Predicting Students' Academic Performance Via Machine Learning Algorithms: An Empirical Review and Practical Application

Edmund F. Agyemang[1,2*], Joseph A. Mensah[2], Obu-Amoah Ampomah[3], Louis Agyekum[4], Justice Akuoko-Frimpong[5], Amma Quansah[1], Oluwaferanmi M. Akinlosotu[6]
[1]School of Mathematical and Statistical Science, University of Texas Rio Grande Valley, USA.
[2]Department of Computer Science, Ashesi University, Berekuso, Eastern Region-Ghana.
[3]Department of Statistics, Western Michigan University, Kalamazoo-USA
[4]Trent University, Peterborough-Canada
[5]Department of Biostatistics, University of Michigan, Ann Abor-USA
[6]Department of Business Administration, Al-Hikmah University, Ilorin-Nigeria.
Corresponding author: edmundfosu6@gmail.com, ORCID ID:0000-0001-8124-4493

ABSTRACT

Predicting academic outcomes is complex and influenced by factors like socioeconomic background, motivation, and learning style. Machine Learning (ML) algorithms have become increasingly important due to their ability to analyze large data volumes and identify prediction patterns. Results show ML's success in predicting academic performance, though effectiveness varies by dataset, algorithm choice, and training features. There is no consensus on the most effective ML method for predicting student performance with broad applicability. Among the five algorithms evaluated in this study, the Random Forest Classifier emerged as the best model, achieving the highest G-Mean and accuracy of 0.9243 and 85.42% respectively. This model's performance emphasizes the importance of balanced sensitivity and specificity in predicting student academic performance. The empirical review highlights several challenges, including a lack of standardization in performance metrics, limited model generalizability, and potential bias in training data. It also notes the impact of individual and environmental factors on academic performance, emphasizing the role of instructors and policymakers in improving educational outcomes. The study provides insights into current trends in using ML algorithms for academic predictions, identifying conceptual, methodological, analytical, and ethical gaps. These gaps affect the validity and reliability of research, underscoring the need to address them for informed decision-making and improved learning outcomes.

Keywords: *Machine Learning Algorithms, Random Forest, Naive Bayes, Logistic Regression, Academic Performance*
DOI: 10.7176/CEIS/15-1-09
Publication date: September 30th 2024

1 Introduction

The use of machine learning (ML) algorithms in academia has gained significant attention in recent years due to the increasing availability of educational data and advancements in ML techniques (Yagcı, 2022). Using ML algorithms to predict students' academic performance can give valuable insights to educators, allowing them to identify at-risk students who may need additional support, modify instructional techniques, boost learning outcomes, tailor teaching approaches to specific students' requirements, and increase student retention rates (Adnan et al., 2021). This procedure promotes the growth of the educational system at higher institutions because educators and policymakers can intervene early to prevent students from falling behind and increase their chances of success (Pinkus, 2008). Applying ML algorithms to predict student academic achievement can dramatically enhance educational results and give valuable insights into the aspects contributing to academic success (Alyahyan and Du¨¸steg¨or, 2020). Therefore, it is critical to carefully assess these algorithms' possible benefits and limitations and ensure they are appropriately utilized. Mechanisms, such as the type of ML algorithm employed, the variables analyzed, and the assessment metrics used to determine prediction accuracy, were included as part of our investigation criteria. Applying ML algorithms in education can transform how we approach teaching and learning with a qualitative or quantitative analysis, or a mix of the two, offering an overall assessment of the condition of the results (Zhai, 2021). The potential benefits of using ML algorithms to predict academic performance extend beyond individual students and can positively impact society (Waheed et al., 2020). By improving education outcomes, individuals are better equipped to contribute to the workforce and society, leading to economic growth and social development (Chan, 2016). A vast majority of the work in educational data analysis has been devoted to developing machine learning models capable of accurately predicting students' performance in specific contexts. However, the existing body of literature often overlooks the crucial aspect of evaluating models for their ability to transcend beyond their original training settings and demonstrate robust generalizability across diverse student populations and learning environments. This oversight raises concerns about the potential biases introduced by relying solely on 'best-performing' models and neglecting the search for models that exhibit superior generalization capabilities. Consequently, a pressing need emerges to address this research gap by identifying and investigating the optimal machine learning model that can be a predictive tool for assessing students' performance. This pursuit emphasizes ensuring that the identified model achieves accurate predictions and avoids any inherent bias stemming from feature selection, thereby ensuring its applicability and effectiveness across various educational contexts. This study contributes to advancing educational data analysis practices by addressing these challenges and encouraging a paradigm shift towards holistic and unbiased model evaluation and selection.

With the increased availability of data from numerous sources, including learning management systems, online platforms, and student records, ML algorithms can give significant insights into student behavior, performance, and learning patterns (Yu et al., 2020). A thorough examination of the literature on ML algorithms for forecasting students' academic achievement may offer a complete knowledge of the various ML approaches employed, the parameters examined, and prediction accuracy (Rastrollo- et al., 2020). Institutions can profit from properly anticipating student performance by concentrating on students who are more likely to perform poorly and helping them improve their performance (Batool et al., 2023). ML algorithms used to predict students' academic achievement can give significant insights to academics, instructors, and educational policymakers (Waheed et al., 2020; Alyahyan and Du¨¸steg¨or, 2020). ML algorithms may effectively predict students' academic achievement by analyzing different academic and non-academic criteria such as previous grades, attendance records, socio-economic background, and student behavior (Batool et al., 2023).

A growing interest has been in using ML algorithms to predict students' academic performance and several studies have explored the use of ML in this area, with promising results. Several schools of thought regarding using ML to predict academic performance have emerged. One school of thought focuses on using traditional statistical methods, such as regression analysis and logistic regression. These methods assume a linear relationship between the predictor and outcome variables. For example, studies by Yaacob et al. (2019) and Waheed et al. (2020) used logistic regression while El Aissaoui et al. (2020) and Ibrahim and Rusli (2007) used linear regression to predict students' academic performance. Another school of thought revolves around using decision trees and random forests. Decision trees are hierarchical models that predict the outcome variable through binary decisions. On the other hand, random forests are an ensemble learning approach combining numerous decision trees. Vijayalakshmi and Venkatachalapathy (2019), Altabrawee et al. (2019) and Zhang et al. (2022) for example, employed decision trees and random forest to predict students' academic performance. A third school of thought focuses on using neural networks (Baashar et al., 2022; Liu et al., 2022), which are computational models inspired by the structure and function of the human brain to predict students' academic performance. Neural networks are beneficial when dealing with complex, non-linear relationships between variables. Hybrid approaches also combine multiple ML methods to improve prediction accuracy. For instance, a study by Francis and Babu (2019) used a hybrid approach that combined logistic regression, decision trees, and neural networks to predict academic performance based on students' demographic information, prior academic performance, and study habits. Each approach has its strengths and weaknesses, and the choice of method depends on the research question and the nature of the data.

