Discover Computing

Research

Data driven decisions in education using a comprehensive machine learning framework for student performance prediction

Muhammad Nadeem Gul[1] · Waseem Abbasi[2] · Muhammad Zeeshan Babar[3] · Abeer Aljohani[4] · Muhammad Arif[1]

Received: 4 January 2025 / Accepted: 29 April 2025
Published online: 18 July 2025

Abstract
Accurately predicting student performance is essential for improving educational outcomes and guiding targeted interventions. This study applies eight advanced machine learning models-Decision Trees, Random Forest, Lasso, K-Nearest Neighbors, XGBoost, CatBoost, AdaBoost, and Gradient Boosting to analyze student performance based on demographic and academic features. Among these, CatBoost achieved the highest accuracy (87.46%) and less error rates, outperforming Gradient Boosting (87.28%) and Decision Trees (82.42%). Model evaluation was conducted using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), demonstrating the robustness of the proposed approach. The results highlight the effectiveness of data-driven methods in early identification of at-risk students, enabling educators to implement personalized learning strategies. This study underscores the transformative potential of machine learning in education, paving the way for more adaptive and student-centered learning environments.

Keywords Machine learning · Performance prediction · Common regression metrics (MAE, MSE and RMSE) · Lasso · K-Neighbours · Decision Trees · Random Forest · XG-Boost · Cat-Boosting · Ada-Boost · Gradient Boosting

1 Introduction

The integration of modern technologies in education plays a crucial role in driving positive change and development. In today's data-driven world, predicting student performance and utilizing educational data to gain valuable insights have become essential for educators. As the volume of student data continues to grow exponentially, educational institutions and administrators are increasingly adopting data-driven decision-making to enhance learning outcomes and institutional effectiveness.

The study uses a rich dataset with several types of information that affect student results. Student performance depends on details about ethnicity and gender combined with parental education levels and how students act during classes. The connection between students' academic results and their testing achievements alongside their learning habits makes an ideal starting ground for building precise forecasting methods and analysis.

Determined by the recognition of the importance of proper education, most states have tried and are still trying to strengthen their educational systems. It is fascinating and very useful in identifying and forecasting students' performance and it supports evidence based decisions. However, most of the existing models have a rich set that includes some irrelevant

✉ Waseem Abbasi, waseemabbasi97@gmail.com; ✉ Muhammad Zeeshan Babar, m.babar@hw.ac.uk | [1]Department of Computer Science & IT, Superior University, Sargodha Campus, Sargodha 40100, Pakistan. [2]Department of Computer Science, The University of Lahore, Sargodha Campus, Sargodha 40100, Pakistan. [3]School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4 AS, UK. [4]Department of Computer Science, Applied College, Taibah University, Medina 42353, Saudi Arabia.

attributes, which makes them very complex and costly. To cope up with this challenge, schools are implementing latest approaches including the deep learning technique to identify the flagged students with learning disabilities.

A similar study did use a Gated Recurrent Unit (GRU) model [1] to observe a student's past academic records to identify the students that needed extra support. Dense layers and max-pooling layers were used, and the ADAM optimization method was applied on the dataset of 15,165 student records. The models compared included RNN and AdaBoost, but the GRU model achieved the high accuracy. This approach enables early intervention of the at risk students and enhancing their academics performance and reducing in their drop out rate.

Academic performance prediction stands as a major role in Educational Data Mining (EDM) and requires machine learning (ML) methods to work with educational data sets. The study [2] used supervised ML methods to forecast student results based on BISE Peshawar, Khyber Pakhtunkhwa data. This research project studied seven geographic areas to test if good education helps nations achieve sustainable development targets. Using 30 student features researchers processed the dataset and tested regression and decision tree algorithms to forecast student results. Exploring Machine Learning shows its ability to forecast student performance through education and helps educational leaders make better choices.

Research work [3] develops a conceptual framework which combines machine learning with (Higher Education Institutions) HEIs' management systems to improve student achievement results. The research implements both strategic planning and analytical reasoning to evaluate and compare Linear Regression with Decision Tree and Random Forest and Multilayer Perceptron (MLP) as machine learning algorithms. The models evaluate three essential performance indicators which consist of student satisfaction levels alongside academic achievement measures with institutional productivity metrics. The Multilayer Perceptron model provided the best performance by reaching an MSE of 0.018 and an MAE of 0.105 and establishing an R^2 score of 0.842. The effectiveness of this framework within higher education institutions has been validated empirically which demonstrates its potential use as a powerful tool for organizational transformation and diversity enhancement and quality development.

The evaluation of e-learning outcomes for engineering students based on predictive learning methods uses both decision trees and random forests algorithms. The models generate forecasts about student attrition rates and their final GPA metrics known as cumulative grade point average (CGPA) [4, 5]. A detailed dataset containing student information and academic records along with classroom involvement spreads and evaluation findings enables the identification of critical performance factors for online learners. The transparent decision paths of the Decision Tree model base their decisions on meaningful feature attributes while the Random Forest model achieves high accuracy through ensemble decision trees which resolve class imbalance problems. The research indicates these models enhance both learner interest and effective e-learning development strategies for educational institutions.

One of the areas of education research is focused on student progress and the ability to meet the changing demands of the learners. Given that the adoption of the digital formats of academic data increases, there is increasing interest around applying machine learning (ML) to study not only academic, but also non-academic factors that could affect student performance. The use of ML in identifying key characteristics and making predictions regarding student outcomes have been looked at in 84 studies which were subjected to a systematic literature review [6]. The review emphasizes most academic and demographic factors are the best predictors of student success.

Among ML methods commonly used in educational research, classification techniques, specifically decision tree models, are ones of widespread use. Despite the review covering the published literature, gaps in the movement science literature exist and fundamental performance analysis and intervention strategies are not present. This study further studies the shortcoming of those indicators by pointing out the need of benchmarked datasets to develop frameworks of early intervention for predicting students' performance.

To forecast student outcomes more accurately, many educational institutions now employ advanced ML algorithms as illustrated in Fig. 1. These predictive models offer substantial benefits to school administrators, educators, parents, and most importantly students to shape their academic trajectories and future aspirations. In line with this, the present study proposes a machine learning-based framework that performs a comprehensive analysis of student performance using advanced predictive techniques. This research underscores the importance of leveraging student data and carefully selecting relevant features for performance prediction and analysis. For this study, we utilized a publicly available dataset containing student records [7].

(i) Problem statement:

Market remains saturated with various academic institutions including online learning institutions, it has become critical for educational institutions to identify correlations between costs and benefits and most importantly student performance predetermination. Despite the various enhancement in the recent years, ML tends to perform well in predicting the amount of tutorials the students will require in order to succeed, there are a few setbacks: Some of the existing

