be limited if they become available late in the learning period.

Third, the studies did not use consistent evaluation measures. Classification studies reported different combinations of accuracy, precision, recall and F1-score. Regression studies used measures such as MAE, MSE and RMSE. Some also reported percentage-based measures that are not directly comparable with classification accuracy. Models should therefore not be ranked solely by their reported numerical performance across studies.

The geographical concentration of the evidence also limits generalisability. Most of the reviewed studies were conducted in Asian educational settings. Other regions had comparatively little representation. Educational systems, assessment practices and student populations differ across settings. Institutional resources and access to digital learning technologies also vary. A model that performs well in one setting may therefore not achieve similar results elsewhere. External validation across institutions and geographical contexts remains important.

The review was also limited to studies retrieved through SpringerLink and IEEE Xplore. Other academic databases were explored. However, the search string used in this review did not produce relevant studies for inclusion from these sources. Differences in database indexing, syntax and search functionality may have contributed to this outcome. Restricting the final search to two databases may therefore have resulted in relevant studies being missed. This is particularly relevant to studies indexed in multidisciplinary and education-focused databases such as Scopus, Web of Science and ERIC. Future reviews could develop database-specific search strings and extend the search across a wider range of bibliographic databases.

Publication bias is another possible limitation. Studies reporting positive or statistically significant findings may be more likely to be published or fully reported than studies with non-significant findings (Dwan et al., 2013). This may contribute to the strong predictive performance observed in the published literature. However, publication bias was not formally assessed in this review. The reported performance should therefore be interpreted with this possibility in mind.

Finally, explainability was not consistently incorporated into the reviewed models. High predictive performance alone provides limited guidance to educators when the reasons behind an individual student's predicted risk are unclear. Explainable approaches can help identify the factors contributing to a prediction. These may include attendance, assessment performance or engagement. Such information can make predictions more useful for targeted student support. Future studies should therefore consider predictive performance alongside validation, explainability and practical usefulness.

CONCLUSION

In conclusion, this systematic literature review highlights the growing integration of machine learning (ML) within higher education as a means of predicting student performance and engagement. With the analysis of 17 research articles published between 2021 and 2025, it is evident that various algorithms have been employed. Random Forest emerged as one of the predominant and higher-performing models, although it did not consistently outperform alternative algorithms. SVM, Extra Trees and MLP also achieved strong performance in several studies. Reported performance ranged from approximately 70% to values approaching 100%. These differences show that performance cannot be attributed to algorithm choice alone. Dataset characteristics, predictor variables, target definition and evaluation approaches also influence reported performance.

Despite the impressive predictive capabilities reported, the review reveals considerable methodological heterogeneity and limited integration of Explainable Artificial Intelligence (XAI). The lack of explainability presents challenges for practical implementation. Educators may be able to identify students at risk but have limited information about the factors contributing to those predictions. This can make it difficult to determine whether support should focus on attendance, academic performance, engagement or other contributing factors.

Furthermore, the geographical concentration of research within Asia highlights the need for studies exploring more diverse contexts, particularly in underrepresented regions. Differences in educational systems, student populations, assessment practices and institutional resources may affect how well predictive models perform across settings. The findings therefore call for more consistent evaluation practices and clearer reporting of dataset characteristics, predictor variables and prediction timing. Future research should also strengthen validation procedures and integrate XAI to improve interpretability. Greater external validation across institutions and geographical settings is also needed. By addressing these gaps, future research can support the development of more transparent, reliable and practically useful predictive systems that enhance educational experiences and student outcomes across varied educational landscapes.

REFERENCES

1. Albahli, S. (2025). Advancing sustainable educational practices through AI-driven prediction of academic outcomes. *Sustainability*, 17(3), 1087.
2. Albreiki, B., Habuza, T., & Zaki, N. (2022). Framework for automatically suggesting remedial actions to help students at risk based on explainable ML and rule-based models. *International Journal of Educational Technology in Higher Education*, 19, 49. <https://doi.org/10.1186/s41239-022-00354-6>
3. Alhazmi, E., & Sheneamar, A. (2023). Early Predicting of Students' Performance in Higher Education. *IEEE Access*, (11) 27579-27589.
4. Ali, F. Z., Ayazuddin, R., & Sanjaya, I. (2025). Predicting Course Engagement with Machine Learning Techniques. *Engineering Proceedings*, 107(1), 46.
5. Alturki, S., Cohausz, L., & Stuckenschmidt, H. (2022). Predicting Master's students' academic performance: an empirical study in Germany. *Smart Learning Environments*, 9(38), <https://doi.org/10.1186/s40561-022-00220-y>.
6. Ayouni, S., Hajjej, F., Maddeh, M., & Al-Otaibi, S. (2021). A new ML-based approach to enhance student engagement in online environment. *PLOS ONE*. 16(11), e0258788. <https://doi.org/10.1371/journal.pone.0258788>