Numerous factors can impact a student's academic performance, including individual factors, such as motivation and self-regulation, and environmental factors, such as socio-economic status and school resources (de la Fuente et al., 2021). Motivation is one of the most significant factors impacting students' academic performance. Research has shown that students who are intrinsically motivated to learn, meaning that they are motivated by their interest and enjoyment of the material, are more likely to perform well academically (Ryan and Deci, 2000). On the other hand, extrinsically motivated students, meaning that they are motivated by external rewards such as grades or praise, may be less likely to perform well if these rewards are not provided (Kusurkar et al., 2013). Another individual factor that can impact academic performance is self-regulation, which refers to the ability to manage one's learning and behaviors (Morrison et al., 2010). Students who can effectively regulate their learning by setting goals, monitoring their progress, and seeking help when needed are likelier to perform well academically (Feeney et al., 2023). Environmental factors can also have a significant impact on academic performance (Asvio et al., 2022). For example, students from lower socio-economic backgrounds may have less access to resources such as high-quality schools, educational materials, and extracurricular activities, which can negatively impact their academic performance (Onyancha et al., 2015). Additionally, students who attend schools with fewer resources, such as low-income schools, may be less likely to have access to experienced teachers or advanced coursework, which can also impact academic performance (Borman and Dowling, 2010). Family support and involvement in education can also have an impact on student performance. Research has shown that students whose families are involved in their education, such as providing support and encouragement, attending parent-teacher conferences, and monitoring homework, are likelier to perform well academically (Epstein and Sheldon, 2002). To improve student learning concentration and collaboration in response to the COVID-19 pandemic, Nyarko et al., (2023) utilize Discrete Choice Experiment to investigate university instructors' preferences for current teaching strategies.

Acknowledging the potential limitations of an empirical literature review of ML algorithms in predicting students' academic performance is essential. These limitations may include publication bias where there is a possibility that the review will favor studies reporting positive results since studies reporting negative findings may be less likely to be published. The heterogeneity of studies (different features considered) is another drawback since various ML techniques, factors considered, and evaluation metrics may differ among the studies included in the review, making it challenging to compare findings. Lastly, the generalizability of findings is also a major limitation since there is a possibility that the findings from the review may not be generalizable to other educational contexts or populations. As a result, it is important to check the overall performance of the various ML models used in predicting students' academic performance taking into consideration its generalizability to other educational contexts and the features used.

The remainder of the paper is organized as follows: Section 2 provides an empirical review of some ML models utilized for the prediction of students' academic performance. Section 3 discusses a practical application of five ML models utilized for the prediction of students' academic performance. Section 4 presents some research gaps in the literature worth investigating. Section 5 concludes the research and provides recommendations for further work.

2 Empirical Literature Review

Many studies have underscored the need for ML algorithms in academia. According to Alzubi et al. (2018), data analysis using ML algorithms can uncover hidden relationships and patterns in complex data sets that may be hard to detect by traditional statistical methods. Researchers can therefore develop more accurate prediction models for various applications, such as predicting disease outcomes or climate change impacts. Also, ML algorithms allow data-driven decisions to be made. They can assist researchers in making better decisions by recommending experimental designs, optimizing experimental designs, and determining which variables or features will be most useful (Singh et al., 2016). Furthermore, the applications of ML algorithms span a wide range of academic disciplines, including healthcare, economics, social sciences, environmental sciences, and computer science Sharma and Juneja (2017); Shailaja et al. (2018). This makes them highly relevant and applicable to academia because they can provide valuable insights and solutions to complex problems. The research work of Alsariera et al. (2022) reviewed the most recent ML algorithms and factors used to predict student academic success. The authors examined 39 research from 2015 to 2021 and discovered that academic characteristics, internal evaluations, demographics, and family/personal attributes all had a substantial impact on student performance prediction. According to the findings of the study, the most efficient predictor of student academic performance was found to be the KNN classifier, followed by the DT approach. To accurately forecast student academic success, however, a full understanding of the elements and qualities that drive student accomplishment is required. According to the authors, there is still a substantial opportunity for improvement in the design of measurement instruments used in instructional performance evaluation. For increased accuracy, new input variables and a larger dataset are required to overcome the methodological and analytical gap. To explore the environment-dependent characteristics not addressed in the present research, data should be collected from diverse institutions. For a more efficient categorization approach, the authors also suggested refining the selection of features based on their relationship.

Ya˘gcı (2022) introduces a novel ML model to predict undergraduate students' final test marks based on their midterm exam grades. The performance of six distinct algorithms, including RF, ANN, SVM, LR, NB, and KNN, were examined and compared. The study focused on two main themes: predicting academic success based on prior attainment grades and comparing ML algorithm performance indicators. The suggested model (educational data mining) attained a classification accuracy of 70-75% utilizing only three parameters: midterm exam marks, department data, and faculty data. The survey indicates that students' midterm test marks are a significant predictor of their final exam grades, and the algorithms utilized in the study predicted final exam grades with a high accuracy rate. The findings of the study were compared to previous research that predicted academic performance using various demographic and socio-economic characteristics. Future studies could investigate integrating new input variables and ML methods in their modelling approach, according to the authors. The author advocates for the use of data mining techniques to analyze students' learning behaviors, detect issues, enhance the educational environment, and make data-driven decisions.

Ouatik et al. (2022) used personal and academic data, as well as artificial intelligence and data mining methods such as KNN, C4.5, and SVM, to forecast students' academic performance. However, when the number of students, specializations, learning approaches, and data sources increased, Big Data technology was used to distribute processing and shorten execution time. The Sequential minimal optimization (SMO) and SVM algorithms were discovered to be the most efficient, with the highest classification rate (87.32%) and the shortest execution time. The most important factors in predicting student accomplishment, according to the study, were academic evaluation, economic position, parent educational level, distance from home, student interest, mental problem, and number of accesses to the virtual classroom. The findings show that in forecasting academic performance, artificial intelligence and data mining can be valuable tools when combined with Big Data technologies to process enormous amounts of data efficiently. The research only adjudged SMO and SVM as the models with high classification rates without its generalizability to other educational contexts.

Obsie and Adem (2018) conducted research at Hawassa University's Faculty of Computer Science on predicting undergraduate students' academic performance using ANN, LR, and SVR. 134 students who graduated between 2015 and 2017, with 52 (38.81%), 39 (29.10%), and 43 (32.09%) completing the last semester of each year made up the dataset. The study revealed that the ANN time prediction was 0.9763, SVR was 0.9805, and LR was 0.9805 after organizing the data in a Microsoft Excel sheet. The best accurate prediction time for neural networks was 0.78 seconds, 0.03 seconds for SVR, and 0.05 seconds for LR. The ANN approach produced the least accurate prediction result for all cases. The tests produced reliable results that may be used to forecast graduation CGPA. SVR and LR approaches, in particular, outperformed ANN in predicting the final Cumulative Grade Point Average (CGPA). The researchers advocated for the use of SVR and LR approaches to forecast final CGPA8, and the models may also be utilized to create a Student Performance Prediction System (SPPS) in an institution. The study conducted by Zhang et al. (2022) utilized transcript data and tree-based ML algorithms to forecast the academic performance of students in their undergraduate program. The courses in the program were divided into six groups, and the average GPAs of each category were used as important inputs for the prediction. Three tree-based ML models, namely, DT, Gradient Boosting Decision Tree (GBDT), and RF, were used for the analysis. The RF model was found to have the capability of identifying over 80% of students who were at risk of low academic performance by the end of the second semester. This is a significant finding because targeted interventions could be implemented promptly to enhance the quality of teaching and learning in the department. Furthermore, the importance of features and the DT structure were analyzed to obtain valuable information for both students and teachers.