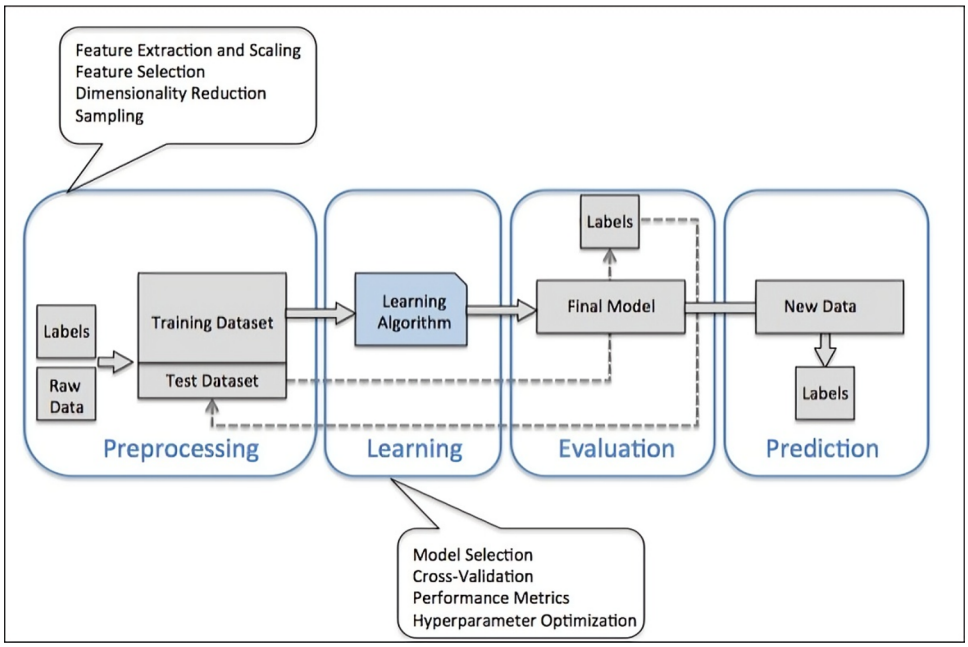

Fig. 1 Proposed model of students for making predictions

research studies do not include enough features or they do not combine all of the overall features of demographic and academic. The prediction of the performance is valuable for educational stakeholders because it allows distinguishing the students who might experience difficulties in achieving good outcomes and who require intervention to check on their studies. This enhances student retention and reduces the likelihood of dropouts.

Education competition and online learning growth make precise student performance forecasting necessary for academic institutions to succeed. Research teams apply machine learning algorithms to student performance prediction but many current studies have restricted parameters and unclear student background statistics. The exact prediction of student results helps educational leaders find students who need help early on. Institution can take prompt action and customize student support when ML detects performance risks which helps students stay enrolled and perform better.

Beyond basic machine learning models, this study employs eight different algorithms-including Decision Tree, Random Forest, and Multi-Layer Perceptron-to predict student performance. The accuracy of these models is evaluated using performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), ensuring a reliable assessment of predictive capabilities. Furthermore, the study investigates the role of machine learning in forecasting student outcomes and highlights the importance of demographic features in enhancing model accuracy.

The proposed model is provided with several advantages. It gives valuable performance predictions to students who can practice what they can do before it as a step to bring their performance to par. It is also a very useful tool for the parents to take part in their children's education and monitor the progress of their children. However, this model can also be used by policymakers and educational institutions as a systematic approach for evaluating the new as well as the current students. The model offers contributions to the existing literature on the use of data driven learning interventions and educational management by being integrated with machine learning into the academic assessment design.

This study identifies and addresses key gaps in the existing literature, particularly the limited exploration of how demographic and academic factors can be combined to predict student success. Furthermore, it proposes a more accurate and comprehensive approach to forecasting student outcomes, offering valuable insights into the integration of machine learning techniques in educational settings.

(ii) Research objectives:

(a) To develop a predictive model for student performance using machine learning techniques:

The primary goal of research study is to design and implement a machine learning framework capable of predicting student performance. This model will utilize a combination of demographic factors (e.g., gender, race/ethnicity, parental education level, lunch type, and test preparation status) and academic indicators (e.g., math, reading, and writing scores) to enhance predictive accuracy.

(b) To evaluate the effectiveness of eight machine learning algorithms for student performance prediction:

This work aims to assess the efficiency of eight different machine learning algorithms, including Decision Trees, Random Forest, and Multi-Layer Perceptron, in predicting student performance. The evaluation will be conducted using standard error metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to ensure a robust performance comparison.

(c) To identify the significance of demographic and academic features in predicting student success:

This objective focuses on examining the impact of various demographic and academic factors on student outcomes. The goal is to identify key risk factors that contribute significantly to variations in student success or failure, providing deeper insights into their influence on academic performance.

(d) To provide actionable insights for educational stakeholders (students, parents, and policymakers):

This study aims to develop a predictive model that serves multiple stakeholders. Students can use it to monitor their academic progress, parents can track their children's performance, and policymakers can leverage it to make informed decisions on student support, retention, and enrollment strategies.

(e) To bridge the gap in current educational research by integrating both demographic and academic data:

The main objective emphasizes the simultaneous use of demographic and academic data to improve the accuracy and significance of student performance prediction. Additionally, it seeks to bridge existing research gaps by applying machine learning techniques to this integrated dataset, offering a more comprehensive approach to student outcome analysis.

(f) To implement error analysis for evaluating model performance:

Basic aims is to compute and compare error metrics-Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) for each machine learning model. This evaluation will provide a comprehensive understanding of the strengths and weaknesses of different models in predicting student performance.

(g) To contribute to the development of data-driven decision-making tools in educational institutions:

To develop a model that serves as an innovative tool for educational institutions to assess current, past, and newly enrolled students. By leveraging data-driven insights, the model will assist in making informed decisions regarding student interventions and academic support, ultimately enhancing educational outcomes.

(iii) Contributions:

This study makes a significant contribution to the field of educational data mining by systematically comparing a diverse range of regression models for student performance prediction. It incorporates a rich set of features, including both demographic and academic variables, ensuring a more holistic analysis than previous works.

Key contributions include:

Comprehensive Model Evaluation: This study rigorously evaluates multiple machine learning models, normalizing errors across techniques and applying standardized feature extraction, data preprocessing, and error analysis methods.

Methodological Advancement: The research provides a structured, step-by-step guide and a conceptual framework for scholars interested in further exploring machine learning applications in education.

Bridging Research Gaps: By integrating both demographic and academic factors, this study extends beyond prior research, offering a more complete understanding of the key predictors influencing student success.

These contributions collectively enhance predictive accuracy, facilitate data-driven decision-making, and provide a valuable reference for future research in educational data mining.

This research utilizes eight machine learning algorithms-Lasso, K-Nearest Neighbors, Decision Trees, Random Forest, XGBoost, CatBoost, AdaBoost, and Gradient Boosting-to predict student performance with high accuracy. Among these, all models exhibited strong predictive capabilities; however, the CatBoost algorithm achieved the highest performance, making it the optimal choice for this study.

This study provides a valuable foundation for educational institutions by enabling data-driven decision-making for both current and former students. Parents can use the findings to monitor their children's academic progress, while students can benefit from the insights for secondary school planning and career guidance.

High-Accuracy Prediction: The study achieves 87.46% accuracy in predicting student performance using the CatBoost regression model.

Data-Driven Decision-Making: The model aids school administrators in assessing student outcomes, while parents can use it as a tool to track their children's academic progress.

Student Guidance: The insights provided can help students make informed decisions regarding secondary education and career planning. Given these contributions, this research establishes a solid foundation for future work in educational

data mining, positioning CatBoost Regression as a highly effective approach for predicting student performance across various educational contexts.

(iv) Organization:

The remainder of this article is organized as follows: Section 2 reviews existing literature on educational data used for predicting and assessing student performance in real-time. Section 3 outlines the dataset and the proposed methodology for predicting student performance. Section 4 discusses the results of the simulations, while Section 5 concludes the study and suggests directions for future research.

2 Related work

Predicting student performance has emerged as a vital research focus in educational data mining due to its potential to enhance learning outcomes and support personalized education. Despite the growing use of predictive models in educational institutions, challenges remain in achieving consistent and reliable results. These difficulties often arise from the limited implementation of advanced analytical methods like machine learning and deep learning, which are crucial for handling large and complex datasets related to student performance.

The performance of students gets significantly influenced by various external conditions and environmental aspects. A universal predictive model faces barriers from student distractions and negative learning conditions and insufficient motivation along with external disturbances during learning. Students who succeed in standardized tests sometimes encounter challenges which impact their performance ratings in other subjects. The use of exam results alone fails to create accurate forecasts about students' academic achievement. Certain predictions about student achievement require numerous educational variables combined with student characteristics to guarantee accurate evaluation results.

