

A SYSTEMATIC APPROACH TO PREDICTING STUDENTS' ACADEMIC PERFORMANCE: A REVIEW OF RECENT LITERATURE

Anselmus Yata Mones✉

Sekolah Tinggi Pastoral St. Petrus
Keuskupan Atambua, Indonesia

✉ anselmojata@gmail.com

ABSTRACT

The rapid expansion of digital learning has generated large volumes of educational data, creating new opportunities to apply machine learning (ML) and data mining techniques to predict student academic performance. This study synthesizes 58 empirical studies that used Decision Trees, Random Forests, Support Vector Machines, Logistic Regression, and Artificial Neural Networks to identify at-risk students and improve educational outcomes.

The review focuses on predictor variables, validation methods, accuracy rates, and performance metrics. Findings suggest that the most effective predictive models combine four categories of variables: demographic factors, academic indicators, digital behavioral features, and psychosocial attributes. Among the algorithms examined, Random Forest and Artificial Neural Networks demonstrated the strongest predictive performance, achieving accuracy rates of 85%–93% across k-fold cross-validation and train-test split validation.

Performance measures such as precision, recall, F1 score, and AUC further confirm the robustness and generalizability of these models. ML-based academic prediction systems can strengthen early warning systems, support data-driven policymaking, and enable personalized learning interventions. The study concludes that combining multidimensional predictors with explainable AI can improve equity, personalization, operational efficiency, and accountability in educational decision-making.

KEYWORDS

Academic performance prediction, artificial neural networks, data mining, educational data analytics, machine Learning, operational efficiency, predictor variables, random forest, validation methods

HOW TO CITE

Yata Mones A. (2026) 'A Systematic Approach to Predicting Students' Academic Performance: A Review of Recent Literature', *Journal on Efficiency and Responsibility in Education and Science*, vol. 19, no. 1, pp. 1–14. <http://dx.doi.org/10.7160/eriesj.2026.190101>

Article history

Received

June 6, 2025

Received in revised form

November 24, 2025

Accepted

February 6, 2026

Available on-line

March 31, 2026

Highlights

- Reviews 58 empirical studies using machine learning to predict student academic performance and identify at-risk learners.
- Finds that the strongest models combine demographic, academic, digital behavioral, and psychosocial predictors.
- Shows Random Forest and Artificial Neural Networks achieved the highest predictive accuracy, typically between 85% and 93%.
- Highlights k-fold cross-validation and train-test split as the most common validation methods in the reviewed studies.

INTRODUCTION

Education plays a fundamental role in human and social development. In the modern education system, the quality of learning and students' academic achievements are two primary indicators of the success of educational institutions. One major challenge educators and educational institutions face is identifying students at risk of academic failure early (Katarya, 2023; Nazir et al., 2023). Some of the challenges identified are difficulties in effectively managing and utilizing data (Baneres et al., 2019), the manual data collection process

is time-consuming and prone to inaccuracies (Pek et al., 2022), and the use of personal data in predictive models raises privacy and ethical concerns, requiring strict data protection measures (Schmidt et al., 2025). This challenge encourages the development of data-driven predictive approaches to understand, anticipate, and improve students' academic performance (Wu et al., 2024).

With advances in information technology and data analysis techniques, new methods have emerged for understanding student behavior and performance, including predicting

academic performance using machine learning algorithms and data mining techniques (Daza et al., 2022; De-La-Cruz et al., 2022; Roslan and Chen, 2022). This is where research on predicting students' academic performance gains significant urgency and relevance, as it can aid policymakers, teachers, and the students themselves in more targeted academic planning (Nazir et al., 2023).

However, the numerous prediction models and methodological differences in previous research create challenges in consistently unifying the findings. Therefore, a systematic literature review (SLR) is an essential approach for summarizing and synthesizing scientific evidence in a methodological, objective, and transparent manner. This approach ensures that the review is conducted thoroughly on relevant scientific publications, providing results that can serve as a foundation for policy development or further research.

Predicting academic performance is not merely a statistical or computational effort but rather a proactive strategy to identify risk factors and opportunities in education. Research shows that factors such as academic records, learning behavior, attendance, socio-economic conditions, and data from Learning Management Systems (LMS) are strongly correlated with students' academic outcomes (Roslan and Chen, 2022).

The application of algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVMs), and Neural Networks has proven effective for processing data to generate accurate predictions. In fact, recent research shows that ensemble learning methods such as boosting and bagging tend to achieve higher prediction accuracy than single algorithms (Wu et al., 2024).

More than just technology, these predictions impact real educational practices. With the presence of an accurate prediction system, educators can design more personalized learning strategies, improve student retention, and reduce dropout rates (Katarya, 2023)

With the increasing volume of scientific publications, narrative literature reviews are no longer sufficient to objectively summarize research findings. The SLR method is a systematic, standardized, and replicable approach that follows guidelines such as PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). SLR allows researchers to filter the literature using strict inclusion and exclusion criteria, evaluate the methodological quality of each study, and synthesize findings thematically and quantitatively.

Several studies have conducted SLR in the context of academic prediction. For example, Daza et al. (2022) in their review found that the Knowledge Discovery in Databases (KDD) method is the most commonly used, with the CGPA attribute as the main prediction indicator, and Naïve Bayes, Neural Network, and Decision Tree as the most frequently applied algorithms (Daza et al., 2022; Karim-Abdallah et al., 2025). Meanwhile, Nawang et al. (2021) noted that academic data and student behavior are the most frequently appearing variables as determinants of academic performance.

Although the literature on predicting academic performance has developed rapidly, there are still several important gaps that need to be addressed:

- Many studies do not detail the procedures for attribute selection and model validation, thereby affecting the replicability and generalizability of results.
- The dominance of higher education contexts, while secondary and primary school contexts are relatively underexplored in depth (Rodrigues et al., 2022).
- There is a lack of studies evaluating the impact of implementing predictive models on real academic outcomes, such as student retention or grade improvement after intervention.
- The lack of commonly used benchmark datasets in research limits the ability to compare model performance across different studies (Balaji et al., 2021).

The identification of this gap emphasizes the need for a systematic literature review that not only presents a list of models and techniques but also assesses the methodological quality, contextual diversity, and practical potential of the research.

Over the past decade, developments in machine learning technology and big data analysis have broadened the range of methodologies for predicting academic success. Initially, the study focused solely on exam results or cumulative grade point averages (GPAs); however, prediction methodologies are currently evolving to become more intricate and multimodal. Recent research indicates that integrating academic, behavioral, demographic, and psychosocial data can substantially enhance the precision of prediction models (Baashar et al., 2021; Valdiviezo-Diaz and Chicaiza, 2024).

Commonly employed strategies in predictive research encompass:

- *Decision Trees and Random Forests* are known for their superior interpretability and effective performance on educational data. Decision trees are known for their transparency and ease of understanding. They recursively partition the data into subspaces, making the decision-making process clear and interpretable (Rokach, 2016; Yassine and Mohammed, 2024). Random Forest is an ensemble method that builds multiple decision trees and combines their predictions to improve accuracy and robustness (He et al., 2018; Shaik and Srinivasan, 2019). Studies consistently show that Random Forests achieve higher accuracy than Decision Trees. For instance, Random Forest achieved an accuracy of 91.7% compared to 88.2% for Decision Trees in predicting student success (Reddy et al., 2024). Another study found that Random Forests had higher R^2 scores and lower error metrics than Decision Trees (He et al., 2018).
- *Support Vector Machine (SVM)* is appropriate for binary classification tasks, such as pass/fail or promote/not promote. They work by finding the optimal hyperplane that separates data points of different classes with the maximum margin. SVMs are particularly useful in applications requiring high accuracy and robustness, such as structural health monitoring and fault diagnosis (Pisner and Schnyer, 2020; Wandekoken et al., 2011). They can handle large datasets and complex classification problems, although they may require significant computational resources for training (Gholami and Fakhari, 2017).

- *Artificial Neural Networks (ANNs)* exhibit exceptional efficacy in managing large, non-linear datasets. This is because these networks can model complex relationships and patterns in the data. They are widely used in various fields, including fog computing, remote sensing, and healthcare, to enhance data processing, classification, and decision-making (Abdolrasol et al., 2021; Abiodun et al., 2019). ANNs can adapt to noisy data and achieve high classification accuracy, making them ideal for applications that require robust, flexible models (Abdolrasol et al., 2021; Aghbashlo et al., 2015).
- *Naïve Bayes and K-Nearest Neighbors (KNN)* are often used as baseline models for comparison due to their simplicity and ease of implementation. Naïve Bayes is based on the Bayes theorem and assumes independence among features, making it computationally efficient but sometimes less accurate than more complex models (Kumar et al., 2024; Wickramasinghe and Kalutarage, 2021). KNN classifies data points based on the majority class of their nearest neighbors, which can be effective for small datasets but may struggle with large or high-dimensional data (Ortiz-Bejar et al., 2020; Ripan et al., 2021).

The influence of using student behavioral data from e-learning platforms is growing more substantial. Data such as the number of clicks, time spent completing assignments, daily login duration, and frequency of contact with learning materials have all proved to be critical markers of students' ultimate academic achievement (Daza et al., 2022; Namoun and Alshanqiti, 2020).