In the study of Alshdaifat et al. (2022), the authors utilized the Mutual Information algorithm in conjunction with five different ML models that employ both classification and regression classifiers. The objective was to predict the academic performance of students. The five models used were Gaussian Naive Bayes, SVM, RF, KNN, and LR. The results showed that the SVM model had the highest prediction accuracy of 81.67%, followed by RF, KNN, LR, and Gaussian Naive Bayes with accuracy scores of 78.33%, 75.00%, 74.17%, and 50.83%

respectively. Additionally, the study found that the top three predictors of academic performance were the frequency of visits to online resources, the number of times the students raised their hands, and the number of days absent from class. The analysis of the study suggested that a strong correlation exists between academic performance and the students' behavior and attitude. However, the study was limited by the small sample size of the dataset leading to a methodological gap, which made it difficult to establish significant relationships within the data. A larger dataset could have led to greater prediction accuracy and provided a larger sample space for training and testing the models. The use of SVM in generalizing to other educational contexts has not been explored presenting a conceptual gap due to lack of attention to historical or cultural differences. In Vijayalakshmi and Venkatachalapathy (2019), a student performance prediction system was proposed, and different ML algorithms such as DT-C5.0, Naïve Bayes, RF, SVM, KNN, and ANN with respective accuracies of 69%, 73%, 79%, 69%, 75% and 84% were applied in R Programming and tested with Kaggle dataset. The results showed that the ANN algorithm with 84% accuracy rate outperformed the other algorithms. Moreover, the study only suggested that ANN algorithm is the best model to predict students' academic performance, whereas empirical literature presents conflicting results on which ML model is best in generalizing to other education population or contexts.

Waheed et al. (2020) evaluated the efficacy of a deep learning model for early prediction of student performance and prompt intervention by institutions to adopt remedial measures for student assistance and coaching. The study discovered that demographic and topographical factors had a considerable effect on educational outcomes. For predicting at-risk students, the deep learning model obtained remarkable accuracy levels, with a sensitivity of 69%, precision of 93%, and total accuracy of 88%. The sensitivity for predicting early withdrawals was 86%, the precision was 96%, and the total accuracy was 93%. Similarly, in discriminating between 'distinction' and 'fail' occurrences, the model attained a sensitivity of 74%, precision of 81%, and total accuracy of 85%. The study, however, encountered a class imbalance problem in 'distinction' situations, which was a restriction. A host of models, including Fuzzy C-means (FCM), Multi-Layer Perceptron (MLP), LR and RF were evaluated by Baig et al. (2023). After preprocessing the dataset, clusters were generated using the FCM approach. Following that, the preprocessed data was categorized using several classification methods such as MLP, LR, and RF. To increase precision, the FCM approach was coupled with each classification algorithm. When compared to typical ML algorithms such as KNN, SVM, and deep neural networks (DNN), the findings revealed that the combination of FCM with MLP and FCM with LR produced the greatest accuracies. FCM-MLP and FCM-LR combinations attained accuracies of 95.833% and 88.333%, respectively. The accuracy of the FCM-RF combination was also 95.833%. The study found that combining FCM with MLP and FCM with RF produced the best accurate predictions of student performance while silent on the model's generalizability to other academic domains and best features to use in prediction.

Kour et al. (2021) employed a variety of ML methods to predict student performance, including LR, DT, KNN, ANN, and SVR. Educational institutions may utilize this data to detect disadvantaged students, minimize dropout rates, and obtain optimal results. When the results of these algorithms were compared, it was shown that LR performed the best in forecasting student performance, with a mean absolute error (MAE) of 0.803 using a 10-fold cross-validation test mode. The ANN, on the other hand, was determined to be the least efficient method, with an MAE of 1.183 utilizing the percentage split test scenario. Additionally, the study found that the grades gained in the second semester were the most useful feature for forecasting student achievement, while the number of failures was the least relevant. To address the methodological gap in future research on student performance, the authors suggested using big data sets, alternative filters in pre-processing in the WEKA explorer software, and ensemble ML techniques by modifying attribute values.

The research work of Chui et al. (2020) offers a new method, the conditional generative adversarial network based deep SVM (ICGAN-DSVM), to predict student performance in supportive learning settings such as school and family tutoring. The approach uses deep learning architecture to address the issue of limited sample size in academic datasets by expanding data volume; accounting for methodological gap and improving prediction accuracy. For 10-fold cross-validation, the specificity, sensitivity, and area under the receiver operating characteristic curve (AUC) of ICGAN-DSVM were 0.968, 0.971, and 0.954, respectively. Integrating both school and family tutoring into the prediction model enhances performance even further. In terms of performance metrics, the suggested ICGAN-DSVM surpasses related works by 8-29%. There is also a comparison between standard conditional generative adversarial networks (CGAN) and kernel design using heuristic-based multiple kernel learning (MKL). The study indicates the importance of ICGAN and DSVM in predicting student performance, as well as the superiority of ICGAN-DSVM over comparable works. Even though the authors accounted for the methodological gap, the study only introduced a new ML algorithm for predicting students' academic performance without testing the best model that can generalize to other academic jurisdictions presenting a conceptual gap.

One study by Altabrawee et al. (2019) employed four different ML algorithms, namely FF-ANN, NB, DT and LR, to aid in identifying students (belonging to archaeology and sociology departments) who have poor academic performance in Al-Muthanna University. The study used sample survey data collected from students and students' previous grades with 160 student records and twenty (20) attributes. A three folds cross-validation method was used; one-fold for the test set and the other two for the training set. The model comparison showed that the FF-ANN model achieved the highest ROC performance index and prediction accuracy of 80.07% and 77.04%, respectively. Moreover, the DT model proved that the most significant predictors of student academic performance are conducive educational environment, accommodation, grade in a computer course, computer studies interest and residency. The small sample considered introduces both a methodological and analytical gap. Moreover, the research was silent on the best model that can be used to improve student learning outcomes and objectives, presenting a conceptual gap.

To determine significant factors that affect students' academic performance in a secondary school setting, Beckham et al. (2023) resorted to using three ML techniques, including Multi-Layer Perceptron (MLP), DT, and RF. The dataset was extracted from Kaggle and consisted of 395 students with 33 features, including age, family support, parental status, interest in further education, internet access, father's occupation, past failures, etc. First, a correlation analysis was conducted to identify the weight of each factor affecting student academic performance. The study observed that student performance is highly skewed toward past failures, age, mother education, and interest in further education. MLP has been judged the best model with an RMSE of 4.32, followed by RF and DT with respective RMSEs of 4.52 and 5.69. Since data from only two Portuguese schools in the mathematics subject were used for the data analysis, the authors recommended using vast amounts of data from around the world to overcome methodological and analytical gaps.