To overcome these challenges it requires that the academic institutions and other stakeholders to consider integrating the continuous form of assessment in the education system to empowering stakeholders most especially parents. The formative assessments help the teachers to detect areas of academic difficulties at the beginning of learning process so they are able to give special additional instructions to the student to address the problem. Parents, guardian, teachers, and counsellors will be better placed to understand children's learning patterns by using complex and wider models that factor in academic, behavioral and other parameters. It also leads to better decision making and encapsulates a data center approach of improving any student outcomes and total academic performance.

In recent years, there has been research [8]focus on predicting student performance in online learning environments by exploring several interaction features, like engagement to learning material, participation in discussion forum, quiz results and collaborative opportunities. Develop one model using the data of the Open University which achieved an accuracy of 75 percent using neural network model that highlights that although quantitative metrics of student interaction are useful, they could be used in conjunction with qualitative insights. This finding points to the need for changing the presentation of content to account for the distinct needs of learners and to offer a more focused instruction for students at risk in order to enhance learning outcomes.

Learning analytics and artificial intelligence (AI) prediction models [9] have been used to solve problems of feedback, interaction, and collaborative learning in online engineering courses. By integrating AI and learning analytics, we found not only that performance and satisfaction of the students has been improved, but also a reasonable amount of valuable information in terms of the synergy between AI and learning analytics and how this synergy can be used in online education.

Multiple research works have studied the application of educational data mining approaches for assessing final examination results. Research conducted with 1,854 Turkish university students utilized Random Forest alongside Support Vector Machines (SVM) and Logistic Regression and k-Nearest Neighbors (kNN) for their examination data analysis [10]. The success rate of the applied models reached 70–75% which demonstrates that accurate early detection of at-risk students requires reliable information sources. The analysis demonstrates the power of predictive analysis to trigger early prevention efforts which leads to better academic results.

Many scholars apply regression models to forecast student marks while decision tree analysis supports the classification of grade levels for performance evaluations. The independent research team analyzed student achievement using predictive models based on B.I.S.E Peshawar data before proposing system-wide solutions for educational management systems in their report [2, 11]. This research shows that educational quality improvement potential of data-driven insights can be utilized across Pakistani educational institutions. Scientific research analyzed how behavioral characteristics within e-learning systems affect academic results. Research analysts upgraded their models' effectiveness using Naïve Bayes

and Decision Tree alongside SVM and KNN algorithms and Bagging and Boosting techniques to show how behavioral characteristics influence student achievement levels [12].

Research reveals that ML play essential role in current learning systems as 33 approaches appeared across 55 studies and 38 countries [13]. Ongoing research will continue studying to anticipate educational results and recognize student weaknesses in order to cut down on class withdrawals. A study evaluated TEL (Technology enhanced learning) system performance by judging student achievements from task times and engagement metrics as recorded by keystroke counts [14]. The analysis showed that artificial neural networks and support vector machines provided the most precise predictive models when cross-validation methods were used because of their decision tree effectiveness evaluation. When TEL systems include these models they track student performance in real-time so that educators can help students earlier.

Research conducted between 2010 and 2020 on student performance prediction assessed evaluation models in combination with forecasting methods and learning achievement determinants [15]. The number of related papers still remains lower than studies encompassing big data perspectives although research continues in this area with present emphasis on 62 papers. A deep artificial neural network (ANN) model predictively analyzed virtual learning environment (VLE) clickstream data from students to determine at-risk students with classification accuracy rates between 84% and 93% [16]. The obtained results demonstrated that students who begin instruction with foundation-level knowledge perform better in education tests thus requiring institutions to develop strategic educational interventions for improved learning outcomes. An evaluation of student performance involved treating a 649-instance and 33-attribute dataset with Naïve Bayes and Decision Tree and Random Forest algorithms [17]. The study proved that data mining techniques successfully reveal academic success determinants which strengthens the position of machine learning in educational analytical applications.

Research work [18] which examines the influence of COVID-19 pandemic in Nigeria with regards to changes made to teaching pedagogy, educational calendar and international education. These disruptions will thus require the government to provide more support in order to ensure continuity of education. A related study [19] on CS students also aims at evaluating the potential of utilizing machine learning algorithms for predicting the students' performance. The study under discussion emphasizes the role of feature selection in the increase of predictive performance discussing superiority of decision tree based techniques over the others. This signifies the importance of choosing informative features to increase the reliability of the models used in education forecasting.

In response to the challenges posed by COVID-19, universities in Thailand transitioned to online learning, prompting concerns about monitoring student behavior. A study [20] analyzed Moodle log data, which consisted of 54,803 events, to evaluate the accuracy of six machine learning classifiers-Neural Networks, Random Forest, Decision Tree, and Support Vector Machine (SVM)-in detecting student performance at various stages of a course. Another study [21] explored student performance prediction using machine learning models such as SVM, Decision Tree, Random Forest, and K-Nearest Neighbors, incorporating both online and offline datasets. The study also examined how factors like sleep patterns, study duration, and leisure activities relate to academic performance. The results emphasized the importance of data normalization to improve prediction accuracy and enhance learning outcomes in e-learning environments.

Recently, a research work [22] conducted the use of ChatGPT in education mainly using Web mining, natural language processing and machine learning techniques to mine 2003 articles. Results showed that ChatGPT enhances the quality of writing, enhances student engagement and supports a personalized learning experience. The study, however, also brought up concerns regarding the possible risks, such as cheating and plagiarism. Another study [23] brought Signed Graph Neural Networks (SGNNs) along with Large Language Model (LLM) embeddings into play to deal with learner data sourcing and noise challenges. Specifically, this approach increased predictive accuracy and robustness to cold start scenarios with Peer-Wise datasets. Moreover, a research study [10] used machine learning models like Random Forest, SVM, Logistic Regression and Naïve Bayes to foretell the final examination outcomes using a data set of 1854 Turkish university students from midterm results. Educational Data Mining can be used for early identification of at-risk students, the models presented obtain an accuracy in the range 70–75%.

A student performance prediction system was assessed using seven machine learning algorithms, including Deep Neural Networks (DNN) [24]. The results revealed that the DNN model achieved an accuracy of 84%, highlighting the effectiveness of deep learning in improving predictive performance in educational data mining. To address challenges posed by imbalanced datasets, a study [25] explored several resampling techniques, including Borderline SMOTE, Random Over Sampler, SMOTE-ENN, SVM-SMOTE, and SMOTE-Tomek. The study, which tested six machine learning classifiers on two datasets, demonstrated that balancing class instances notably improved classifier performance. Among the methods evaluated, SVM-SMOTE combined with the Random Forest algorithm yielded the best results. Additionally, the use of Learning Coefficients [26] was explored to enhance student outcome predictions through trajectory-based

adaptive assessments. By analyzing attribute relationships with Pearson's correlation coefficient, the study identified Linear Regression as the most accurate model, achieving an accuracy of 97%, thus validating its effectiveness in predicting student performance.

In Hong Kong, multiple regression models and machine learning algorithms were compared and their predictive capabilities were evaluated to a cross-sectional survey of 425 undergraduate students [27]. The Elastic Net Regression model explained 65.2% of student satisfaction proving itself to be a good option for modeling academic experiences. Furthermore, a Deep neural network (DNN) model [28] was proposed to determined student performance on higher education. The model trained on labelled data had MAE score of 0.593 and RMSE score of 0.785 indicating that the model has strong predictive power. The model's potential to support decision making of educators and policy makers also comes out of these results.

The application of machine learning techniques in Turkey [29] uses various factors including academic reputation and university facilities to predict student placements. XGBoost proved to be the most accurate model among all the tested variants. The analysis showed the previous placement statistics and faculty-to-student ratios as essential factors for university resource planning through decision support tools. In addition, a performance prediction model is developed for the students at IBN ZOHR University in Morocco [30]. Random Forest Recursive Feature Elimination with Cross-Validation (RFECV-RF) was used as an optimal feature selection technique in this study. When we trained the Support Vector Machine (SVM) model on eight features selected via RFECV-RF, an accuracy of 87% is obtained, which shows the great influence of feature selection on improving prediction performance.