7. Borges, G. A., Pedro, C. F. D. S., Dos Anjos, J. C. S., Rodrigues, A., Boavida, F., & Silva, J. S. (2025). A platform for early class dropout prediction of university students. *IEEE Access*, 13, 109116-109133. <https://doi.org/10.1109/ACCESS.2025.3581751>.
8. Brahim, G.B. Predicting Student Performance from Online Engagement Activities Using Novel Statistical Features. *Arab J Sci Eng* 47, 10225–10243 (2022). <https://doi.org/10.1007/s13369-021-06548-w>
9. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
10. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.
11. Daimi, S. A., & Iqbal, A. (2023). Ensemble Machine Learning to Predict and Feature Engineering to Identify Factors for Academic Success. *Proceedings of the Fourth International Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023)*. Atlantis Press. https://doi.org/10.2991/978-94-6463-314-6_28
12. Dwan, K., Gamble, C., Williamson, P. R., & Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias—An updated review. *PLoS ONE*, 8(7), e66844.
13. Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., ... & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of research in education*, 44(1), 130-160.
14. Fu, J., Zhong, Z., Qin, J., & Luo, L. (2025). AHP-Guided Stacked Ensemble Modeling for Student Engagement Level Prediction in Online Education. *IEEE Access*, 13(doi: 10.1109/ACCESS.2025.3575529), 114230-114240.
15. Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64-71.
16. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3–42.
17. Gul, M. N., Abbasi, W., & Wani, M. Y. (2025). Revolutionizing educational decision-making: a robust machine learning mechanism for predicting student performance. *Journal of Electrical Systems and Information Technology*, 12(1), 32. <https://doi.org/10.1186/s43067-025-00230-z>
18. Gupta, V., Singhal, P., & Khattri, V. (2024). Enhancing Predictive Accuracy in Education: A Detailed Analysis of Student Performance Using Machine Learning Models. *Tuijin Jishu/Journal of Propulsion Technology*, 45(3), 2024.
19. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3–42.
20. Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 359–366
21. Alsumaideie, M. I., Nafea, A. A., Mukhlif, A. A., Jalal, D. R., & AL-ani, M. M. (2024). Intelligent System for Student Performance Prediction Using Machine Learning. *Baghdad Science Journal*, 3877-3891.
22. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804.
23. Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review* (Online).
24. Marín, Y. R., Huatangari, L. Q., Tuesta, J. N. A., Caro, O. C., Guevara, J. L. M., Bardales, E. S., & Santos, R. C. (2025). Analysis of factors affecting the academic performance of university students using machine learning. *Scientific Reports*, 15(1), 44027.
25. Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, 13
26. Onwuzurike, M. A., & Kpogli, S. A. (2025). Predictive modeling of student engagement and behavioral outcomes using machine learning techniques in technology-enhanced classrooms. *International Journal of Scientific Research in Humanities and Social Sciences*, 2(6), 58-79.
27. Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

28. Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 15, 1829.
29. Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American behavioral scientist*, 57(10), 1510-1529.
30. Wang, W. (2025). Machine learning-based engagement prediction for online courses. *ITM Web of Conferences*, 70, 04014. <https://doi.org/10.1051/itmconf/20257004014>.
31. Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(11), <https://doi.org/10.1186/s40561-022-00192-z>.
32. Zawacki-Richter, O., & Latchem, C. (2018). Exploring four decades of research in Computers & Education. *Computers & Education*, 122, 136-152.
33. Zhang, L., Gao, J., Zhao, L., Liu, Z., & Guan, A. (2025). Predicting College Student Engagement in Physical Education. *Applied sciences*, 15, 3884.
34. Zhao, L., Chen, K., Song, J., Zhu, X., Sun, J., Caulfield, B., & Mac Namee, B. (2021). Academic performance prediction based on multisource, multifeatured behavioral data. *IEEE Access*, 9, 5453-5465.
35. Zou, W., Zhong, W., Du, J., & Yuan, L. (2025). Prediction of student academic performance utilizing a multi-model fusion approach in the realm of machine learning. *Applied Sciences*, 15(7), 3550.