The academic performance of students in a university setting was conducted by Ahmed et al. (2021) using four ML algorithms (LR, NB, SVM and GBDT). Particularly, the core of the research was to predict the exam score of students at an earlier date so that measures can be taken to help students at risk of failing. The dataset consisted of 450 university students and 20 features including age, college, department, sex, extra work, midsemester grade, etc. Respective accuracies of 82.6%, 86.2%, 88.80% and 89.1% were observed for LR, NB, SVM and GBDT ML models which implies that GBDT is a better predictor of academic performance. There was no statement of ethical consideration in the study leading to the ethical gap which is of high priority to researchers. In predicting student academic performance at the end of a semester, Hasan et al. (2020) employed supervised data classification models such as RF and CN2 Rule Inducer algorithm. The researchers made room for ethical approval from respondents by means of informed consent before data collection to take care of ethical gap. 772 samples of students record with 12 features from one academic year were used. RF outperformed the other ML algorithm, CN2 Rule Inducer algorithm (87.4%) in predicting successful students with a prediction accuracy of 88.3%. Features with greater importance were extracted using PCA (with 95.5% variance) and genetic algorithm. In general, it was observed that categorical attributes performed better than continuous attributes. The authors only suggested predicting students' performance on a weekly basis to make room to cater for the needs of underperforming students but were silent on the model's generalizability to different disciplines.

2.1 Table of ML Models with their accuracy, precision and recall

Table 1 summarizes some ML algorithms used in the prediction of students' academic performance.

Table 1: ML algorithms with their references, dataset, accuracy, precision and recall

Algorithm	Type	References	Dataset	Accuracy	Precision	Recall
DT	Non-Linear	Ahmed et al. (2021)	450 university students	89.1%	89.6%	99.3%
		Bernacki et al. (2020)	337 undergraduate students	63.2%	47.6%	75.3%
		Altabrawee et al. (2019)	161 university student records	76.9%	77.9%	77.8%
		Francis and Babu (2019)	Student data	66.0%	34.5%	55.6%
		Adnan et al. (2021)	OULAD	91.8%	92.0%	91.8%
RF	Non-linear	Sixhaxa et al. (2022)	Educational dataset from kaggle	78.3%	75.3%	77.9%
		Ya˘gcı (2022)	1854 students of Turkey	74.6%	75.2%	74.6%
		Adnan et al. (2021)	OULAD	92.0%	92.2%	92.0%
		Sixhaxa et al. (2022)	xAPI-Edu data from Kaggle	78.3%	77.9%	75.3%
SVM	Non-linear	Ahmed et al. (2021)	450 university students	88.8%	88.9%	99.3%
		Sixhaxa et al. (2022)	Educational dataset from kaggle	81.7%	81.4%	83.7%
		Ya˘gcı (2022)	1854 students of Turkey	73.5%	73.5%	73.5%
		Adnan et al. (2021)	OULAD	81.6%	90.3%	80.8%
		Francis and Babu (2019)	Student data	64.2%	32.8%	55.6%
		Sixhaxa et al. (2022)	xAPI-Edu data from Kaggle	81.7%	83.7%	81.4%
LR	Linear	Ahmed et al. (2021)	450 university students	82.6%	90.7%	89.8%
		Sixhaxa et al. (2022)	Educational dataset from kaggle	74.2%	74.6%	75.4%
		Ya˘gcı (2022)	1854 students of Turkey	71.7%	70.0%	71.7%
		Bernacki et al. (2020)	337 undergraduate students	67.4%	57.1%	75.3%
		Altabrawee et al. (2019)	161 university student records	74.5%	78.9%	71.9%
		Sixhaxa et al. (2022)	xAPI-Edu data from Kaggle	74.2%	75.4%	74.6%
NB	Linear	Ahmed et al. (2021)	450 university students	86.2%	94.5%	94.5%
		Sixhaxa et al. (2022)	Educational dataset from kaggle	50.8%	60.6%	59.1%
		Ya˘gcı (2022)	1854 students of Turkey	71.3%	70.6%	71.3%
		Bernacki et al. (2020)	337 undergraduate students	65.0%	53.1%	74.2%
		Altabrawee et al. (2019)	161 university student records	66.5%	67.3%	64.3%
		Francis and Babu (2019)	Student data	52.8%	27.1%	63.9%
		Sixhaxa et al. (2022)	xAPI-Edu data from Kaggle	50.8%	59.1%	60.6%
ANN	Non-linear	Ya˘gcı (2022)	1854 students of Turkey	74.6%	74.8%	74.6%
		Altabrawee et al. (2019)	161 university student records	77.0%	79.2%	77.9%
		Francis and Babu (2019)	Student data	64.2%	34.3%	63.9%
KNN	Non-Linear	Sixhaxa et al. (2022)	Educational dataset from kaggle	75.0%	75.8%	77.6%
		Ya˘gcı (2022)	1854 students of Turkey	69.9%	69.1%	69.9%
		Adnan et al. (2021)	OULAD	89.7%	89.9%	89.7%
		Sixhaxa et al. (2022)	xAPI-Edu data from Kaggle	75.0%	77.6%	75.9%

DT: Decision Tree; RF: Random Forest; SVM: Support Vector Machine, LR: Logistic Regression; NB: Naïve Bayes; ANN: Artificial Neural Networks; KNN: K-Nearest Neighbor; OULAD: Open University Learning Analytics Dataset.

3 Practical Application

This section provides a statistical analysis of student academic performance using the xAPI-Edu-Data from Kaggle.

3.1 Data and Methods

The data consists of 480 observations each containing 16 features with a common target (class) with three categories. The data is freely available at Kaggle and can be assessed through the repository at https://www.kaggle.com/search?q=xAPI-Edu+. The description of the data is given in Table 2:

Table 2: Description of Extract of the Student's Academic Performance Dataset

Attribute	Description	Unit	Data Type
Gender	1 = Male, 0= Female	1,0	Nominal
Class	1: Lower Level 2: Middle School 3: High School	1,2,3	Nominal
Section ID	0: Class A 1: Class B 2: Class C	0,1,2	Nominal
Semester	0: First Semester 1: Second Semester	0,1	Nominal
Responsible for student	0: Father 1: Mum	0,1	Nominal
Parent Answering Survey	0: No 1: Yes	0,1	Nominal
Parent School Satisfaction	0: No 1: Yes	0,1	Nominal
Student Absence Days	0: Under 7 1: Above 7	0,1	Nominal
Raised hand	Number of times student raise hands	0-100	Numeric
Visited resources	Number of times student visits a course content	0-100	Numeric
Viewing announcements	Number of times student checks announcement	0-100	Numeric
Discussion groups	Number of times student participated in discussions	0-100	Numeric

Before employing the various machine learning algorithms to the data, the data was split into training (80%=384) and testing (20%=96). The train data was used to develop the ML models while the test data was used to assess the model's performance. A standard scaler was adopted as a technique to scale the features to make all the features have equal contributions to the result of the study. We also conducted extensive hyperparameter tuning for the algorithms under consideration. For example, for the K-Nearest Neighbors (KNN) classifier, error rates across various k values (from 1 to 40) were evaluated to identify the optimal number of neighbors for the model. We first initialize an array to store error rates, then iterates through each k value, creating and training a KNN classifier for each iteration. Predictions were made on the test data, and the error rates were calculated and stored. The error rates were then plotted against the corresponding k values to visually determine the k value that minimizes the error rate for the KNN classifier by balancing model accuracy and performance. Various ML algorithms considered in this research included Logistic Regression, Random Forest Classifier, Support Vector Classifier, Decision Tree Classifier and K Nearest Neighbor Classifier. We relied on evaluation metrics such as recall, specificity, precision, F1 score, accuracy and AUC scores to determine the optimal models.