Similarly a research [31] also investigated predicting students' engagement and performance in the above mentioned online environments using machine learning methods. It was found that the Random Forest classifier performed well in predicting grades (85%) and prediction engagement (83%), indicating that machine learning holds promise to increase the effectiveness of the online education. A survey of over 70 researches on predictive models of student performance prediction using machine learning, collaborative filtering, recommender systems, and artificial neural networks is undertaken by Azar [32]. From being able to better predict student performance and plan academic progress, to improve the educational strategy used for instruction, these AI driven methods have proven to be vital components for an instituiton.

Research [33] used artificial intelligence methods to discover whether textbooks create gender prejudices among educators and students. The research team studied 470 participant responses through machine learning approaches that included ANN, RF, and SVM. The Neural Network Application reached 87.2% accuracy which exceeded the other models including Random Forest at 84% and Support Vector Machine at 80%. This research demonstrates the power of AI to find and resolve education systems' social prejudice. Machine learning helps develop customized education methods [34] by replacing basic teaching models with adaptive systems according to research. The systems update learning materials and adjust educational paths as each student progresses. Research finds ways to improve machine learning models so they help students experience education better in physics, mathematics, and language topics.

Researchers studied how separate science curriculum delivery affected student scientific literacy using TIMSS 2019 data from 44 countries and publishing their results in [35]. An integrated curriculum format led Grade 8 students to achieve better science outcomes when compared to students who had separate curricular programs. The study demonstrated that an integrated curriculum could not establish itself as one-dimensional factor determining academic outcomes. The Random Forest algorithm used within machine learning achieved superior modeling outcomes over traditional statistical methods when studying science achievement results. The results demonstrate how machine learning technology can create substantial value for educational research studies and assessment of teaching programs. During the COVID-19 pandemic researchers constructed a machine learning regression model according to [36] for predicting student outcomes from their study activity. The tool lets students monitor their educational advancement while stimulating students to interact with online classes leading to better academic achievements.

A team presented University of Baluchistan (UOBEDM dataset) Pakistan which included information for 23,492 students and implemented computer learning algorithm testing in their exploration. Rephrase the authors determined tree-based models reached 96% accuracy through their creation of the Early Intervention Model (EIM). The UOBEDM model demonstrates how statistical approaches to decision-making help increase education efficiency by establishing fair and accessible educational establishments. Another research [38] predicted college student dropouts through the evaluation of demographic, socioeconomic, academic, and financial variables. The most accurate model among machine learning tests proved to be Gradient Boosting which successfully identified 94.4% of students at risk. The research results show machine learning tools hold promise for developing specific support measures which automatically sustain student population and promote fair educational opportunities.

Machine learning algorithms e.g. K-Nearest Neighbors (KNN) and Multiple Regression can be used [39] to predict and visualize students' cognitive and psychomotor skills in the context of Learning Management Systems (LMS). In particular, the Multiple Regression model demonstrated an accuracy level of 99%, thus indicating the efficacy of data driven strategies in evaluating student performance. Besides this, machine learning has also significantly contributed to the shift of the traditional education into personalized learning systems [34]. Machine learning algorithms analyze students' learning histories, interests, abilities to create personalized learning materials and provide personalized feedback to students. It improves learning outcomes from early childhood education to adult learner education in traditional educational settings as well as settings in second language learning, adult education, and personalized education.

The student performance prediction has become a critical research area within Educational Data Mining (EDM), particularly as institutions worldwide increasingly rely on data-driven strategies to enhance educational quality. Various machine learning (ML) techniques have been employed to forecast student performance using demographic and academic data. Prior studies have demonstrated that predictive models can effectively identify students at risk of underperformance by analyzing historical academic records.

Supervised ML approaches, such as regression and decision trees, have been widely used to estimate student grades based on demographic and academic predictors. These studies emphasize the importance of feature selection and data preprocessing in improving predictive accuracy. Additionally, research has explored the application of decision trees, random forests, and Multilayer Perceptron (MLP) models to evaluate key performance indicators, including student satisfaction and institutional efficiency. This underscores the versatility of tree-based approaches in higher education management.

However, as shown in Table 1, many existing models remain constrained by the scope of their feature sets and the quality of data used, limiting their predictive potential.

This study builds upon previous research by incorporating a comprehensive set of demographic and academic features, thereby providing a more holistic approach to predicting educational outcomes. Eight advanced machine learning algorithms-Decision Trees, Random Forest, Lasso, K-Neighbors, XGBoost, CatBoost, AdaBoost, and Gradient Boosting (GB)-are utilized to forecast student performance. The performance of these models is evaluated using multiple error metrics, including mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE).

Among these models, the CatBoost regression model achieved the highest accuracy at 87.46%, outperforming Decision Trees (82.42%) and Gradient Boosting (87.28%). These findings underscore the significance of integrating both demographic and academic data to enhance prediction accuracy and enable early identification of at-risk students. The proposed model has substantial implications for various educational stakeholders-including students, parents, and policymakers-by serving as a valuable tool for data-driven decision-making and targeted intervention strategies to improve student outcomes.

3 Methodology

This section provides a comprehensive explanation of the methodology used in this study, detailing the processes involved in data collection, preprocessing, model selection, evaluation, and result interpretation. Additionally, Fig. 1 illustrates the conceptual model designed for predicting the performance of upper secondary students. The experimental design and analytical conclusions drawn from the dataset serve as the foundation of this research.

3.1 Dataset collection and required packages

The dataset used in our research, which focuses on student performance and predictions, is widely accessible and holds significant importance. It was sourced from Kaggle, providing a comprehensive foundation for our analysis [7]. To develop and evaluate the proposed model, the necessary Python libraries were imported into Google Colab, along with the dataset. This step was critical for data exploration, preprocessing, and model development. The dataset consists of 8 columns and 1000 rows, comprising 5 categorical variables and 3 numerical variables.

These are:

Gender: Denotes the gender of the student (male or female).

Race: Categorizes students into five distinct groups: A, B, C, D, or E.

Parent Education Level: Specifies the highest level of education attained by the student's parents, including categories such as school level, high school level, college level, associate level, bachelor's level and master's degree level.

Table 1 Overview of related work

Author & Year	Study focus	Machine learning techniques used	Key findings	Accuracy/Results	Dataset
Baniata LH, Kang S. 2024 [1]	Predicting student performance in e-learning	Neural Networks, Random Forest, Decision Tree, SVM	Analyzed Moodle log data to predict student performance at different stages of course completion	Decision Tree: 81.10%, SVM: 86.90% at 25% completion	Moodle log data
Hussain S, Khan MQ. 2023 [2]	Predicting student success using study habits	SVM, Decision Tree, Random Forest, K-Neighbors	Correlated student behaviors (sleep, study time) with class performance	Improved prediction of e-learning outcomes	Student behavior surveys, Academic performance data
Khan MI, Khan ZA, Imran A. 2023 [6]	Predicting final exam scores based on midterm results	Random Forest, SVM, Logistic Regression, Naïve Bayes	Predicts final exam performance using midterm grades and student data	70–75% classification accuracy	Academic performance records
Vijayalakshmi V, Venkatachalapathy K.2019 [7]	Deep learning for student performance prediction	DNN, Decision Tree (C5.0), Naïve Bayes, Random Forest, SVM, K-NN	Deep learning models provided better predictions for student performance	DNN achieved 84% accuracy	Academic performance datasets
Nanavaty S, Khuteta A. 2024 [8]	Handling imbalanced data in performance prediction	SMOTE, SVM-SMOTE, Random Forest, KNN, ANN, XGBoost, Logistic Regression	Resampling techniques improve classifier accuracy in imbalanced data	SVM-SMOTE with Random Forest achieved the best performance	Academic datasets with imbalanced classes
Ouyang F, Wu M, Zheng L. 2023 [9]	Improving prediction using learning coefficients	Decision Trees, Random Forest, SVM, Linear Regression, ANN	Learning coefficients used for more accurate performance prediction	Linear Regression: 97% accuracy	Academic performance data
Yağcı M. 2022 [10]	Predicting student satisfaction in educational environments	Elastic Net Regression, Multiple Regression, Machine Learning	Predicts student satisfaction based on learning conditions	Elastic Net Regression explained 65.2% of satisfaction variance	Student satisfaction surveys
Arashpour M, Golafshani EM. 2023 [11]	Using deep learning for student performance prediction	Deep Neural Network	Used deep learning for predicting student performance in educational contexts	0.593 MAE, 0.785 RMSE	Student performance data
Hussain S, Khan MQ. 2023 [2]	Predicting student placement in higher education	XGBoost	Predicts student placement based on academic and institutional factors	XGBoost outperformed other models with high accuracy	Higher education placement data
Ajibade SS, Dayupay J. 2022 [12]	Predicting student performance with feature selection	Random Forest, SVM, Logistic Regression, Naïve Bayes, k-NN	SVM with Recursive Feature Elimination achieved highest accuracy	SVM with RFECV-RF: 87% accuracy	Student performance data
Forero-Corba W, Bennasar FN. 2024 [13]	Predicting student engagement and grades	Random Forest	Predicts student engagement and final grades based on platform interaction data	85% accuracy for grades, 83% for engagement	E-learning platform data
Namoun A, Alshanqiti A. 2020 [15]	Analyzing gender bias in education using AI	ANN, Random Forest, Decision Trees, SVM	AI models used to analyze perceptions of gender bias in textbooks	ANN: 87.2% accuracy	Educational content data