The ability to accurately predict student performance has strategic value for all stakeholders in education. For educational institutions, predictive models can be used to: 1) Identify low-risk students (Khan and Menezes, 2020; Seo et al., 2024) performance early and providing appropriate interventions; 2) Directing educational resources (teachers, guidance, facilities) more efficiently (Hoyos Osorio and Daza Santacoloma, 2023); 3) Improving student retention and reducing dropout rates (Deleña et al., 2025; Salloum et al., 2024); 4) Monitoring the impact of data-driven academic policies (Matz et al., 2023). For students, prediction results can serve as a tool for self-reflection, allowing them to independently strategize their learning, manage their time better, and receive personalized guidance tailored to their needs. Research by Chakrapani and Chitradevi (2022) concluded that the use of data mining-based predictive models enables faster and more targeted academic interventions and can promote academic success despite students' diverse backgrounds (Chakrapani and Chitradevi, 2022).

The urgency of this study is growing, given that most schools and universities now gather substantial amounts of data through academic information systems and online learning platforms. Without a systematic framework for analyzing and utilizing this data, the potential for evidence-based early intervention will be lost.

In the global context, education is increasingly driven by the principles of evidence-based policy, in which important decisions on curriculum, teaching methods, resource allocation, and remedial programs are grounded in data and scientific findings (Salmi, 2015). One of the main drivers of

this approach is Sustainable Development Goal (SDG) 4, which states, “ensure inclusive and equitable quality education and promote lifelong learning opportunities for all” (García-Martín and Pérez Fernández, 2025; Reimers, 2024).

Predicting academic performance is an integral part of this effort, as it enables quick and accurate decision-making to prevent learning failures and dropouts. Educational institutions that implement predictive systems based on machine learning, as extensively researched in various scientific publications, can enhance the efficiency of academic and remedial guidance programs, adjust individual learning strategies based on performance predictions, and compile comprehensive student risk profiles (Bilquise et al., 2024; Deleña et al., 2025; Vaarma and Li, 2024). The use of predictive algorithms extends beyond performance classification to include the design of early interventions tailored to students' behavioral and academic profiles (Wu et al., 2024).

Moreover, in the post-pandemic era, many schools and universities have switched to hybrid learning systems, which generate large amounts of digital data—such as interactions with learning materials, forum participation, task completion speed, and online resource usage. All this data, if processed effectively, can become a strategic asset for mapping students' potential and challenges in real time. Considering the complexity of the theme, practical urgency, and the abundance of available literature, this research aims to:

- Conduct a systematic review of recent scientific literature (2015–2024) discussing the prediction of student academic performance.
- Identifying the main methods, algorithms, and attributes used in academic prediction research,
- Evaluating the methodological quality and publication trends in this field,
- Uncovering research gaps and recommending directions for further studies to develop more effective and inclusive prediction models.

The main contribution of this article lies in its effort to create a comprehensive knowledge map of student academic prediction approaches grounded in SLR principles. This map is expected to:

- Serve as a reference for novice researchers in selecting algorithms and predictive variables,
- Serving as a basis for decision-making for education policymakers,
- Providing practical insights for teachers and educational counsellors in implementing early detection systems in schools.

This research also addresses the calls of many previous studies that lament the lack of systematic synthesis in the literature on student performance prediction (Nabil et al., 2022; Waheed et al., 2020), as well as the limitations in developing benchmark predictive models that can be tested across educational contexts (Pelima et al., 2024).

The Systematic Literature Review (SLR) approach was chosen for this study as a robust framework to assess and synthesize scientific contributions to predicting students' academic

performance. The author hopes to make scientific contributions that are not only academically relevant but also practically impactful in developing adaptive, predictive, and inclusive educational systems.

METHOD

This research uses the SLR approach as the primary method to systematically identify, evaluate, and synthesize scientific publications on predicting student academic

performance (Fundoni et al., 2023; Krüger et al., 2020; Prill et al., 2021). SLR allows this research to be conducted transparently, structurally, and replicably so that the results can be academically accountable and serve as a reference for further research.

Research design

The design of this research aligns with the PRISMA framework, which consists of four primary stages:

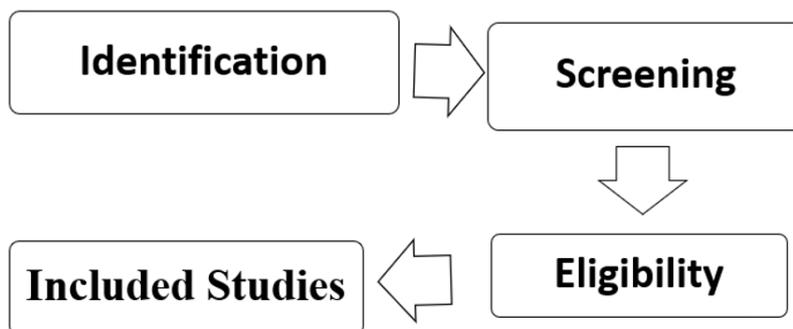


Figure 1: Stages of the PRISMA framework

The process illustrated in the image above follows a series of strict steps, beginning with the formulation of the research question and culminating in the thematic analysis of the study’s findings.

Literature search strategy

The search was conducted across several reputable scientific databases, including Scopus, Publish or Perish, Google Scholar, ScienceDirect, and ACM Digital Library. The keywords employed include academic performance prediction, student performance, machine learning, educational data mining, and systematic literature review. Several criteria guide researchers in selecting articles for this study, including those published between 2015 and 2024, focusing on student academic performance through computational methods and data-driven empirical models, and written in English. Conversely, articles that do not meet the criteria include those with unclear writing style, those that are not peer-reviewed, and studies that focus solely on dropout rates without addressing academic performance. Following data extraction, the researchers conducted a thematic analysis based on the following categories: the prediction algorithms used; data types (academic, behavioral, demographic); the main objectives of the study; model validation (*cross-validation*, *hold-out*, *k-fold*); and the implications of the research findings. Additionally, assessing methodological quality is crucial as an indicator that warrants evaluation. These indicators include the clarity of research objectives and problems, data validity and source quality, clarity of analysis methods, comprehensive reporting of results, and relevance to the objectives of this research.

Study selection process

The literature identification and selection process in this study followed the PRISMA guidelines to ensure transparency and replicability of the results.

At the *identification stage*, the researcher collected data from various sources, namely databases ($n = 2,000$) and registers ($n = 700$), yielding a total of 2,700 records. Before the screening process, some records were removed due to duplication ($n = 230$), unsuitability as determined by automated tools ($n = 500$), and other reasons, such as incorrect format or incomplete data ($n = 50$). Thus, 1,920 records were advanced to the screening stage.

The *screening stage* involves reviewing titles and abstracts to assess their relevance to the research topic. A total of 1360 recordings were excluded as irrelevant, leaving 560 reports for further review.

Next, in the *eligibility stage*, the researcher attempts to obtain the full text of the remaining reports. However, 473 reports were inaccessible, leaving only 87 that could be successfully reviewed in their entirety. Of that number, 29 reports were excluded, with the following breakdown: 14 reports were not relevant to the research focus, 10 were not in English, and 5 had unclear research methods. Fourteen articles were excluded because they were not relevant to the research focus and did not directly relate to the research variables or parameters. These articles generally discussed topics outside the scope of the study and therefore could not provide the necessary data. This process was carried out to maintain data consistency and ensure that only literature that directly contributed to the primary analysis was included.

The *final stage*, inclusion, yielded 58 studies that met all selection criteria and were deemed suitable for this systematic review.

Overall, this selection process reflects a systematic, transparent, and replicable approach, aiming to ensure that the literature used is truly relevant and possesses adequate methodological quality to support the analysis of predicting students’ academic performance.

The PRISMA diagram is displayed as follows:

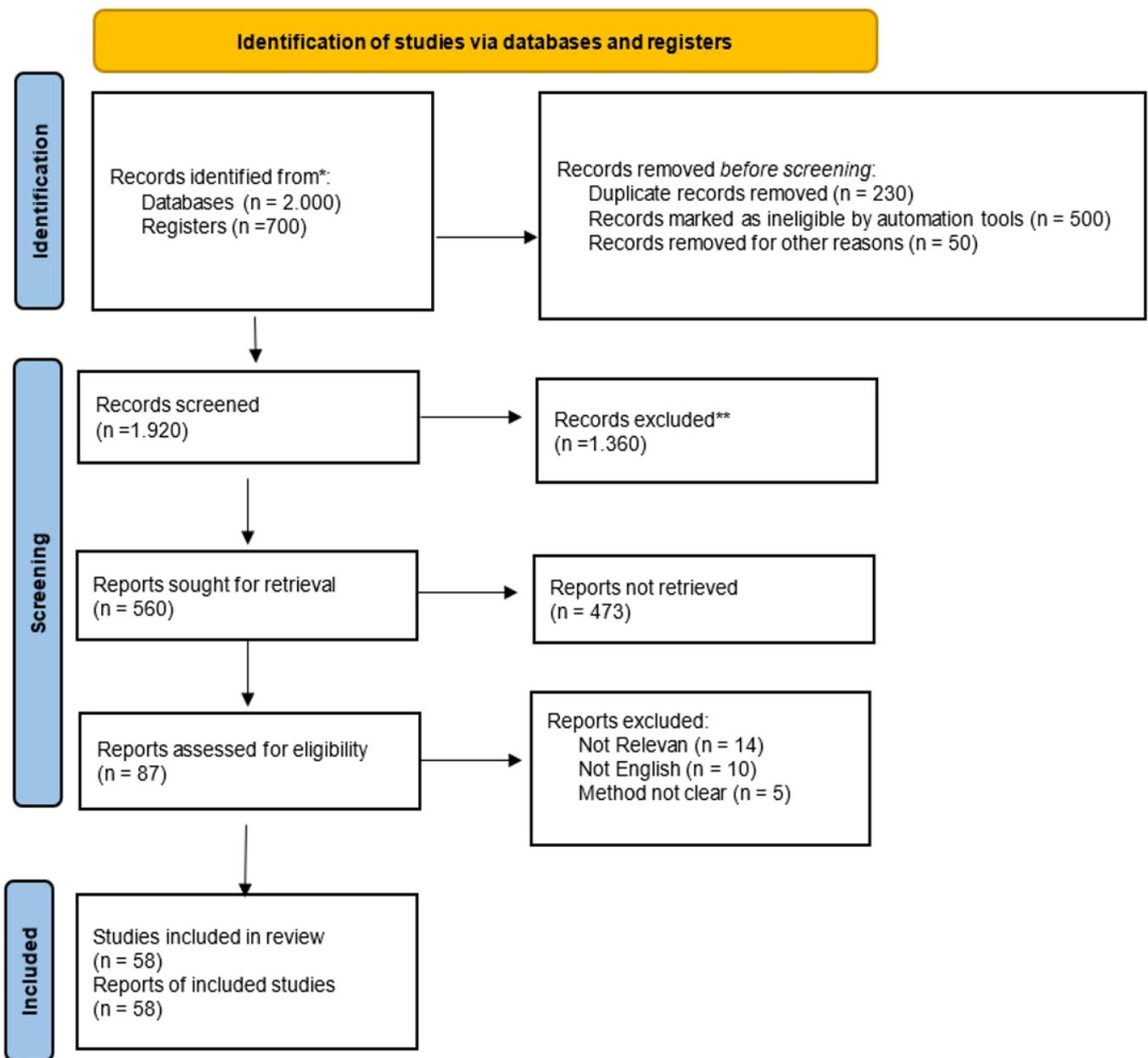


Figure 2: PRISMA flowchart

RESULTS

General trends in research related to predicting student academic performance

Research on predicting student academic performance has advanced rapidly over the past decade, primarily through ML and *Educational Data Mining (EDM)*. An analysis of 58 selected scientific articles reveals a growing trend in publications on predicting student academic performance, particularly since 2020. The most frequently utilized techniques include *Decision Trees*, *Random Forests*, *Artificial Neural Networks*, and *Support Vector Machines* (De-La-Cruz et al., 2022; Molina and Cancell, 2021; Nazir et al., 2023; L. Zhang et al., 2021). These methods are favored for their ability to manage complex data and deliver accurate predictions. Additionally, ensemble methods

such as Gradient Boosting and Naïve Bayes are often employed to enhance prediction accuracy (Chakrapani and Chitradevi, 2022; Kamal and Ahuja, 2019). This research indicates that machine learning models can effectively predict academic performance by leveraging various factors such as academic records, demographics, and student behavior (Valdiviezo-Diaz and Chicaiza, 2024). This progress is driven by the adoption of machine learning technology and the growing availability of digital educational data. Studies conducted by Hellas et al. (2018) and Roslan and Chen (2022) highlight that research in this area is not only increasing in quantity but also diversifying in the methods and variables utilized (Hellas et al., 2018; Roslan and Chen, 2022).

A comparison of the strengths and weaknesses of each Machine Learning Model can be shown in the following table.

Algorithm	Strengths	Use Cases
Decision Tree	High interpretability, easy to understand, and visualize	Identifying key factors in student performance (Niranjala et al., 2024)
Random Forest	High accuracy, robust to overfitting, handles large datasets, and missing data.	High accuracy, robust to overfitting, handles large datasets and missing data Predicting student performance, identifying at-risk students (Abdelaziz et al., 2025; Salman et al., 2024)
SVM	High accuracy in binary classification, effective for well-separated classes	Pass/fail prediction, grade movement prediction (Niranjala et al., 2024; Strub, 2020)
ANN	Handles non-linear and extensive datasets, adaptable to noisy data	Requires large datasets, can be computationally intensive (Lutsenko and Zgonnikov, 2024)
Naïve Bayes	Simple, fast, and efficient for small datasets	Assumes feature independence, may be less accurate for complex relationships (Antonakis and Sfakianakis, 2009)
K-Nearest Neighbors (KNN)	Simple, intuitive, effective for small datasets	Computationally expensive for large datasets, sensitive to feature scaling (Ripan et al., 2021)

Table 1: A comparison of the strengths and weaknesses of each Machine Learning Model

Frequently used algorithms

Some commonly used algorithms include Decision Trees, Random Forests, SVMs, Logistic Regression, and ANNs. Decision Trees work by dividing data into branches based on specific variables to produce easily interpretable decisions. Random forests are an extension of Decision Trees that use multiple decision trees to improve prediction accuracy (Akhatkulov et al., 2024; Alhassan et al., 2020; Nazir et al., 2023). Meanwhile, SVM is used to determine the optimal separating hyperplane between data classes, which is suitable for classifying students as “at risk of failing” or “successful.” Logistic regression is often used to predict the probability of an event, such as the likelihood that a student will not graduate.

At the same time, ANN mimics the way the human brain works to find complex patterns in educational data.

In practice, the analysis results from these algorithms can be used to provide personalized recommendations for each student. For example, the system can provide early warnings to teachers or parents if a student shows a pattern of declining performance based on their historical data. Additionally, schools can develop adaptive learning strategies that tailor the material and teaching methods to each student’s unique needs. Thus, the application of machine learning and data mining in education not only improves decision-making efficiency but also helps create a more inclusive and data-driven learning environment. Here is a quantitative table from various studies related to the use of machine learning algorithms to predict students’ academic performance:

Algorithm	Usage in Research (%)	Main Advantages	Main Disadvantages	References
Decision Tree	20%	Easy to interpret, suitable for categorical data	Prone to overfitting	(Al-Khlifeh et al., 2025)
Random Forest	25%	High accuracy, reduces overfitting by combining multiple trees	Hard to interpret	(Albreiki et al., 2021)
Support Vector Machine (SVM)	18%	Effective for high-dimensional data and clear class boundaries	Less efficient for large datasets	(Xu et al., 2019)
Logistic Regression	15%	Simple, fast, and easy to interpret	Not suitable for complex non-linear relationships	(Y. Zhang et al., 2021)
Artificial Neural Network (ANN)	22%	Can learn complex and non-linear patterns	Requires large amounts of data and a long training time	(Adejo and Connolly, 2018)

Table 2: Frequently used algorithms

Key algorithms and their applications

ML and data mining techniques have become powerful tools for predicting academic performance and identifying students at risk. This section highlights the key algorithms commonly used in educational data mining: Decision Trees, Random Forest, SVM, Logistic Regression, and ANN, and discusses their main applications supported by empirical studies.

The analysis of these algorithms indicates that ensemble and deep learning models, including Random Forest and ANN, often surpass classical models in prediction accuracy (85–93%). However, Decision Tree and Logistic Regression retain significance due to their interpretability and simplicity. In educational settings, the trade-off between

accuracy and explainability is crucial, as transparent models are frequently favored for policy and intervention decisions.

Predictor variables in academic performance prediction

Understanding the characteristics that most strongly predict academic achievement is vital for building accurate, interpretable ML models in education. Across 58 empirical studies, four primary groups of predictor variables appear consistently: demographic, academic, digital behavioral, and psychological characteristics. Each category significantly adds to the model’s explanatory power and provides insights into educational policy and intervention design.

Algorithm	Description	Main Applications	Representative Studies
Decision Tree (CART, ID3, C4.5)	A rule-based model that recursively splits data into hierarchical nodes to make decisions.	Widely applied in early identification of at-risk students, dropout detection, and performance classification due to its interpretability.	<i>Predicting Academic Performance Using Machine Learning Techniques</i> (Akhatkulov et al., 2024; Bhimavarapu et al., 2025; Dai and Lu, 2024; Kumar et al., 2022)
Random Forest	An ensemble learning method that combines multiple decision trees to improve accuracy and robustness.	Used in student performance prediction, satisfaction analysis, and anomaly detection in educational datasets.	<i>Application of Machine Learning in Predicting Student Performance</i> (Boujmiraz et al., 2026; Cruz and Lumauag, 2024)
Support Vector Machine (SVM)	A supervised learning algorithm that separates classes using an optimal hyperplane with maximum margin.	Applied for classifying student outcomes, analyzing attendance patterns, and predicting graduation success.	<i>A Review on Predicting Student Academic Performance Using Data Mining Techniques</i> (Akhatkulov et al., 2024; Masangu et al., 2021; Pelima et al., 2024)
Logistic Regression	A statistical model that estimates the probability of a binary outcome (e.g., pass/fail, dropout/non-dropout).	Commonly used as a baseline model for academic success prediction based on demographic and socioeconomic variables.	<i>Using Data Mining Techniques to Predict Student Academic Performance</i> (Akhatkulov et al., 2024; Alhassan et al., 2020; Cruz and Lumauag, 2024)
Artificial Neural Network (ANN)	A computational model inspired by the human brain, consisting of interconnected layers of artificial neurons.	Highly effective in modeling complex nonlinear relationships in educational data, including e-learning behavior analysis and personalized learning recommendations.	<i>Analysis of Machine Learning Algorithms for Predicting Student Performance</i> (Chakrapani and Chitradevi, 2022; Daza et al., 2022; Rajendran et al., 2022; Santiketa et al., 2024)

Table 3: Key algorithms and their applications

Demographic predictors

Demographic variables form the foundation of most predictive models, as they describe the social and economic circumstances in which learning occurs. Variables such as gender, age, parental education, household income, and socioeconomic status (SES) are among the most often utilized. Multiple studies using Decision Tree, Random Forest, and Logistic Regression algorithms suggest that demographic characteristics, notably parental education and wealth, are strongly associated with student success (De-La-Cruz et al., 2022; Sarker et al., 2024; L. Zhang et al., 2021).