3.2 Descriptive Statistics

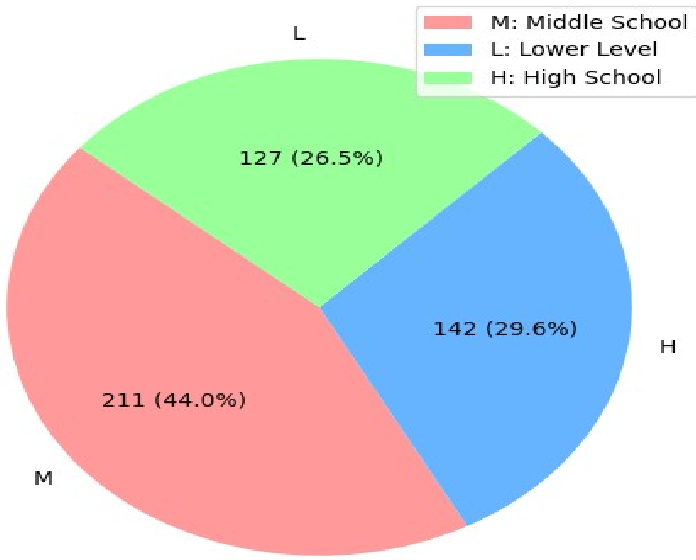


Figure 1: Number and Percentage of Students in Each Class

The pie chart in Figure 1 illustrates the distribution of students across three educational levels: Middle School (M), Lower Level (L), and High School (H). Middle School has the highest number of students, with 211 students, making up 44.0% of the total student population. High School follows with 142 students, representing 29.6%, while Lower School has 127 students, accounting for 26.5%.

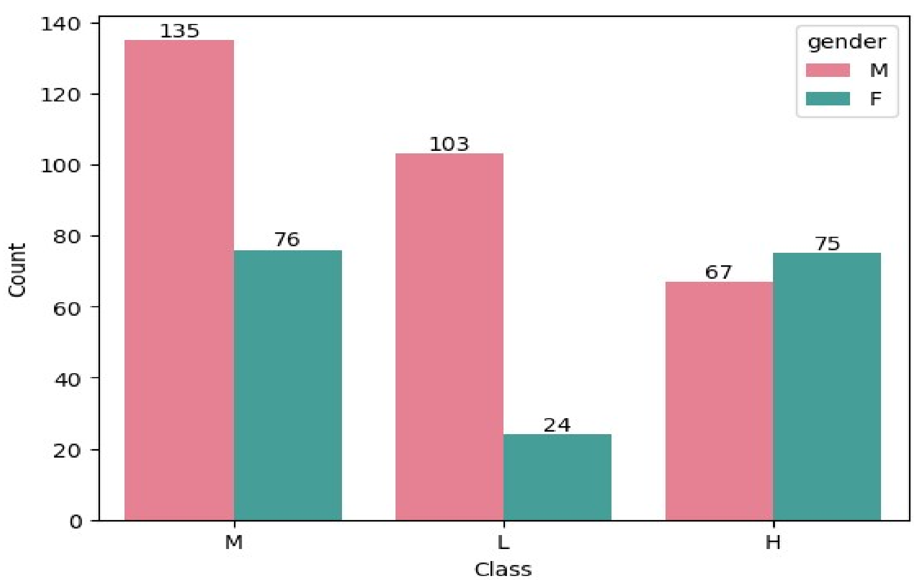
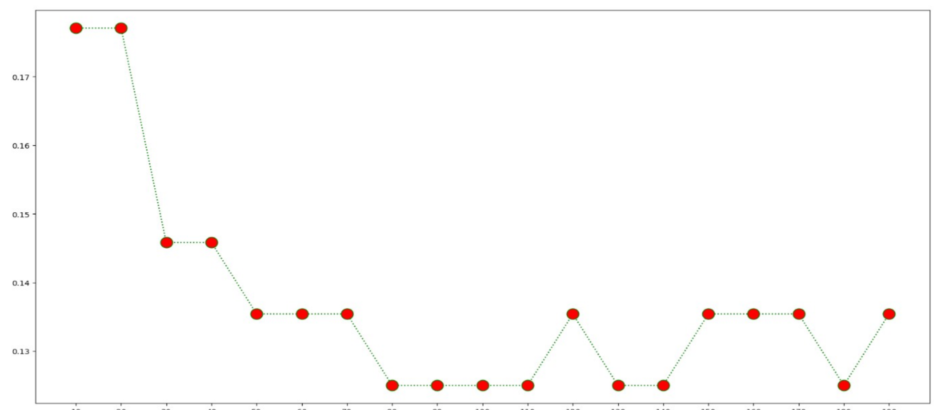


Figure 2: Number of Students in Each Class by Gender

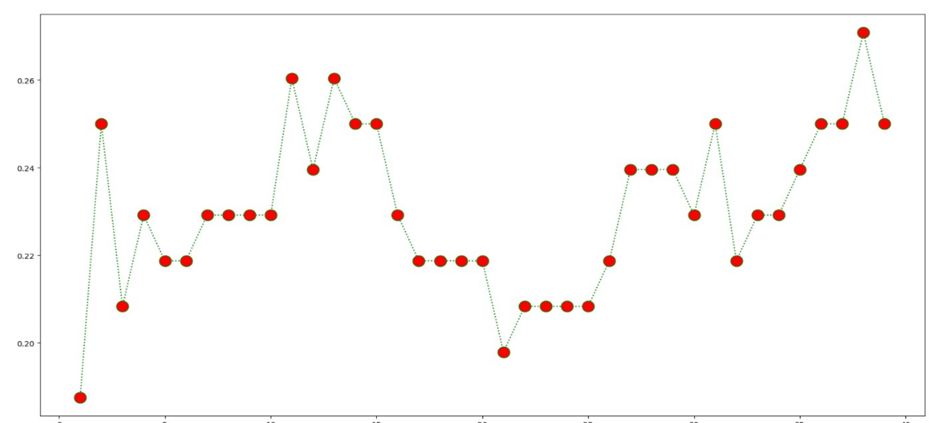
Figure 2 depicts the distribution of students across different school classes (Middle, Lower, and High) categorized by gender. Middle School (M) has the highest number of students, with 135 males and 76 females. Lower School (L) follows, with 103 males and significantly fewer females at 24. High School (H) has a more balanced distribution, with 67 males and 75 females. The chart clearly indicates a higher male student population across all classes, particularly pronounced in Lower School. Each bar is annotated with the exact count of students, providing a clear and immediate understanding of the gender distribution within each class category.

3.3 Analysis of the ML Algorithms-Hyperparameter Tunings

Using grid search cross validation with five (5) folds, the optimal hyperparameters for the various ML algorithms were derived as follows. For the random forest, fitting 5 folds for each of 120 candidates, with a total of 600 fits yielded the best validation score of 78.91%. The optimal hyperparameters chosen were n_estimators=80, max_depth=10, min_samples_leaf=1, min_samples_split=2. For the KNN classifier, even though k=1 was the nearest neighbor with the least error rate we did not chose 1 as it is so sensitive to just rely on 1 neighbor. We chose the next k with the least error which was $k = 21$. For the Decision Tree Classifier, fitting 5 folds for each of 90 candidates, with a total of 450 fits yielded the best validation score of 70.84%. The optimal hyperparameters chosen were 'criterion': 'gini', 'max_depth': 8, 'min_samples_leaf': 2, 'min_samples_split': 5. For the Logistic regression, a maximum iteration of 100 was considered and a moderate regularization strength of 1 was chosen to find a balance between fitting the training data well and maintaining good generalization to unseen data. Lastly, the SVC employed the radial basis function (RBF) as its kernel. The plots of the choice of the number of estimators in the Random Forest model and the number of nearest neighbors is depicted in Figure 3.