Table 1 (continued)

Author & Year	Study focus	Machine learning techniques used	Key findings	Accuracy/Results	Dataset
Waheed H, Hassan SU. 2020 [16]	Personalizing education with machine learning	ANN, Random Forest, SVM	Uses AI to adapt learning materials based on individual student needs	Improved student engagement and performance	Personalized learning datasets
Salal YK, Abdullaev SM, Kumar M. 2019 [17]	Curriculum impact on science literacy prediction	Random Forest, Educational Data Mining	Machine learning outperforms traditional methods in science literacy prediction	Machine learning improved science achievement modeling	Educational data on science literacy
Kaensar C, Wongnin W. 2023 [20]	Predicting dropout risk in higher education	Logistic Regression, Random Forest, Decision Tree, SVM, Gradient Boosting	Predicts dropout risk based on various student and academic factors	Gradient Boosting: 94.4% accuracy	Student academic data
Holicza B, Kiss A. 2023 [21]	Predicting cognitive and psychomotor skills	KNN, Multiple Regression	Evaluates cognitive and psychomotor skills with machine learning models	Multiple Regression: 99% accuracy	Psychomotor and cognitive skill datasets
Rejeb A, Rejeb K, Appolloni A. 2024 [22]	Customizing education using machine learning	Various ML algorithms	Personalizes education using machine learning algorithms for real-time feedback	Improved educational outcomes through dynamic learning adjustments	Educational systems data
Ni L, Wang S, Zhang Z, Li X.2023 [23]	Predicting success in online education	Random Forest, Logistic Regression, SVM	Focuses on predicting student success in online courses	Random Forest achieved the highest accuracy	Online learning platform data
Adu-Twum HT, Sarfo EA.2024 [24]	Academic success prediction using student behavior data	Random Forest, SVM, Naïve Bayes	Uses student behavior data to predict academic success	Enhanced prediction accuracy in predicting academic performance	Student behavior data
Ghorbani R, Ghousi R.2020 [25]	Developing adaptive learning systems	SVM, ANN, Random Forest	Develops adaptive learning systems that modify content based on student performance	Improved engagement and learning outcomes through AI-based adaptations	Adaptive learning data
Asthana P, Mishra S, Gupta N.2023 [26]	Predicting performance in blended learning environments	SVM, Decision Tree, KNN	Predicts student performance in blended learning settings	High predictive accuracy using machine learning classifiers	Blended learning environment data
Ho IM, Cheong KY, Weldon A.2021 [27]	Big data analytics for performance prediction	Random Forest, SVM, Gradient Boosting	Uses big data to predict student performance and academic success	Gradient Boosting performed best in predicting outcomes	Big data on student performance

Discover Computing (2025) 28:153 | https://doi.org/10.1007/s10791-025-09585-3

Lunch Consumption: Indicates whether students receive a Standard or Reduced lunch before classes or exams.
Exam Preparation: Denotes whether the student has prepared for the exam (marked as "prepared" or "none").
Grades: The final three columns represent the marks obtained by students in mathematics, reading, and writing assessments, each scored out of a maximum of 100.

The dataset is free from duplicate entries which ensures a higher degree of accuracy in the analysis. The data types for each attribute were identified and checks for null values were performed to ensure the integrity and completeness of the dataset. With 1,000 non-missing values, the dataset forms a fully connected network or graph, making it suitable for further analysis without additional data cleaning.

3.2 Visualization and plotting

This section of the paper performs a graphical analysis of the students dataset to observe the mean distribution and end up with some reasonable conclusions. The data visualizations used Histogram, Kernel Density Estimation (KDE) plots and combined Histogram + KDE plots. Such categorizations portray the information as being in order, drawing inferences about groupings of data and providing a dashboard-style overview of the data. Separate male and female datasets were created to allow for gender specific analysis. For instance, different distribution in the data points for each gender, as shown in Fig. 2, manifests in different patterns of student performance metrics.

Gender distribution shown as a histogram, there are total 518 males and 482 females we have total 1000 individuals in the dataset. From this plot we can see that 51.8% of dataset is males and 48.2% is females. At this point of the analysis, it is time to visualize "race" characteristics shown in the dataset and try to gather some insights/conclusions based on the distribution of racial or ethnical groups in the dataset. The data of this representation is illustrated in Fig. 3.

Figure 3 offers a clear representation of the distribution of the "race" attribute within the dataset. The plot reveals that groups C and D are significantly more prevalent than groups A, B, and E, with group A being the least represented. Additionally, Fig. 3 provides a breakdown of the proportion of each racial or ethnic category, offering a visual summary of the overall percentage distribution by category.

Following this, the dataset was further analyzed by examining the distribution of parental education levels. A graphical representation was used to illustrate the various educational categories including School Level, High School Level, College Level, Associate Level, Bachelor Level, and Master Level. This visualization depicted in Fig. 4, provides valuable insights into the varying levels of parental education and their distribution within the dataset, facilitating a deeper understanding of potential influences on student performance.

The comparative analysis plot indicates that the largest proportion of students' parents fall into the "College Level" category followed by the "Associate Level" and then the "High School Level" category. Figure 5 illustrates that students whose parents have attained a "Master Level" or "Bachelor Level" education demonstrate significantly better academic performance compared to those from other educational backgrounds. Following this analysis of parental education levels, students' scores across three subjects- i) Mathematics (Math), ii) Reading, and iii) Writing are compiled and visually

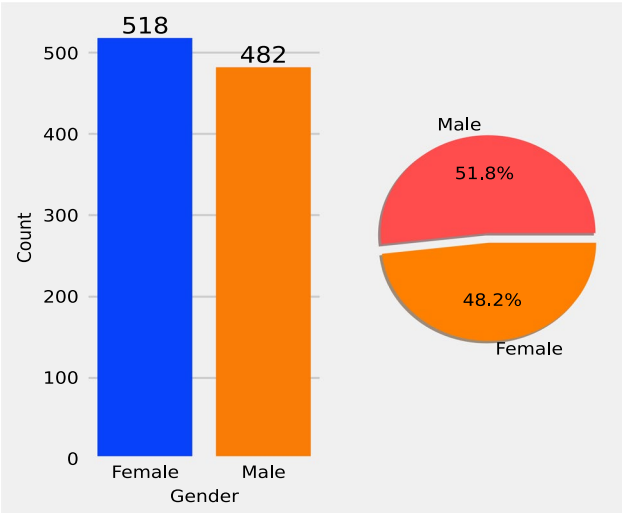

Fig. 2 Gender visualization

represented in Fig. 5. Figure 6 reveals that a substantial proportion of students have achieved Mathematics scores ranging from 60 to 80. In contrast, in Reading and Writing, most students' scores fall within the 50 to 80 range.