Students from wealthier socioeconomic backgrounds tend to achieve better academic performance due to access to superior educational resources, technological support, and favorable learning environments (Albreiki et al., 2021; Molina and Cancell, 2021; Nabil et al., 2022). For instance, decision tree models commonly highlight parental education as a significant node in the categorization of high- vs. low-performing kids (Balaji et al., 2021; Nawang et al., 2021). Similarly, Random Forest models boost prediction accuracy by mixing demographic factors with academic indicators such as attendance and past grades (He et al., 2018; Salman et al., 2024).

The implications are clear: incorporating demographic factors into predictive models enables institutions to identify structural inequities and design targeted interventions for disadvantaged groups, thereby promoting fairness and inclusivity in educational analytics.

Academic predictors

Academic factors provide the strongest and most consistent predictors of student achievement across practically all

algorithms. These include past GPA, test results, attendance records, assignment submissions, and involvement in learning activities. Studies utilizing Random Forest and ANN models have repeatedly identified these traits as the most significant contributors to model accuracy (Abiodun et al., 2018; Shaik and Srinivasan, 2019; Yassine and Mohammed, 2024).

SVM and Logistic Regression algorithms also perform effectively when applied to academic indicators, especially for binary or ordinal classifications (e.g., pass/fail, dropout/non-dropout) (García-Martín and Pérez Fernández, 2025; Gholami and Fakhari, 2017). Academic predictors directly reflect prior achievement and learning consistency, making them the strongest statistical indicators of future success (Santiketa et al., 2024; Sarker et al., 2024).

Digital behavioral predictors

With the advent of e-learning environments and LMS, digital behavioral data have become more crucial for evaluating student engagement (Y. Zhang et al., 2021). Variables in this area include login frequency, time spent on learning platforms, number of forum conversations, resource downloads, video viewing duration, and quiz attempts (Farissi et al., 2023).

Machine learning models such as ANNs and SVMs excel in this domain because they can capture nonlinear and high-dimensional relationships in behavioral data (Baneres et al., 2019; Nguyen and Jones, 2022) especially in countries like Vietnam, with rich biodiversity and a high population growth rate. One of the main causes of biodiversity loss

in Vietnam is the unsustainable bushmeat consumption rate in urban areas. To help mitigate the demand for bushmeat, this study aims to examine the associations between biodiversity loss perceptions, perception toward the prohibition of illegal wildlife consumption, and bushmeat consumption behaviors among urban residents in Vietnam. The investigation employed the Bayesian Mindsponge Framework (BMF). For example, an ANN model achieved over 90% accuracy in predicting final grades using patterns of LMS interaction, highlighting the role of online engagement as a proxy for motivation and effort (Al-Khlifeh et al., 2025; Daza et al., 2022).

These predictors provide educators with real-time insights into student engagement, enabling proactive interventions (e.g., early warnings for disengaged students or adaptive learning pathways). Moreover, integrating digital behavioral variables with academic records increases the accuracy and timeliness of predictive models.

Psychosocial predictors

Psychosocial factors are often overlooked but are gaining increased recognition; they encompass the emotional, motivational, and interpersonal aspects of learning (Karim-Abdallah et al., 2025). Variables such as motivation, self-efficacy, emotional intelligence, learning strategies, stress level, and family support enhance the interpretability of prediction models (Bilquise et al., 2024; Daza et al., 2022).

Decision tree and random forest models often identify motivation and study habits as major contributors to performance differentiation, particularly when combined with academic data (Reddy et al., 2024; Salman et al., 2024; Shaik and Srinivasan, 2019). Meanwhile, ANN models can represent more subtle, nonlinear dependencies between psychosocial factors and academic success, such as the indirect effects of stress and peer collaboration on performance (Aghbashlo et al., 2015; Hellas et al., 2018).

Validation and accuracy methods for predicting student academic performance

The dependability and applicability of ML models for forecasting academic achievement are fundamentally contingent on the validation techniques and accuracy measures used. Strong validation ensures that the predicted outcomes aren't solely the result of overfitting the model but instead reflect patterns that can be applied to new or unknown student data. The 58 papers used a wide range of validation methods and performance indicators. The most common ones were k-fold cross-validation, train-test splits, and hold-out procedures.

1. Validation Techniques

a) K-Fold Cross-Validation

The *k*-fold cross-validation method is the most widely adopted technique in educational prediction studies, particularly with Decision Tree, Random Forest, and SVM algorithms (Salman et al., 2024; Strub, 2020; Sushma and Sriramakrishnan, 2025).

- In this method, the dataset is divided into *k* equally sized folds (most commonly 5 or 10).
- The model is trained on *k*-1 folds and tested on the remaining one, repeating the process *k* times.
- The final performance score is averaged across all folds to minimize bias and variance.

This approach ensures stability and prevents the model from being overly dependent on specific data partitions. Studies by Farissi et al. (2023), Lam et al. (2024), Salman et al. (2024), and Sarker et al. (2024) demonstrated that *k*-fold cross-validation enhances the consistency of predictive results, especially when datasets are limited in size.

b) Train-Test Split Method

The train-test split method remains common in larger datasets, particularly for ANN and Logistic Regression models. Typical ratios such as 70:30 or 80:20 are used to split the data into training and test sets (Bhimavarapu et al., 2025; Lou and Colvin, 2025; Pelima et al., 2024). While simpler than cross-validation, this approach provides a rapid estimation of model performance and is suitable when datasets exceed several thousand records (Albreiki et al., 2021; Cheng et al., 2024; De-La-Cruz et al., 2022). However, its disadvantage is the potential variance introduced by random data division, which can affect model reliability if not repeated or stratified properly.

c) Hold-Out and Nested Validation

A smaller proportion of studies utilized hold-out validation (where one subset is used exclusively for final evaluation) or nested cross-validation (for hyperparameter tuning); (Adejo and Connolly, 2018; Alalawi et al., 2023). These methods are particularly relevant in SVM and ANN models, which require extensive parameter optimization (Alhassan et al., 2020; Fundoni et al., 2023). Nested validation prevents data leakage during tuning and ensures a more realistic estimation of generalization performance.

2. Accuracy Metrics and Evaluation Criteria

To assess model performance, studies commonly employ a combination of accuracy-based, error-based, and probabilistic metrics. The selection of metrics depends on the model type (classification or regression) and research objective (e.g., predicting pass/fail outcomes vs. continuous GPA values).

a) Accuracy and Classification Rate

Accuracy, defined as the ratio of correctly classified instances to the total number of instances, remains the most widely reported metric (Abdelaziz et al., 2025; Daza et al., 2022). Across studies, Random Forest and ANN models consistently achieved the highest accuracy rates, ranging between 85% and 93% (Khan and Ghosh, 2021; Lam et al., 2024). Decision Trees and Logistic Regression models typically achieved accuracy between 75% and 85% (Meghji et al., 2023), while SVM models performed comparably in binary classification tasks (Syed Mustapha, 2023).

b) *Precision, Recall, and F1-Score*

Other studies evaluate precision (positive predictive value), recall (sensitivity), and F1 score (the harmonic mean of precision and recall) to assess a model’s balance between false positives and false negatives (Dai and Lu, 2024; Deleña et al., 2025; Kamal and Ahuja, 2019). These metrics are important in academic prediction tasks where misclassification can lead to inappropriate interventions, for example, mislabeling successful students as “at risk” (Dai and Lu, 2024). ANN and SVM models tend to achieve higher F1 scores due to their ability to model nonlinear relationships and complex data distributions (Khan and Ghosh, 2021; Syed Mustapha, 2023).

c) *ROC Curve and AUC*

The Receiver Operating Characteristic (ROC) Curve and the Area Under the Curve (AUC) are frequently used to evaluate a model’s discrimination ability, especially for binary outcome models (e.g., pass/fail, dropout/non-dropout) (Matz et al., 2023). Higher AUC values (> 0.85) were reported in Random Forest and ANN models, reflecting superior sensitivity-specificity trade-offs (Namoun and Alshantqi, 2020; Nguyen and Jones, 2022) especially in countries like

Vietnam, with rich biodiversity and a high population growth rate. One of the main causes of biodiversity loss in Vietnam is the unsustainable bushmeat consumption rate in urban areas. To help mitigate the demand for bushmeat, this study aims to examine the associations between biodiversity loss perceptions, perception toward the prohibition of illegal wildlife consumption, and bushmeat consumption behaviors among urban residents in Vietnam. The investigation employed the Bayesian Mindsponge Framework (BMF).

d) *Error-Based Metrics*

For regression-oriented prediction tasks (e.g., predicting continuous GPA), metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are widely employed (Khan and Menezes, 2020; Pek et al., 2022). Lower MAE and RMSE values indicate better model precision. ANN and Random Forest models consistently produce the lowest RMSE values across datasets, confirming their robustness in handling nonlinear and high-dimensional features (Pisner and Schnyer, 2020; Ripan et al., 2021; Salman et al., 2024).