(a) Error Rates versus Number of Estimators of Random Forest Model



(b) Error Rates versus K Values of KNN Classifier

Figure 3: Plots of optimal number of estimators and nearest neighbors.

3.4 Machine Learning Models Performance Evaluations

Table 3: Weighted Averages of ML Model Comparison Results for Testing Data

Models	Precision	Sensitivity	Specificity	F1-Score	G-Mean	Accuracy (%)
RF	0.8337	0.8542	1.0000	0.8534	0.9243	85.42
KNN	0.8172	0.8125	1.0000	0.8106	0.9014	81.25
DT	0.8386	0.8333	1.0000	0.8326	0.9129	83.83
Logistic	0.8135	0.8125	1.0000	0.8103	0.9014	81.25
SVM (RBF)	0.8268	0.8229	1.0000	0.8226	0.9071	82.29

Table 3 presents the performance metrics of five different machine learning algorithms—Logistic Regression, Random Forest Classifier (RF), Support Vector Machine (SVM) with RBF kernel, Decision Tree Classifier (DT), and K-Nearest Neighbor Classifier (KNN)—in predicting students' academic performance. Among these models, the Random Forest Classifier demonstrates the highest G-Mean value of 0.9243, indicating it provides a balanced performance between sensitivity and specificity, making it the best model in this context. It also shows the highest accuracy at 85.42%, the highest F1-score at 0.8534, and competitive precision and sensitivity. The Decision Tree Classifier follows with a G-Mean of 0.9129 and an accuracy of 83.83%, indicating robust performance as well. SVM with RBF kernel follows with a G-Mean of 0.9071. The Logistic and KNN models exhibit the same G-Mean, sensitivity and accuracy values of 0.9014, 0.8125 and 81.25% respectively but Logistic Regression lags slightly in precision. Overall, while all models maintain high specificity, the Random Forest Classifier stands out as the best performer based on the G-Mean metric, highlighting its effectiveness in accurately predicting academic performance with a balanced approach.

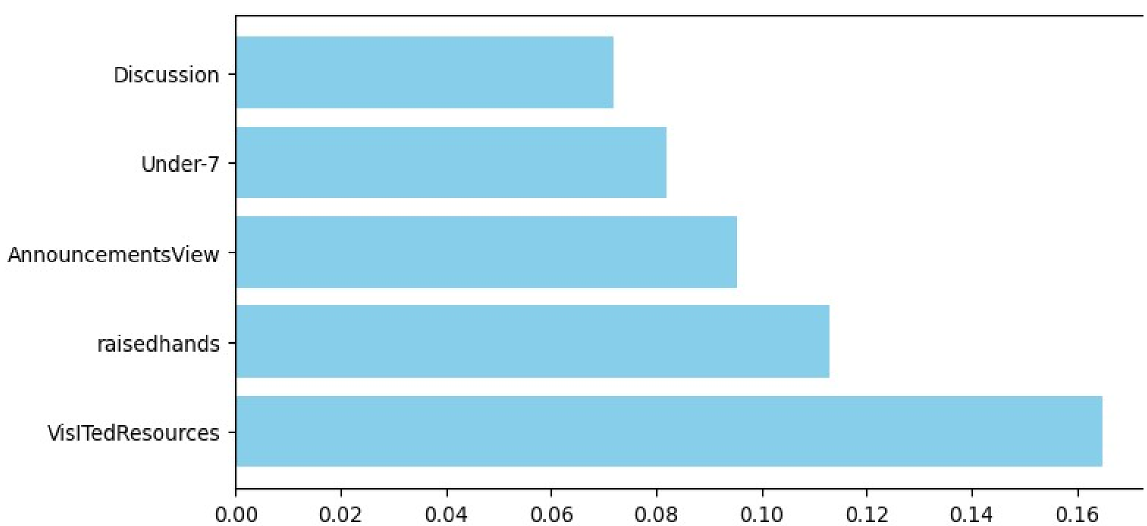


Figure 4: Feature Importance plot of the Random Forest Model

The feature importance plot from the Random Forest model in Figure 4 gives the relative significance of the first five most important features in predicting students' academic performance. From Figure 4, it can be observed that the most important feature for predicting students' academic performance is the number of times a student visited a course content, with an importance score of approximately 0.16, indicating it has the highest impact on the model's predictions. This is followed by the " times the student raises his/her hand in the classroom" feature with an importance score of approximately 0.12. Other significant features include "Announcements View," which represents the number of times a student checks new announcements, "Under-7" which is a category of the variable “student absent days”, and "Discussion," with descending importance scores. These scores suggest that the number of resources visited, and the number of hands raised are crucial factors for the model while discussions and announcements also play notable roles, albeit to a lesser extent. It is worth noting that all the quantitative predictors form part of the most important features of students’ academic performance.

4 Research Gaps

There is currently no consensus on which ML model is most effective in predicting students' academic performance, even though ML models have the potential to provide valuable insights into students’ performance. This lack of clarity can be attributed in part to the fact that student academic performance is influenced by several factors, such as socioeconomic status, cultural background, teaching styles, and individual learning differences. To train ML models, large and diverse datasets are required, and obtaining these datasets can be challenging. Despite these challenges, many researchers continue to investigate ML potential in education. Models have been developed that predict academic performance based on features like student behavior, engagement, demographics and the like but there is no consensus on the most effective features for prediction.

Ouatik et al. (2022) indicated that SVM algorithm is the most efficient ML model to predict students’ performance, with the highest classification rate (87.32%) and the shortest execution time among other ML algorithms such as KNN and C4.5. However, according to Hasan et al. (2020), RF compared to CN2 Rule Inducer algorithm was more accurate in predicting student’s academic performance with a prediction accuracy of 88.3%. Ahmed et al. (2021) noted that DT was a better predictor of students’ academic performance with a prediction accuracy of 89.1% compared to LR (82.6%), NB (86.2%) and SVM (88.8%). Vijayalakshmi and Venkatachalapathy (2019) indicated that DNN algorithm with 84% accuracy rate outperformed other ML algorithms such as DT-C5.0 (69%), Na¨ıve Bayes (73%), RF 79%, SVM (69%) and KNN (75%) when a student performance prediction system was modelled. In Alshdaifat et al. (2022), Mutual Information algorithm in conjunction with five different ML models that employ both classification and regression classifiers showed that SVM had the highest prediction accuracy of 81.67%, followed by RF, KNN, LR, and Gaussian Naive Bayes with respective accuracy scores of 78.33%, 75.00%, 74.17%, and 50.83%. In Aman et al. (2019), LR achieved the highest prediction accuracy of 83.8% when used as an ML algorithm for students’ academic performance relative to RF and J48 with respective accuracy scores of 83.3% and 79.4%.