3.3 Pie plot and multivariate analysis

Figure 7 employs a pie chart to effectively encapsulate the selected dataset, providing a clear and comprehensive overview of all categorical features. This visualization aids in informed decision-making regarding student attributes and their associated factors, highlighting the intricacies and significance of the dataset. The plot also reveals a linear relationship between scores, indicating that performance increases consistently across all subjects. Notably, student performance closely correlates with factors such as lunch consumption, race/ethnicity and parental education level while gender and test preparation show weaker correlations with academic outcomes. Based on these findings, researchers can make informed decisions. Female students demonstrate higher pass rates and top scores, while the analysis indicates that test preparation courses may have a limited impact on overall student achievement. Therefore graduation levels and course completion may yield greater benefits for students.

3.4 Dataset pre-processing

Our analysis requires removing less helpful features because they affect model efficiency before machine learning classifiers and regression models can work on the information. By choosing the most important features our models become faster to process and produce more accurate results. The model produces better results and handles new data better when it works only with important data features instead of all attributes. This step avoids unnecessary features from training the models which helps them perform at their best.

3.5 Feature selection and bias mitigation

In this study, we selected demographic and academic attributes including race and parental education to analyze their impact on student performance. These features are commonly used in educational research to understand disparities in academic outcomes. However, we acknowledge that such attributes can introduce potential bias in machine learning models.

To assess and mitigate bias, we examined the impact of each feature using feature importance analysis. Additionally, we ensured that predictions were not disproportionately influenced by any single demographic variable. Our evaluation primarily relied on performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE), which provide an unbiased assessment of model accuracy.

While this study focuses on predictive performance, future research could incorporate fairness-aware algorithms to further mitigate bias. Techniques such as feature reweighting, adversarial debiasing, or fairness constraints in model training could enhance the ethical and equitable use of machine learning in educational decision-making.

3.6 Dataset feature engineering

Raw datasets, when directly applied to classifiers and machine learning algorithms, often produce suboptimal results due to inconsistencies, missing values, and unstructured data. Therefore, data preprocessing is a crucial step to enhance the efficiency and accuracy of machine learning models. During the feature engineering phase, the dataset was refined to optimize classification performance. A key transformation involved introducing a new column labeled "Total" which aggregates the Mathematics, Reading, and Writing scores to provide a comprehensive measure of overall student performance.

The updated dataset, incorporating the "Total" column, is displayed in Table 2. Additionally, each attribute assigned an appropriate data type, with categorical attributes represented as strings and numerical attributes as integers. Since not all values in the dataset were whole numbers, necessary transformations were applied to align them with the requirements of the proposed model. One of these transformations involved converting categorical or non-numeric features into a binary format, where "1" signifies "true" and "0" represents "false". This conversion facilitated a structured dataset, ensuring that each feature was effectively represented and ready for analysis using machine learning algorithms.

Table 2 Students attributes sample dataset

Gender	Race	Parental Level of Education	Lunch	Preparations	Computer Science Score	Reading Score	Writing Score	Total
Male	Group B	Bachelor Level	Standard	None	74	73	76	223
Male	Group C	College Level	Standard	Completed	67	88	87	242
Female	Group B	Master Level	Standard	None	91	96	94	281
Male	Group A	Associate Level	Free/Reduced	None	49	59	60	168
Female	Group C	College Level	Standard	None	77	79	76	232

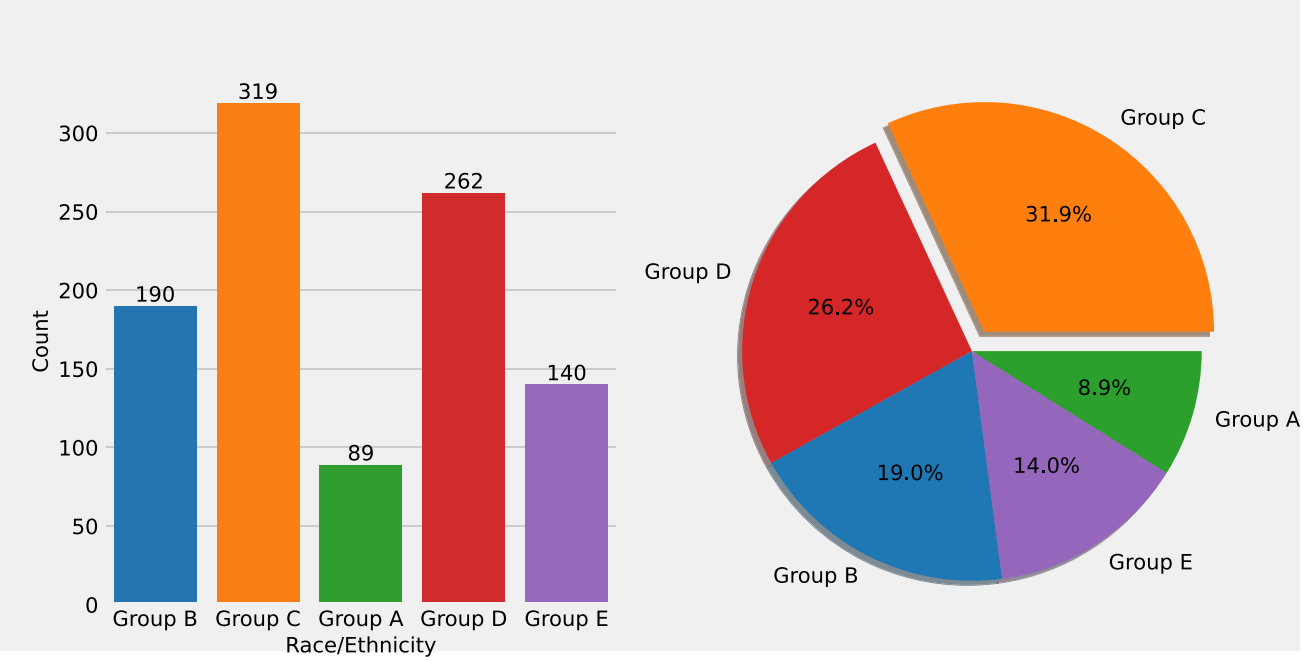

Fig. 3 Race/ethnicity visualization

3.7 Visualization of dataset after feature engineering

The data is transformed and feature engineering is performed to re-visualize the dataset, with the goal of extracting accurate results, facilitating predictions, and assessing student performance. Figure 8 presents a visualization depicting the relationship between gender and the results obtained by the students.

Figure 8 clearly depicts that the female students on average obtain higher and better grades than male students. Here this model denotes gender based disparity in performance in the dataset and the kind of factors affecting performance such as study habits, engagement level or support system.

Figure 9 gives a visual of the relationship between the races and ethnicities, and the students' related grades. The purpose of this graphical representation is to find identifiable patterns or differences in academic performance between the different racial or ethnic groups. The study will analyze these trends to potentially identify disparities as well as the factors that might contribute to success in terms of the various demographic groups.

Figure 9 demonstrates that students who belong to race/ethnicity group C score better than the rest of racial or ethnic groups present in the dataset. The data reveals a distinct achievement pattern that occurs within this ethnic group thus demonstrating potential indicators such as wealth and educational resources availability and cultural academic influences.