3. *Comparative Analysis of Model Accuracy*

Algorithm	Common Validation Method	Reported Accuracy	Supporting Studies
Decision Tree (DT)	K-Fold CV (10-fold)	75–85%	(Akhatkulov et al., 2024; Bhimavarapu et al., 2025; Dai and Lu, 2024; Kumar et al., 2022)
Random Forest (RF)	K-Fold CV, Hold-Out	85–93%	(He et al., 2018; Salman et al., 2024)
Support Vector Machine (SVM)	K-Fold CV	80–88%	(Syed Mustapha, 2023)
Logistic Regression (LR)	Train–Test Split (80:20)	70–82%	(De-La-Cruz et al., 2022; Sarker et al., 2024; L. Zhang et al., 2021)
Artificial Neural Network (ANN)	Train–Test Split, Nested CV	88–92%	(Khan and Ghosh, 2021; Lam et al., 2024)

Table 4: Comparative analysis of model accuracy

Validation and accuracy assessment form the backbone of reliable educational data mining. The research data above demonstrates that Random Forest and ANN models, when validated through k-fold cross-validation or a nested approach, achieve the highest and most stable predictive performance. Furthermore, a multidimensional evaluation framework that combines accuracy, precision, recall, and AUC provides a more nuanced and ethical basis for the use of predictive analytics in education. This ensures that algorithms not only perform statistically well but also support fair, actionable, and student-centered outcomes.

DISCUSSION

The results of this study highlight that the effectiveness of ML models in predicting academic performance is not only determined by the complexity of the algorithm but also by the quality of the validation process, the diversity of predictor variables, and the suitability of evaluation metrics (Nazir et al., 2023; Patel and Patel, 2024). This research reveals several important theoretical and practical implications for

educational research, institutional decision-making, and data-driven policy development.

One of the most significant benefits of accurate predictive modeling is the early detection of at-risk students, especially if teachers have the necessary personal and professional skills to effectively interpret and act on data-driven insights (Cocca et al., 2025). Models such as Random Forests and ANNs achieve accuracy rates above 90% (Aghbashlo et al., 2015; Khan and Ghosh, 2021; Lam et al., 2024). By combining academic, behavioral, and psychosocial predictors, institutions can intervene promptly and allocate resources to the most vulnerable students (Chakrapani and Chitradevi, 2022; Rodríguez-Hernández et al., 2021). The success of implementing predictive systems depends heavily on developing teachers’ digital skills and adopting technology-based teaching methods (Can, Kerkez, and Manav, 2025). This predictive ability transforms machine learning models from solely diagnostic tools into proactive instruments for academic support and retention (Abdrakhmanov et al., 2024), ultimately enhancing graduate employability and

the efficiency of higher education systems (Blašková and Staňková, 2023).

The ethical and validated use of accurate machine learning models has significant implications for educational equity and policy design. Integrating demographic and socioeconomic predictors can reveal systemic disparities among students from different backgrounds (Cheng, Liu, and Jia, 2024; Sarker et al., 2024). In this process, the construction of teachers' professional identities plays a crucial role; Giles and Yazan (2023) highlight that teacher collaboration in addressing student diversity is highly influential in shaping how interventions are implemented. Teachers do not simply act as algorithm executors but as moral agents who ensure data is used to uncover patterns of injustice without falling into deterministic predictions (Giles and Yazan, 2023). To avoid reinforcing bias, demographic data should not be used to predict outcomes deterministically but rather to reveal patterns of inequality and inform equitable interventions that respect the diverse backgrounds and learning needs of all students.

Although high-accuracy models like ANNs and Random Forests offer strong predictive performance, their interpretability is often limited compared to simpler models like Decision Trees or Logistic Regression (Hoyos Osorio and Daza Santacoloma, 2023; Kamal and Ahuja, 2019). In an educational environment where transparency and accountability are paramount, this exchange between accuracy and interpretation must be carefully managed (Haneem et al., 2017). Therefore, a hybrid approach combining interpretable models with ensemble methods is recommended. Analyzing feature importance and Shapley Additive exPlanations (SHAP) can help explain model predictions in human-understandable terms. Additionally, teachers must be trained to critically interpret model outputs, ensuring decisions remain pedagogically grounded rather than dictated by algorithms. This training should build upon teachers' existing professional skills while developing new competencies specific to data-driven decision-making in various educational contexts (Cocca et al., 2025).

The results of this study have several important implications for institutional implementation. Machine learning models must be retrained periodically with new data to account for evolving patterns and learning curricula. The model should also be adaptable across various educational contexts, including secondary schools, universities, and online learning platforms, to ensure scalability and maximize its contribution to graduates' employability outcomes at different educational levels (Blašková and Staňková, 2023).

As an implementation step, this predictive system must be embedded in the LMS to provide real-time monitoring and be updated regularly to remain relevant to the dynamic curriculum (Abdrakhmanov et al., 2024; Can, Kerkez, and Manav, 2025). This integration will only be successful if a human-in-the-loop approach is implemented, with educators remaining central to interpretation and action (Haneem et al., 2017). Thus, the synergy between technological sophistication, educators' collaborative identity, and the validation of professional skills will create an educational ecosystem that is not only technically accurate but also efficient, equitable, and responsible toward the future of its graduates.

CONCLUSION

The current study synthesizes evidence from 58 empirical investigations regarding the application of ML and data mining techniques to predict students' academic performance.

Some of the algorithms compared are Decision Tree, Random Forest, SVM, Logistic Regression, and ANN. The results of this analysis reveal a comprehensive understanding of how data-driven models can improve educational decision-making, identify at-risk students, and enhance learning outcomes.

Four main categories of predictor variables were identified as the most influential in modeling academic performance: (1) Demographic factors such as socioeconomic status and parental education, (2) Academic indicators such as GPA, previous grades, and attendance, (3) Digital behavioral variables derived from e-learning interactions, and (4) Psychosocial attributes including motivation, stress, and emotional well-being.

This heterogeneous predictor integration consistently yields superior model performance, with Random Forest and ANN algorithms achieving accuracy exceeding 90% when validated through rigorous cross-validation procedures.

In terms of validation and evaluation, the findings indicate that k-fold cross-validation and train-test split remain the most reliable approaches for ensuring generalization. Comprehensive performance evaluation using accuracy, precision, recall, F1-score, and AUC provides a more profound understanding of model behavior beyond simple correctness. Such methodological rigor prevents overfitting and strengthens confidence in the model's applicability in real-world educational contexts.

There are several practical implications of these findings, including:

- Enabling early detection of at-risk students through continuous monitoring of academic and behavioral data;
- Supporting evidence-based policy design to address structural inequalities; and
- Facilitating personalized learning interventions tailored to individual needs and learning styles.

One consideration is that ethical and pedagogical principles must guide the application of ML in education. Transparency, fairness, and student privacy must remain paramount in all model implementation processes.

The model must also remain understandable to educators, ensuring that predictions enhance, rather than replace, human pedagogical insights.

This synthesis also identifies some research gaps and future directions.

There is an urgent need for:

- Increased inclusion of psychosocial and affective factors in predictive modeling.
- Integration of explainable AI frameworks to make ML models more transparent and trustworthy; and
- Development of adaptive learning analytics systems capable of providing real-time feedback and personalized intervention recommendations.

Finally, machine learning provides a robust and evidence-based framework for transforming educational assessment

and student support. When based on strong validation, ethical governance, and human-centered design, predictive models have the potential to make education more equitable, responsive, and effective, ensuring that every student receives

the support needed to reach their full academic potential.

ACKNOWLEDGEMENT

No funding was received for this study.