A significant number of literature focuses on different features as predictors of students’ academic performance in ML. Sixhaxa et al. (2022) predicted students’ academic performance with ML approach using 16 features (4 quantitative and 12 qualitative) including demographic (gender, country of origin and place of birth), academic (grade level, educational stages, semester, etc.) and behavioral (days absent from school, group discussion, raised hands, viewing announcements, etc.) from students dataset extracted from Kaggle. However, Ahmed et al. (2021) predicted students’ performance using 20 features of which 6 of them were numerical (age, site number, family size, number of repetition years, mid school degree and rate of internet usage) and 14 nominal features including department affiliated to, sex, home address, father’s job, mother’s job, college affiliation etc. On the contrary, in Ya˘gcı (2022), 1854 students who took Turkish Language-I exam in the 2019–2020 fall semester were selected to be included in the dataset used for assessing students’ performance. The features used for prediction consisted of midterm and final exam grades, Faculty and Department of the students studied. Cruz-Jesus et al. (2020) made use of features such as age, number of years enrolled in high school, access to internet, scholarship, class size, economic level, number of courses offered to predict students’ academic performance in a ML setting. In another research by Hoffait and Schyns (2017), with the usage of a ML algorithm to predict students of low risk of passing, features such as gender, country of origin, prior schooling, math scholarships were used. Features such as location of school, school type, school size, gender, socio-economic status, parental pressure, parent educational status were employed by Rebai et al. (2020) to identify the key features that impact student academic performance using a ML algorithm. Nevertheless, Bernacki et al. (2020) made use of log records in the management system to predict undergraduate student achievement.

It is evident from previous studies and this present study that different ML models employed give different prediction accuracies of student’s academic performance. Also, the different ML models employed use different features in the prediction of students' academic performance. Several studies have investigated using ML models to create personalized learning experiences and to provide targeted interventions for struggling students. Even though no clear model has been developed to predict students' performance, the quality of data, interpretability, generalizability and feature selection of the ML models employed in predicting students' academic performance should be of high priority to researchers. The availability of quality data plays a crucial role in the success of the ML algorithm. Therefore, data cleaning and preprocessing techniques are essential to ensure that the input data is accurate and reliable. While ML models are effective in predicting student academic performance, they are often difficult to interpret. Having an in-depth knowledge of how the model arrived at its predictions is critical in developing core objectives. Also, selecting the right set of features that have the most significant impact on student academic performance is critical in developing effective ML models. However, determining which features to include in the model can be challenging. While ML models for predicting academic performance have not yet given us a clear alternative, this does not necessarily indicate that ML algorithms are worthless in predicting student academic performance. Instead, it highlights the need for continued research and development to improve the accuracy and reliability of ML models in predicting students' academic performance with improved learning outcomes while generalizing to different populations and other learning environments.

5 Conclusions and Recommendations

The burgeoning utilization of machine learning (ML) algorithms in predicting students' academic performance marks a pivotal shift in the educational landscape. As delineated in this study's empirical literature review, a diverse array of ML techniques—including decision trees, random forest, support vector machine, logistic regression, Naive Bayes, KNN, and ANN—has been effectively employed to forecast academic success. These algorithms consider a variety of critical factors, such as demographic characteristics, prior academic achievements, study habits, and socio-economic status, to provide a better understanding of student performance. The capability of ML algorithms to not only accurately predict academic outcomes but also identify students who are at risk and in need of additional support is particularly noteworthy. This advancement holds great promise for revolutionizing educational strategies and interventions. However, the journey to fully integrating ML in education is not without its challenges and limitations. A primary concern is the need for high-quality, relevant data, which can be a significant hurdle in certain contexts. The quality of data directly impacts the effectiveness of ML predictions, necessitating a robust data collection and management infrastructure within educational institutions. Another critical issue is the potential bias inherent in ML algorithms. If the data used to train these models is biased, the resulting predictions will be skewed, leading to inaccurate and potentially unfair outcomes. This necessitates a concerted effort from policymakers and educators to establish rigorous standards and guidelines that ensure the fairness and accuracy of ML applications. Regular audits and assessments of these algorithms are essential to identify and rectify any biases. Moreover, the integration of ML into educational systems requires careful consideration. Policymakers must navigate the complex interplay of technology, pedagogy, and student welfare to create an ecosystem where ML tools are used ethically and effectively. This involves not only leveraging technology to enhance learning outcomes but also ensuring that the human aspect of education is not overshadowed.

The practical analysis of student academic performance using the xAPI-Edu-Data from Kaggle has demonstrated that machine learning algorithms can effectively predict academic outcomes. Among the five algorithms evaluated, the Random Forest Classifier emerged as the best model, achieving the highest G-Mean and accuracy of 0.9243 and 85.42% respectively. This model's performance emphasizes the importance of balanced sensitivity and specificity in predicting academic performance. The feature importance analysis highlighted that students' engagement with course content, as measured by the number of resources visited, is a crucial determinant of their academic success. These findings suggest that educational institutions can leverage such predictive models to identify and support at-risk students, thereby enhancing overall academic achievement. The research identifies significant gaps in existing literature, encompassing conceptual, methodological, analytical, and ethical aspects. These gaps can profoundly affect the validity, reliability, and overall impact of research in this field. Addressing these issues is crucial for ensuring that the research not only contributes to academic discourse but also translates into practical strategies that can significantly improve student learning outcomes. Additionally, numerous factors influencing a student's academic performance have been recognized, ranging from individual aspects like motivation and self-regulation to environmental elements like socio-economic status and school resources. Understanding and addressing these multifaceted factors is key to improving educational outcomes. Educators and policymakers are thus encouraged to adopt a holistic approach that considers the diverse needs and backgrounds of students, fostering an educational environment that promotes academic success for all. While ML algorithms offer transformative potential in predicting and enhancing student academic performance, the path forward involves addressing critical challenges related to data quality, algorithmic bias, and the integration of technology within educational frameworks. By doing so, educators and policymakers can harness the power of ML to not only predict academic outcomes but also to create more equitable and effective educational systems.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data used to support the findings of this research is freely available at the Kaggle and can be assessed through the repository https://www.kaggle.com/search?q=xAPI-Edu+.

Acknowledgement

The first and corresponding author also acknowledges the enormous support of the University of Texas Rio Grande Valley (UTRGV) Presidential Research Fellowship fund.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., and Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *Ieee Access*, 9:7519–7539.

Ahmed, D. M., Abdulazeez, A. M., Zeebaree, D. Q., and Ahmed, F. Y. (2021). Predicting university's students performance based on machine learning techniques. In *2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS)*, pages 276–281. IEEE.

Ahmed, I. A. I., Cheng, W., et al. (2020). The performance of robust methods in logistic regression model. *Open Journal of Statistics*, 10(01):127.

Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., and Ali, N. (2022). Assessment and evaluation of different machine learning algorithms for predicting student performance. *Computational Intelligence and Neuroscience*, 2022.

Alshdaifat, E., Zaid, A., and Alshdaifat, A. (2022). Predicting student performance: a classification model using machine learning algorithms. *International Journal of Business Information Systems*, 39(3):349–364.

Altabrawee, H., Ali, O. A. J., and Ajmi, S. Q. (2019). Predicting students' performance using machine learning techniques. *JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences*, 27(1):194–205.

Alyahyan, E. and Du¨¸steg¨or, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17:1–21.

Alzubi, J., Nayyar, A., and Kumar, A. (2018). Machine learning from theory to algorithms: an overview. In *Journal of physics: conference series*, volume 1142, page 012012. IOP Publishing.

Aman, F., Rauf, A., Ali, R., Iqbal, F., and Khattak, A. M. (2019). A predictive model for predicting students academic performance. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4. IEEE.