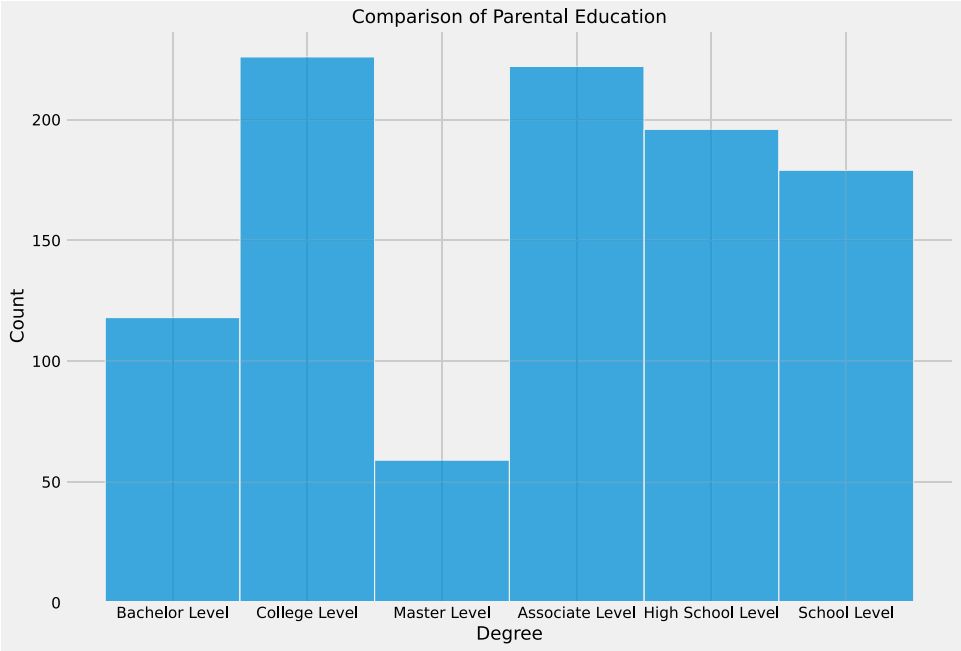

Fig. 4 Comparison of parental education

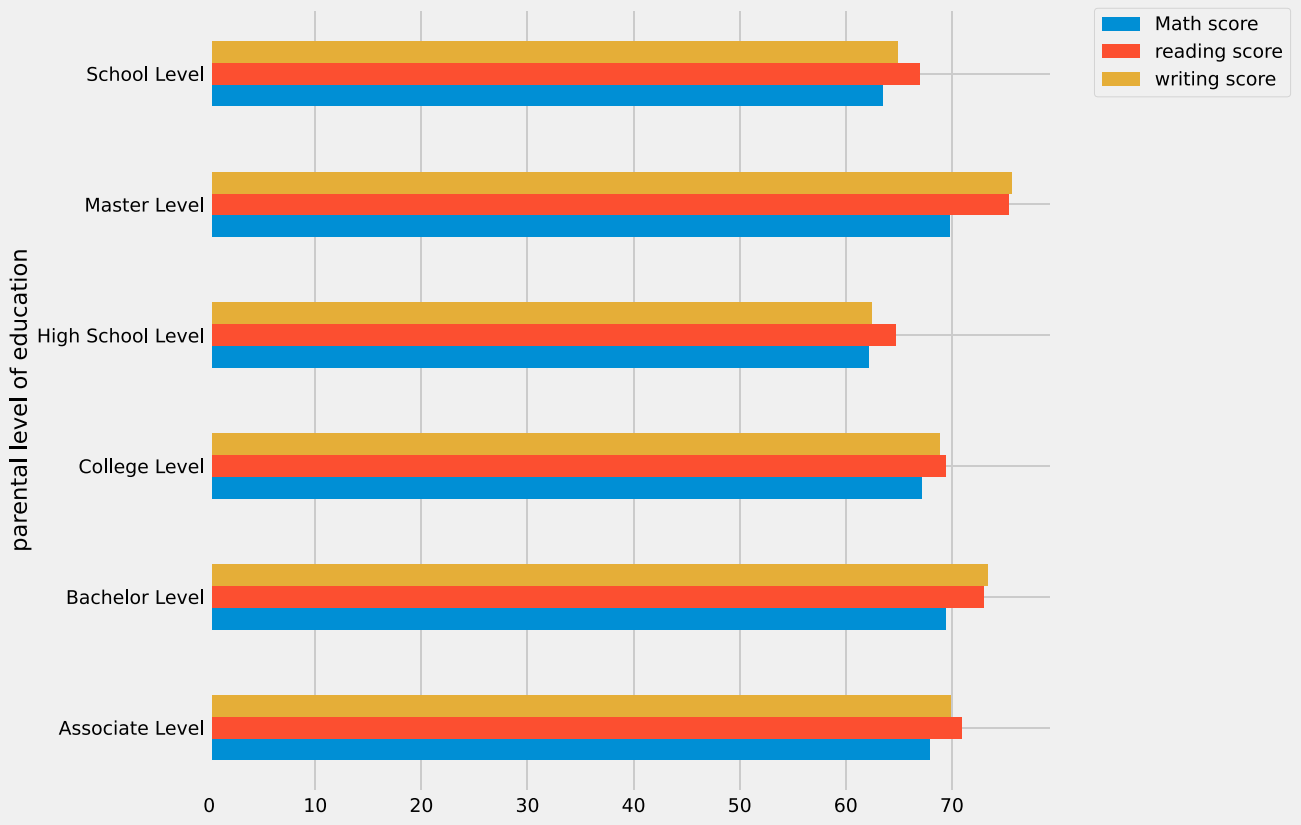

Fig. 5 Parental level of education

A statistical analysis of educational parent levels versus student test scores occurs after completing data transformation and feature engineering steps. The visual representation serves as a key element for revealing patterns or trends that would support understanding how parental education affects student performance. The study investigates academic