REFERENCES

- Abdrakhmanov, R., Zhaxanova, A., Karatayeva, M., Niyazova, G. Z., Berkimbayev, K. and Tuimebayev, A. (2024) 'Development of a Framework for Predicting Students' Academic Performance in STEM Education using Machine Learning Methods', *International Journal of Advanced Computer Science and Applications*, Vol. 15, No. 1, pp. 38–46. <https://doi.org/10.14569/IJACSA.2024.0150105>
- Abdelaziz, A. A., Shaker, H., Tolba, A. S. and Abdelfattah, F. (2025) 'Predictive learning analytics: Analyzing student participation and performance on Moodle', in: Hassanien, A. E., Darwish, A., and El-Askary, H. (eds.), *The International Conference on Advanced Intelligent Systems and Informatics, Studies in Systems, Decision and Control*, Vol. 601, Cham: Springer, pp. 373–384. https://doi.org/10.1007/978-3-031-92240-4_34
- Abdolasol, M. G. M., Hussain, S. M. S., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., Ali, J. A., Mekhilef, S. and Milad, A. (2021) 'Artificial neural networks based optimization techniques: A review', *Electronics*, Vol. 10, No. 21, pp. 26–89. <https://doi.org/10.3390/electronics10212689>
- Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. and Arshad, H. (2018) 'State-of-the-art in artificial neural network applications: A survey', *Heliyon*, Vol. 4, No. 11, pp. 09–38. <https://doi.org/10.1016/j.heliyon.2018.e00938>
- Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Umar, A. M., Linus, O. U., Arshad, H., Kazaure, A. A., Gana, U. and Kiru, M. U. (2019) 'Comprehensive review of artificial neural network applications to pattern recognition', *IEEE Access*, Vol. 7, pp. 158820–158846. <https://doi.org/10.1109/ACCESS.2019.2945545>
- Adejo, O. W. and Connolly, T. (2018) 'Predicting student academic performance using multi-model heterogeneous ensemble approach', *Journal of Applied Research in Higher Education*, Vol. 10, No. 1, pp. 61–75. <https://doi.org/10.1108/JARHE-09-2017-0113>
- Aghbashlo, M., Hosseinpour, S. and Mujumdar, A. S. (2015) 'Application of artificial neural networks (ANNs) in drying technology: a comprehensive review', *Drying Technology*, Vol. 33, No.12, pp. 1397–1462. <https://doi.org/10.1080/07373937.2015.1036288>
- Akhatkulov, S., Yusupov, O. and Omonov, A. (2024) 'Predicting students' future final exam results using machine learning algorithms: A comparative analysis', *AIP Conference Proceedings*, Vol. 3244, p. 030071. <https://doi.org/10.1063/5.0241786>
- Al-Khlifeh, E., Tarawneh, A. S., Almohammadi, K., Alrashidi, M., Hassanat, R. and Hassanat, A. B. (2025) 'Decision tree-based learning and laboratory data mining: an efficient approach to amebiasis testing', *Parasites and Vectors*, Vol. 18, No.1, pp. 1–18. <https://doi.org/10.1186/s13071-024-06618-6>
- Alalawi, K., Athauda, R. and Chiong, R. (2023) 'Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review', *Engineering Reports*, Vol. 5, No. 12, pp. 1–25. <https://doi.org/10.1002/eng2.12699>
- Albreiki, B., Zaki, N. and Alashwal, H. (2021) 'A systematic literature review of student' performance prediction using machine learning techniques', *Education Sciences*, Vol. 11, No. 9, p. 552. <https://doi.org/10.3390/educsci11090552>
- Alhassan, A., Zafar, B. and Mueen, A. (2020) 'Predict students' academic performance based on their assessment grades and online activity data', *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 4, pp. 185–194. <https://doi.org/10.14569/IJACSA.2020.0110425>
- Antonakis, A. C. and Sfakianakis, M. E. (2009) 'Assessing naïve Bayes as a method for screening credit applicants', *Journal of Applied Statistics*, Vol. 36, No. 5, pp. 537–545. <https://doi.org/10.1080/02664760802554263>
- Baashar, Y., Alkaws, G., Ali, N., Alhussian, H. and Bahbouh, H. (2021) 'Predicting student's performance using machine learning methods: A systematic literature review', in: *2021 International Conference on Computer and Information Sciences (ICCOINS)*, pp. 357–362. <https://doi.org/10.1109/ICCOINS49721.2021.9497185>
- Balaji, P., Alelyani, S., Qahmash, A. and Mohana, M. (2021) 'Contributions of Machine Learning Models towards Student Academic Performance Prediction: A Systematic Review', *Applied Sciences*, Vol. 11, No. 21, p. 10007. <https://doi.org/10.3390/app112110007>
- Baneres, D., Rodríguez-Gonzalez, M. E. and Serra, M. (2019) 'An Early Feedback Prediction System for Learners At-Risk within a First-Year Higher Education Course', *IEEE Transactions on Learning Technologies*, Vol.12, No.2, pp. 249–263. <https://doi.org/10.1109/TLT.2019.2912167>
- Bhimavarapu, N., Prasanthi, B. V., Lakshmi Veenadhari, C. H., Durga Satish, M., Matta, V. D. R. and Pradeep, I. K. (2025) 'Predicting student academic performance using machine learning: A comparison of classification algorithms', in: Smys, S., Tavares, J. M. R. S. and Balas, V. E. (eds.), *Springer Proceedings in Mathematics and Statistics*, Vol. 441, Singapore: Springer, pp. 703–716. https://doi.org/10.1007/978-3-031-51338-1_56
- Bilquise, G., Ibrahim, S. and Salhieh, S. M. (2024) 'Investigating student acceptance of an academic advising chatbot in higher education institutions', *Education and Information Technologies*, Vol. 29, No. 5, pp. 6357–6382. <https://doi.org/10.1007/s10639-023-12076-x>
- Blašková, V. and Staňková, M. (2023) 'Graduate Employability as a Key to the Efficiency of Tertiary Education', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 16, No. 4, pp. 262–274. <https://doi.org/10.7160/eriesj.2023.160401>
- Boujmiraz, S., Darhmaoui, H. and Drissi el Maliani, A. (2026) 'Predicting student performance: A comprehensive review of machine learning, deep learning, and explainable AI approaches', *Computers and Education: Artificial Intelligence*, Vol. 10, No. 1, p. 100548. <https://doi.org/10.1016/j.caeai.2026.100548>

- Can, S., Kerkez, F. İlker and Manav, G. . (2025) ‘Physical Education and Sports Teachers’ Perceptions to Benefit from Web 2.0 Tools in Face-to-face Education after Emergency Remote Teaching Process: A Mixed Method Research’, *Journal on Efficiency and Responsibility in Education and Science*, Vol. 18, No. 1, pp. 1–12. <https://doi.org/10.7160/eriesj.2025.180101>
- Chakrapani, P. and Chitradevi, D. (2022) ‘Academic performance prediction using machine learning: A comprehensive and systematic review’, in: *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, pp. 335–340. <https://doi.org/10.1109/ICESIC53714.2022.9783512>
- Cheng, B., Liu, Y. and Jia, Y. (2024) ‘Evaluation of students’ performance during the academic period using the XG-Boost Classifier-Enhanced AEO hybrid model’, *Expert Systems with Applications*, Vol. 238, p. 122136. <https://doi.org/10.1016/j.eswa.2023.122136>
- Cocca, A., Ciesralová, M., Cocca, M., Greier, K., Uchytíl, J. and Ruedl, G. (2025) ‘Validation of the Teachers’ Personal and Professional Skills Questionnaire in the Czech Physical Education Setting’, *Journal on Efficiency and Responsibility in Education and Science*, Vol. 18, No. 1, pp. 58–63. <https://doi.org/10.7160/eriesj.2025.180107>
- Cruz, M. M. P. and Lumauag, R. G. (2024) ‘Comparative analysis of machine learning algorithms for predicting student academic performance in higher education’, in: *Proceedings of the 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS 2024)*, pp. 888–896. <https://doi.org/10.1109/ICUIS64676.2024.10866086>
- Dai, Z. and Lu, P. J. (2024) ‘The application of machine learning in student performance prediction in higher education institutions: A systematic literature review’, in: *Proceedings of the 2024 13th International Conference on Computer Technologies and Development (TechDev)*, pp. 57–62. <https://doi.org/10.1109/TechDev64369.2024.00019>
- Daza, A., Guerra, C., Cervera, N. and Burgos, E. (2022) ‘Predicting Academic Performance through Data Mining: A Systematic Literature’, *TEM Journal*, Vol. 11, No. 2, pp. 939–949. <https://doi.org/10.18421/TEM112-57>
- De-La-Cruz, P., Rojas-Coaquira, R., Vega-Huerta, H., Pérez-Quintanilla, J. and Lagos-Barzola, M. (2022) ‘A Systematic Review Regarding the Prediction of Academic Performance’, *Journal of Computer Science*, Vol. 18, No.12, pp. 1219–1231. <https://doi.org/10.3844/JCSSP.2022.1219.1231>
- Deleña, R. D., Dia, N. J., Sacayan, R. R., Sieras, J. C., Khalid, S. A., Macatotong, A. H. T. and Gulam, S. B. (2025) ‘Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors’, *Systems and Soft Computing*, Vol. 7, p. 200352. <https://doi.org/10.1016/j.sasc.2025.200352>
- Farissi, A., Dahlan, H. M. and Shah, Z. A. (2023) ‘High accuracy feature selection using metaheuristic algorithm for classification of student academic performance prediction’, in: Abraham, A., Gandhi, N., Hanne, T. and Hong, T. P. (eds.), *Proceedings of International Conference on Intelligent Systems Design and Applications, Lecture Notes on Data Engineering and Communications Technologies*, Vol. 179, Cham: Springer, pp. 399–409. https://doi.org/10.1007/978-3-031-36258-3_35
- Fundoni, M., Porcu, L. and Melis, G. (2023) ‘Systematic literature review: Main procedures and guidelines for interpreting the results’, in: Bell, E., Bryman, A. and Harley, B. (eds.), *Researching and Analysing Business: Research Methods in Practice*, Abingdon: Routledge, pp. 55–74. <https://doi.org/10.4324/9781003107774-5>
- García-Martín, J. and Pérez Fernández, L. M. (2025) ‘A Review of Global Strategies for Achieving Sustainable Development Goal 4 in Higher Education (2020–2024): Key Actions in the Education for Sustainable Development Framework’, *Sustainable Development*, Vol. 14, No. S2, pp. 843–856. <https://doi.org/10.1002/sd.70374>
- Gholami, R. and Fakhari, N. (2017) ‘Support vector machine: Principles, parameters, and applications’, in: Saha, P., Saha, S. and Balas, V. E. (eds.), *Handbook of Neural Computation*, London: Academic Press, pp. 515–535. <https://doi.org/10.1016/B978-0-12-811318-9.00027-2>
- Giles, A. and Yazan, B. (2023) ‘Constructing teacher identity in teacher collaboration: What does it mean to be a teacher of culturally and linguistically diverse English learners?’, *Journal on Efficiency and Responsibility in Education and Science*, Vol. 16, No. 1, pp. 36–45. <https://doi.org/10.7160/eriesj.2023.160104>
- Haneem, F., Kama, N., Ali, R. and Selamat, A. (2017) ‘Applying data analytics approach in systematic literature review: Master data management case study’, *Frontiers in Artificial Intelligence and Applications*, Vol. 297, pp. 705–715. <https://doi.org/10.3233/978-1-61499-800-6-705>
- Lingjun, H., Levine, R. A., Fan, J., Beemer, J. and Stronach, J. (2018) ‘Random forest as a predictive analytics alternative to regression in institutional research’, *Practical Assessment, Research and Evaluation*, Vol. 23, no. 1, pp. 1–16. <https://doi.org/10.7275/1wpr-m024>
- Hellas, A., Ihtantola, P., Petersen, A., Ajanovski, V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C. and Liao, S. N. (2018) ‘Predicting academic performance: A systematic literature review’, in: *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018)*, pp. 175–199. <https://doi.org/10.1145/3293881.3295783>
- Hoyos Osorio, J. K. and Daza Santacoloma, G. (2023) ‘Predictive model to identify college students with high dropout rates’, *Revista Electrónica de Investigación Educativa*, Vol. 25, No. e13, pp. 1–10. <https://doi.org/10.24320/redic.2023.25.e13.5398>
- Kamal, P. and Ahuja, S. (2019) ‘An ensemble-based model for prediction of academic performance of students in undergrad professional course’, *Journal of Engineering, Design and Technology*, Vol.17, No. 4, pp. 769–781. <https://doi.org/10.1108/JEDT-11-2018-0204>
- Karim-Abdallah, B., Ayitey Junior, M., Appiahene, P., Harris, E. and Binful, D. K. (2025) ‘Application of Machine Learning Algorithms in Predicting Academic Performance of Students in Higher Education Institutes (HEIS): A Systematic Review and Bibliographic Analysis’, *African Journal of Applied Research*, Vol. 11, No. 1, pp. 536–559. <https://doi.org/10.26437/ajar.v11i1.869>
- Katarya, R. (2023) ‘A Systematic Review on Predicting the Performance of Students in Higher Education in Offline Mode Using Machine Learning Techniques’, *Wireless Personal Communications*, Vol. 133, No. 3, pp. 1643–1674. <https://doi.org/10.1007/s11277-023-10838-x>
- Khan, A. and Ghosh, S. K. (2021) ‘Student performance analysis and prediction in classroom learning: A review of educational data mining studies’, *Education and Information Technologies*, Vol. 26, No. 1, pp. 205–240. <https://doi.org/10.1007/s10639-020-10230-3>
- Khan, S. and Menezes, J. (2020) ‘Predictive modelling to illustrate factors influencing students at risk’, *International Journal of Technology Transfer and Commercialisation*, Vol. 17, No. 1, pp. 68–75. <https://doi.org/10.1504/IJTTC.2020.106574>
- Krüger, J., Lausberger, C., von Nostitz-Wallwitz, I., Saake, G. and Leich, T. (2020) ‘Search. Review. Repeat? An empirical study of threats to replicating SLR searches’, *Empirical Software Engineering*, Vol. 25, pp. 627–677. <https://doi.org/10.1007/s10664-019-09763-0>