Asvio, N. et al. (2022). The influence of learning motivation and learning environment on undergraduate students' learning achievement of management of islamic education, study program of iain batusangkar in 2016.

Alzubi, J., Nayyar, A., & Kumar, A. (2018, November). Machine learning from theory to algorithms: an overview. In *Journal of physics: conference series* (Vol. 1142, p. 012012). IOP Publishing.

Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., Alsariera, Y. A., Ali, A. Q., Hashim, W., and Tiong, S. K. (2022). Toward predicting student's academic performance using artificial neural networks (anns). *Applied Sciences*, 12(3):1289.

Baig, M. A., Shaikh, S. A., Khatri, K. K., Shaikh, M. A., Khan, M. Z., and Rauf, M. A. (2023). Prediction of students performance level using integrated approach of ml algorithms. *International Journal of Emerging Technologies in Learning*, 18(1).

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., and Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1):905–971.

Beckham, N. R., Akeh, L. J., Mitaart, G. N. P., and Moniaga, J. V. (2023). Determining factors that affect student performance using various machine learning methods. *Procedia Computer Science*, 216:597–603.

Bernacki, M. L., Chavez, M. M., and Uesbeck, P. M. (2020). Predicting achievement and providing support before stem majors begin to fail. *Computers & Education*, 158:103999.

Borman, G. D. and Dowling, M. (2010). Schools and inequality: A multilevel analysis of coleman's equality of educational opportunity data. *Teachers College Record*, 112(5):1201–1246.

Chan, R. Y. (2016). Understanding the purpose of higher education: An analysis of the economic and social benefits for completing a college degree. *Journal of Education Policy, Planning and Administration*, 6(5):1–40.

Chui, K. T., Liu, R. W., Zhao, M., and De Pablos, P. O. (2020). Predicting students' performance with school and family tutoring using generative adversarial network-based deep support vector machine. *IEEE Access*, 8:86745–86752.

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., and Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a european union country. *Heliyon*, 6(6):e04081.

de la Fuente, J., Malpica-Chavarria, E. A., Garz´on-Umerenkova, A., and Pach´on-Basallo, M. (2021). Effect of personal and contextual factors of regulation on academic achievement during adolescence: the role of gender and age. *International Journal of Environmental Research and Public Health*, 18(17):8944.

El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., Dakkak, A., and El Allioui, Y. (2020). A multiple linear regression-based approach to predict student performance. In *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019) Volume 1-Advanced Intelligent Systems for Education and Intelligent Learning System*, pages 9–23. Springer.

Epstein, J. L. and Sheldon, S. B. (2002). Present and accounted for: Improving student attendance through family and community involvement. *The Journal of Educational Research*, 95(5):308–318.

Feeney, D. M., Morgan, J. J., Rodgers, W. J., Brown, M. R., and Relles, S. R. (2023). Self-talk monitoring and goal-oriented thinking: Effects on middle schoolers' behavior during academics. *Learning Disability Quarterly*, page 07319487221145189.

Francis, B. K. and Babu, S. S. (2019). Predicting academic performance of students using a hybrid data mining approach. *Journal of medical systems*, 43:1–15.

Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., and Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences*, 10(11):3894.

Hoffait, A.-S. and Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, 101:1–11.

Ibrahim, Z. and Rusli, D. (2007). Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. In *21st Annual SAS Malaysia Forum, 5th September*. Kuala Lumpur.

Kotsiantis, S., Pierrakeas, C., and Pintelas, P. (2004). Predicting students'performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5):411–426.

Kour, S., Kumar, R., and Gupta, M. (2021). Analysis of student performance using machine learning algorithms. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1395–1403. IEEE.

Kusurkar, R., Ten Cate, T. J., Vos, C., Westers, P., and Croiset, G. (2013). How motivation affects academic performance: a structural equation modelling analysis. *Advances in health sciences education*, 18:57–69.

Liu, C., Wang, H., Du, Y., and Yuan, Z. (2022). A predictive model for student achievement using spiking neural networks based on educational data. *Applied Sciences*, 12(8):3841.

Morrison, F. J., Ponitz, C. C., and McClelland, M. M. (2010). Self-regulation and academic achievement in the transition to school.

Ning, H. K. and Downing, K. (2012). Influence of student learning experience on academic performance: The mediator and moderator effects of self-regulation and motivation. *British Educational Research Journal*, 38(2):219–237.

Nyarko, E., Agyemang, E. F., & Arku, D. (2023). COVID-19 and online teaching in higher education: A discrete choice experiment. *Model Assisted Statistics and Applications*, *18*(2), 115-123.

Obsie, E. Y. and Adem, S. A. (2018). Prediction of student academic performance using neural network, linear regression and support vector regression: A case study. *International Journal of Computer Applications*, 180(40):39–47.

Onyancha, F. K., Njoroge, D. K., Newton, D. M., et al. (2015). The influence of parents' socio-economic status on students' academic performance in public secondary schools in keumbu division, kisii county, kenya.

Ouatik, F., Erritali, M., Ouatik, F., and Jourhmane, M. (2022). Predicting student success using big data and machine learning algorithms. *International Journal of Emerging Technologies in Learning (Online)*, 17(12):236.

Pinkus, L. (2008). Using early-warning data to improve graduation rates: Closing cracks in the education system. *Washington, DC: Alliance for Excellent Education*.

Rastrollo-Guerrero, J. L., G´omez-Pulido, J. A., and Dur´an-Dom´ınguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied sciences*, 10(3):1042.

Rebai, S., Yahia, F. B., and Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in tunisia. *Socio-Economic Planning Sciences*, 70:100724.

Ryan, R. M. and Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67.

Shailaja, K., Seetharamulu, B., and Jabbar, M. (2018). Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pages 910–914. IEEE.

Sharma, N. and Juneja, A. (2017). Combining of random forest estimates using lsboost for stock market index prediction. In *2017 2nd International conference for convergence in technology (I2CT)*, pages 1199–1202. IEEE.

Singh, A., Thakur, N., and Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee.

Sixhaxa, K., Jadhav, A., and Ajoodha, R. (2022). Predicting students performance in exams using machine learning techniques. In *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 635–640. IEEE.

Vijayalakshmi, V. and Venkatachalapathy, K. (2019). Comparison of predicting student's performance using machine learning algorithms. *International Journal of Intelligent Systems and Applications*, 11(12):34.

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., and Nawaz, R. (2020). Predicting academic performance of students from vle big data using deep learning models. *Computers in Human behavior*, 104:106189.

Yaacob, W. F. W., Nasir, S. A. M., Yaacob, W. F. W., and Sobri, N. M. (2019). Supervised data mining approach for predicting student performance. *Indones. J. Electr. Eng. Comput. Sci*, 16(3):1584–1592.

Yaˇgcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1):11.

Yu, R., Li, Q., Fischer, C., Doroudi, S., and Xu, D. (2020). Towards accurate and fair prediction of college success: Evaluating different sources of student data. *International educational data mining society*.

Zhai, X. (2021). Practices and theories: How can machine learning assist in innovative assessment practices in science education. *Journal of Science Education and Technology*, 30(2):139–149.

Zhang, W., Wang, Y., and Wang, S. (2022). Predicting academic performance using tree-based machine learning models: A case study of bachelor students in an engineering department in china. *Education and Information Technologies*, pages 1–16.