- Kumar, M., Singh, A. J., Sharma, B. and Cengiz, K. (2022) 'Evaluation of machine learning algorithms on academic big dataset by using feature selection techniques', in: Balas, V. E., Solanki, V. K. and Kumar, R. (eds.), *Intelligent Network Design Driven by Big Data Analytics, IoT, AI and Cloud Computing*, London: Institution of Engineering and Technology, pp. 61–92. https://doi.org/10.1049/PBPC054E_ch4
- Kumar, R., Goswami, B., Mhatre, S. M. and Agrawal, S. (2024) 'Naive bayes in focus: a thorough examination of its algorithmic foundations and use cases', *International Journal of Innovative Science and Research Technology*, Vol. 9, No. 5, pp. 2078–2081. <https://doi.org/10.38124/ijisrt/IJSRT24MAY1438>
- Lam, P. X., Mai, P. Q. H., Nguyen, Q. H., Pham, T., Nguyen, T. H. H. and Nguyen, T. H. (2024) 'Enhancing educational evaluation through predictive student assessment modeling', *Computers and Education: Artificial Intelligence*, Vol. 6, p. 100244. <https://doi.org/10.1016/j.caeai.2024.100244>
- Lou, Y. and Colvin, K. F. (2025) 'Performance prediction using educational data mining techniques: a comparative study', *Discover Education*, Vol. 4, No. 112, pp. 1–14. <https://doi.org/10.1007/s44217-025-00502-w>
- Lutsenko, V. and Zgonnikov, M. (2024) 'Fault tolerant system for data storage, transmission and processing in fog computing using artificial neural networks', in: Kotenko, I., Badica, C. and Taratukhin, V. (eds.), *Proceedings of International Conference on Intelligent Data Engineering and Automated Learning, Lecture Notes in Networks and Systems*, Vol. 744, Cham: Springer, pp. 199–212. https://doi.org/10.1007/978-3-031-64010-0_19
- Masangu, L., Jadhav, A. and Ajoodha, R. (2021) 'Predicting student academic performance using data mining techniques', *Advances in Science, Technology and Engineering Systems*, Vol. 6, No.1, pp.153–163. <https://doi.org/10.25046/aj060117>
- Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A. and Stachl, C. (2023) 'Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics', *Scientific Reports*, Vol. 13, No. 1, pp. 5705. <https://doi.org/10.1038/s41598-023-32484-w>
- Meghji, A. F., Shaikh, F. B., Wadho, S. A., Bhatti, S. and Ayyasamy, R. K. (2023) 'Using educational data mining to predict student academic performance', *VFAST Transactions on Software Engineering*, Vol. 11, No. 2, pp. 43–49. <https://doi.org/10.21015/vtse.v11i2.1475>
- Molina, O. E., and Cancell, D. R. F. (2021) 'Is it possible to predict academic performance? An analysis from educational technology', *Revista Fuentes*, Vol. 3, No. 23, pp. 363–375. <https://doi.org/10.12795/REVISTAFUENTES.2021.14278>
- Nabil, A., Seyam, M. and Elfetouh, A. A. (2022) 'Predicting students' academic performance using machine learning techniques: a literature review', *International Journal of Business Intelligence and Data Mining*, Vol. 20, No. 4, pp. 456–479. <https://doi.org/10.1504/IJBIDM.2022.123214>
- Namoun, A. and Alshantqi, A. (2020) 'Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review', *Applied Sciences*, Vol. 11, No. 1, p. 237. <https://doi.org/10.3390/app11010237>
- Nawang, H., Makhtar, M. and Hamza, W. M. A. F. W. (2021) 'A systematic literature review on student performance predictions', *International Journal of Advanced Technology and Engineering Exploration*, Vol. 8, No. 84, pp. 1441–1453. <https://doi.org/10.19101/ijatec.2021.874521>
- Nazir, M., Noraziah, A., Rahmah, M. and Sharma, A. (2023) 'Examining the potential of machine learning for predicting academic achievement: A systematic review', *Fusion: Practice and Applications*, Vol. 13, No. 2, pp.71–90. <https://doi.org/10.54216/FPA.130207>
- Nguyen, M. H. and Jones, T. E. (2022) 'Predictors of support for biodiversity loss countermeasure and bushmeat consumption among Vietnamese urban residents', *Conservation Science and Practice*, Vol. 4, No. 12, p. e12822. <https://doi.org/10.1111/csp2.12822>
- Niranjala, S. H., Alobaedy, M. M. and Goyal, S. B. (2024) 'A comparative study of machine learning techniques for predicting student academic performance', in: Kotenko, I., Badica, C. and Taratukhin, V. (eds.), *Proceedings of International Conference on Intelligent Data Engineering and Automated Learning, Lecture Notes in Networks and Systems*, Vol. 811, Cham: Springer, pp. 307–315. https://doi.org/10.1007/978-3-031-73318-5_31
- Ortiz-Bejar, J., Tellez, E. S., Graff, M., Moctezuma, D. and Miranda-Jimenez, S. (2020) 'Improving k nearest neighbors and naïve Bayes classifiers through space transformations and model selection', *IEEE Access*, Vol. 8, pp. 221669–221688. <https://doi.org/10.1109/ACCESS.2020.3042453>
- Patel, H. I. and Patel, D. (2024) 'Exploratory Data Analysis and Feature Selection for Predictive Modeling of Student Academic Performance Using a Proposed Dataset', *International Journal of Engineering Trends and Technology*, Vol. 72, No. 11, pp. 131–143. <https://doi.org/10.14445/22315381/IJETT-V72I11P116>
- Pek, R. Z., Özyer, S. T., Elhage, T., Özyer, T. and Alhadj, R. (2022) 'The role of machine learning in identifying students at-risk and minimizing failure', *IEEE Access*, Vol. 11, pp. 1224–1243. <https://doi.org/10.1109/ACCESS.2022.3232984>
- Pelima, L., Sukmana, Y. and Rosmansyah, Y. (2024) 'Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review', *IEEE Access*, Vol. 12, pp. 23451–23465. <https://doi.org/10.1109/ACCESS.2024.3361479>
- Pisner, D. A. and Schnyer, D. M. (2020) 'Support vector machine', in: Mechelli, A. and Vieira, S. (eds.), *Machine Learning*, London: Academic Press, pp. 101–121. <https://doi.org/10.1016/B978-0-12-815739-8.00006-7>
- Prill, R., Karlsson, J., Ayeni, O. R. and Becker, R. (2021) 'Author guidelines for conducting systematic reviews and meta-analyses', *Knee Surgery, Sports Traumatology, Arthroscopy*, Vol.29, No. 9, pp. 2739–2744. <https://doi.org/10.1007/s00167-021-06631-7>
- Rajendran, S., Chamundeswari, S. and Sinha, A. A. (2022) 'Predicting the academic performance of middle- and high-school students using machine learning algorithms', *Social Sciences and Humanities Open*, Vol. 6, No. 1, p. 100357. <https://doi.org/10.1016/j.ssaho.2022.100357>
- Reddy, A. L., Sathish, T. and Sangeetha, N. (2024) 'Prediction of student results using novel random forest in comparison with decision tree to improve accuracy', *AIP Conference Proceedings*, Vol. 2853, p. 020053. <https://doi.org/10.1063/5.0198498>
- Reimers, F. M. (2024) 'The sustainable development goals and education, achievements and opportunities', *International Journal of Educational Development*, Vol. 104, p. 102965. <https://doi.org/10.1016/j.ijedudev.2023.102965>
- Ripan, R. C., Sarker, I. H., Hasan Furhad, M., Musfique Anwar, M. and Hoque, M. M. (2021) 'An effective heart disease prediction model based on machine learning techniques', in: Hassanien, A. E., Bhatnagar, R., Darwish, A. and Hameed, K. (eds.), *Proceedings of International Conference on Advanced Intelligent Systems and Informatics, Advances in Intelligent Systems and Computing*, Vol. 1375, Cham: Springer, pp. 280–288. https://doi.org/10.1007/978-3-030-73050-5_28

- Rodrigues, L. S., Santos, M., De Araújo Costa, I. P. and Moreira, M. (2022) 'Student Performance Prediction on Primary and Secondary Schools-A Systematic Literature Review', *Procedia Computer Science*, Vol. 214, pp. 680–687. <https://doi.org/10.1016/j.procs.2022.11.229>
- Rodríguez-Hernández, C. F., Musso, M., Kyndt, E. and Cascallar, E. (2021) 'Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation', *Computers and Education: Artificial Intelligence*, Vol. 2, p. 100018. <https://doi.org/10.1016/j.caeai.2021.100018>
- Rokach, L. (2016) 'Decision forest: Twenty years of research', *Information Fusion*, Vol. 27, No. 1, pp. 111–125. <https://doi.org/10.1016/j.inffus.2015.06.005>
- Roslan, M. H. B. and Chen, C. J. (2022) 'Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021)', *International Journal of Emerging Technologies in Learning*, Vol. 17, No. 5, pp. 147–179. <https://doi.org/10.3991/ijet.v17i05.27685>
- Salloum, S. A., Basiouni, A., Alfaisal, R., Salloum, A. and Shaalan, K. (2024) 'Predicting student retention in higher education using machine learning', in: Hassanien, A. E., Darwish, A. and El-Askary, H. (eds.), *Proceedings of International Conference on Advanced Intelligent Systems and Informatics, Communications in Computer and Information Science*, Vol. 2162, Cham: Springer, pp. 197–206. https://doi.org/10.1007/978-3-031-65996-6_17
- Salman, H. A., Kalakech, A. and Steiti, A. (2024) 'Random Forest Algorithm Overview', *Babylonian Journal of Machine Learning*, Vol. 2024, pp. 69–79. <https://doi.org/10.58496/BJML/2024/007>
- Salmi, J. (2015) 'Evidence-based policies in higher education: Data analytics, impact assessment and reporting', in: Curaj, A., Matei, L., Pricopie, R., Salmi, J. and Scott, P. (eds.), *The European Higher Education Area: Between Critical Reflections and Future Policies*, Cham: Springer International Publishing, pp. 807–813. https://doi.org/10.1007/978-3-319-20877-0_49
- Santiketa, N., Chaikhan, S., Ninrutsirikun, U. and Wattanakitrungraj, N. (2024) 'Student academic performance prediction using machine learning with various features and scenarios', in: *International Computer Science and Engineering Conference (ICSEC 2024)*, pp. 1–6. <https://doi.org/10.1109/ICSEC62781.2024.10770729>
- Sarker, S., Paul, M. K., Thasin, S. T. H. and Hasan, M. A. M. (2024) 'Analyzing students' academic performance using educational data mining', *Computers and Education: Artificial Intelligence*, Vol. 7, p. 100263. <https://doi.org/10.1016/j.caeai.2024.100263>
- Schmidt, A., Cechinel, C., Queiroga, E. M., Primo, T., Ramos, V., Bordin, A. S., Mello, R. F. and Munoz, R. (2025) 'Analyzing Intervention Strategies Employed in Response to Automated Academic-Risk Identification: A Systematic Review', *IEEE Revista Iberoamericana de Tecnologías Del Aprendizaje*, Vol. 20, pp. 77–85. <https://doi.org/10.1109/RITA.2025.3540161>
- Seo, E.-Y., Yang, J., Lee, J.-E. and So, G. (2024) 'Predictive modelling of student dropout risk: Practical insights from a South Korean distance university', *Heliyon*, Vol. 10, No. 11, pp. 1–17. <https://doi.org/10.1016/j.heliyon.2024.e30960>
- Shaik, A. B. and Srinivasan, S. (2019) 'A brief survey on random forest ensembles in classification model', in: Gunjan, V. K., Zurada, J. M. and Raman, B. (eds.), *Proceedings of International Conference on Recent Trends in Machine Learning, Lecture Notes in Networks and Systems*, Vol. 56, Singapore: Springer, pp. 253–260. https://doi.org/10.1007/978-981-13-2354-6_27
- Strub, O. (2020) 'Optimal feature selection for support vector machine classifiers', in: *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM 2020)*, pp. 304–308. <https://doi.org/10.1109/IEEM45057.2020.9309859>
- Sushma, P. G. and Sriramakrishnan, G. V. (2025) 'Exploring predictive algorithms: Linear regression and decision tree in student performance', *AIP Conference Proceedings*, Vol. 3270, p. 020154. <https://doi.org/10.1063/5.0264437>
- Syed Mustapha, S. (2023) 'Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods', *Applied System Innovation*, Vol. 6, No. 5, p. 86. <https://doi.org/10.3390/asi6050086>
- Vaarma, M. and Li, H. (2024) 'Predicting student dropouts with machine learning: An empirical study in Finnish higher education', *Technology in Society*, Vol. 76, p. 102474. <https://doi.org/10.1016/j.techsoc.2024.102474>
- Valdiviezo-Diaz, P. and Chicaiza, J. (2024) 'Prediction of academic outcomes using machine learning techniques: A survey of findings on higher education', in: Hassanien, A. E., Darwish, A. and El-Askary, H. (eds.), *Proceedings of International Conference on Advanced Intelligent Systems and Informatics, Communications in Computer and Information Science*, Vol. 2049, Cham: Springer, pp. 206–218. https://doi.org/10.1007/978-3-031-58956-0_16
- Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J. and Nawaz, R. (2020) 'Predicting academic performance of students from VLE big data using deep learning models', *Computers in Human Behavior*, Vol. 104, p. 106189. <https://doi.org/10.1016/j.chb.2019.106189>
- Wandekoken, E. D., Varejão, F. M., Batista, R. and Rauber, T. W. (2011) 'Support vector machine ensemble based on feature and hyperparameter variation for real-world machine fault diagnosis', in: Iliadis, L., Jayne, C. and Angelov, P. (eds.), *Engineering Applications of Neural Networks, Advances in Intelligent and Soft Computing*, Vol. 96, Berlin: Springer, pp. 271–282. https://doi.org/10.1007/978-3-642-20505-7_24
- Wickramasinghe, I. and Kalutaraage, H. (2021) 'Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation', *Soft Computing*, Vol. 25, No. 3, pp. 2277–2293. <https://doi.org/10.1007/s00500-020-05297-6>
- Wu, M., Subramaniam, G., Zhu, D., Li, C., Ding, H. and Zhang, Y. (2024) 'Using machine learning-based algorithms to predict academic performance: A systematic literature review', in: *4th International Conference on Innovative Practices in Technology and Management (ICIPTM 2024)*, pp. 1–8. <https://doi.org/10.1109/ICIPTM59628.2024.10563566>
- Xu, X., Wang, J., Peng, H. and Wu, R. (2019) 'Prediction of academic performance associated with internet usage behaviors using machine learning algorithms', *Computers in Human Behavior*, Vol. 98, pp.166–173. <https://doi.org/10.1016/j.chb.2019.04.015>
- Yassine, E. A. and Mohammed, K. (2024) 'A Comparative Analysis of Decision Trees, Bagging, and Random Forests for Predictive Modeling in Monetary Poverty: Evidence from Morocco', *Applied Mathematics and Information Sciences*, Vol. 18, No. 2, pp. 233–240. <https://doi.org/10.18576/amis/180203>
- Zhang, L., Li, K. F. and Bourguiba, I. (2021a) 'Recent advances in academic performance analysis', in: *International Conference on Higher Education Advances (HEAd21)*, pp. 607–614. <https://doi.org/10.4995/HEAd21.2021.13196>
- Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H. and Shang, X. (2021b) 'Educational data mining techniques for student performance prediction: method review and comparison analysis', *Frontiers in Psychology*, Vol. 12, p. 698490. <https://doi.org/10.3389/fpsyg.2021.698490>