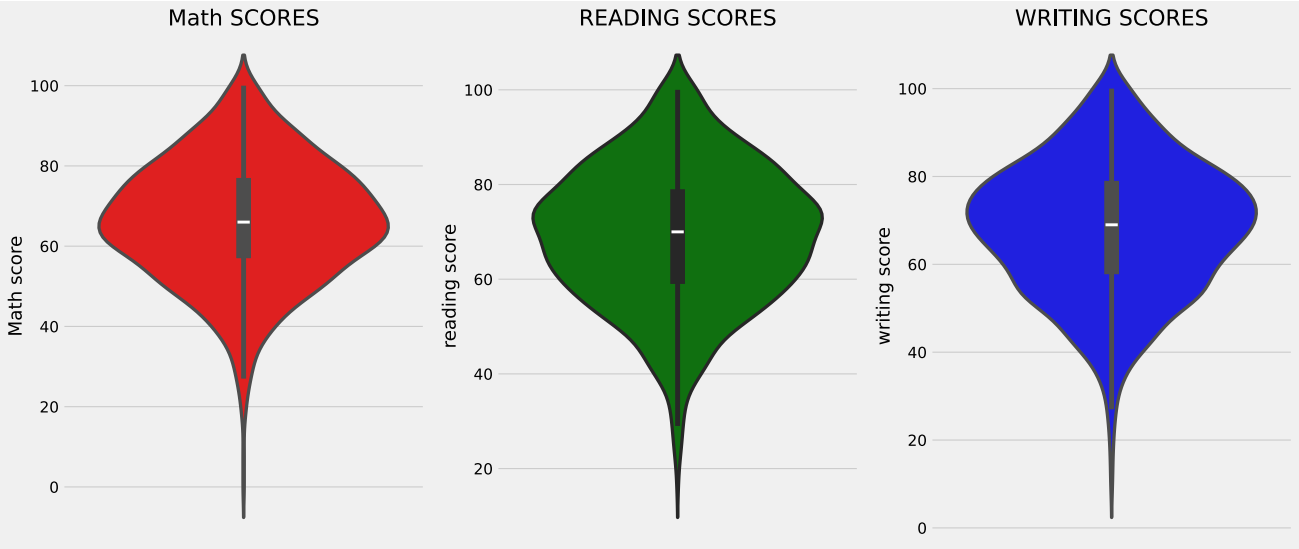

Fig. 6 Students scores in all 3 subjects

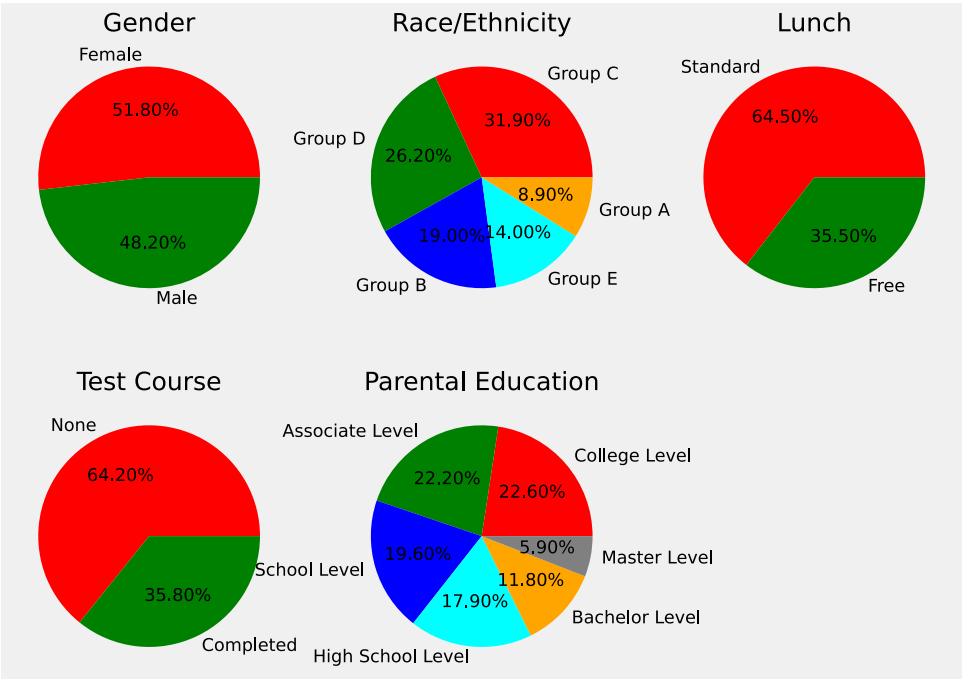

Fig. 7 Multi variant analysis and pie plot

achievement relationships to determine whether better student performance happens with increased parental education levels.

The correlation between students' grades and their parents' level of education is depicted graphically in Fig. 10. While it is difficult to derive definitive conclusions from this plot alone, the data suggests that students whose parents have attained college or graduate-level education, as well as those whose parents have completed high school, tend to perform better on tests. This observation indicates a potential relationship between higher parental educational attainment and improved student achievement. However, further research and advanced modeling techniques would be necessary to establish a more precise and statistically significant correlation.

Additionally, Fig. 11 presents a histogram illustrating the relationship between student performance and lunch preferences. By analyzing the distribution of total scores across different lunch plans, this visualization aims to identify whether

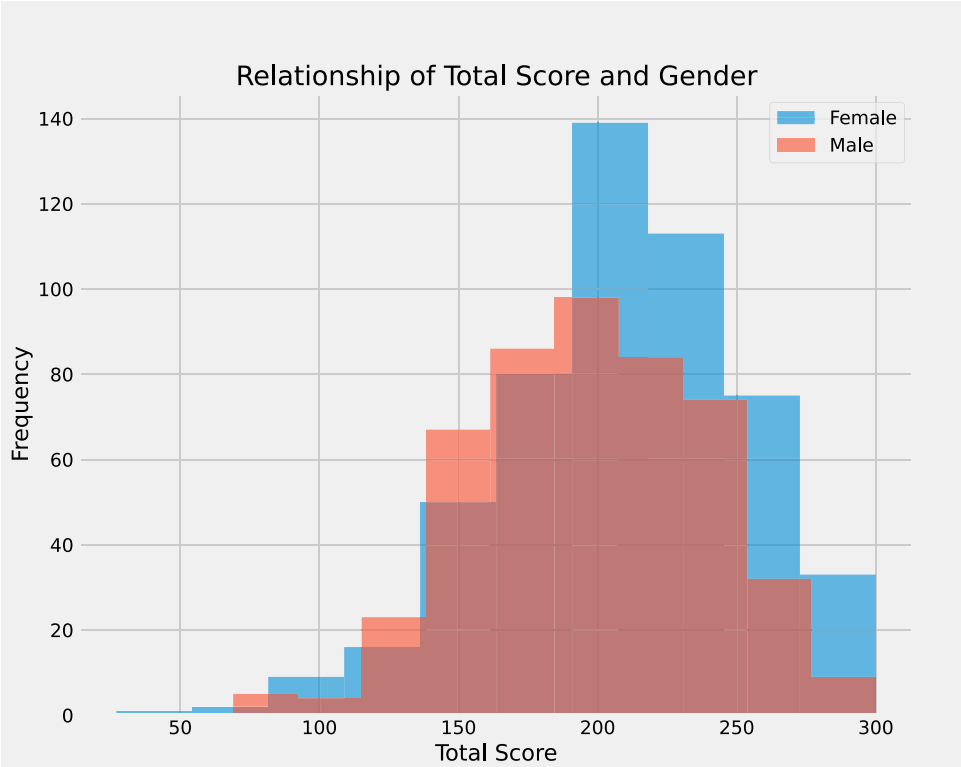

Fig. 8 Relationship of total score and gender

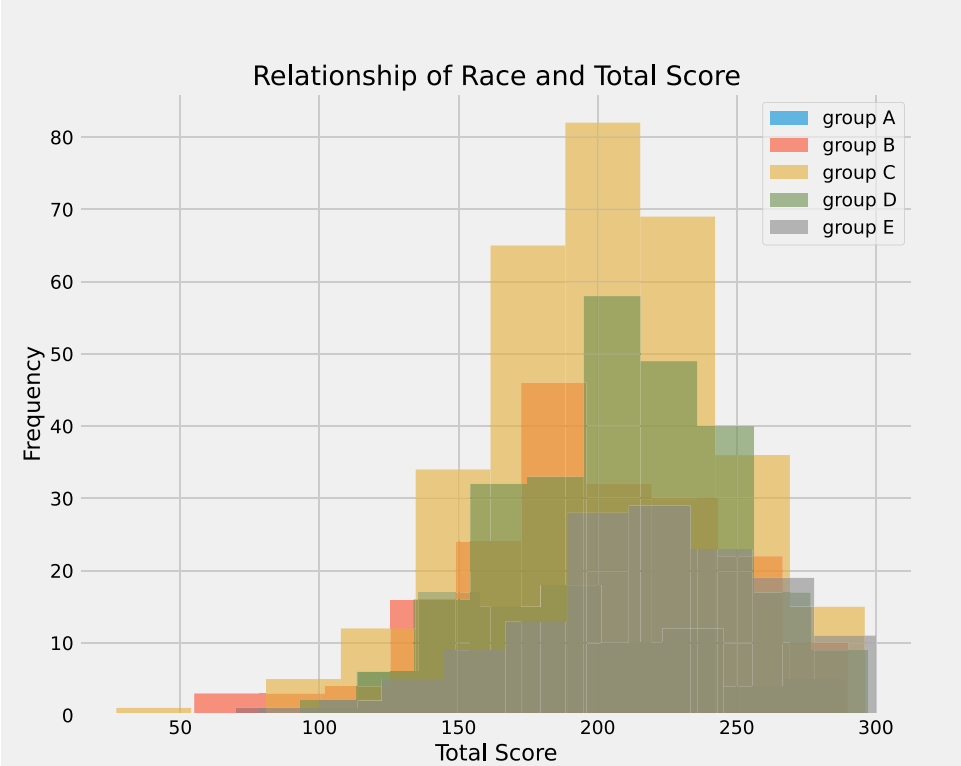

Fig. 9 Relationship of race and total score

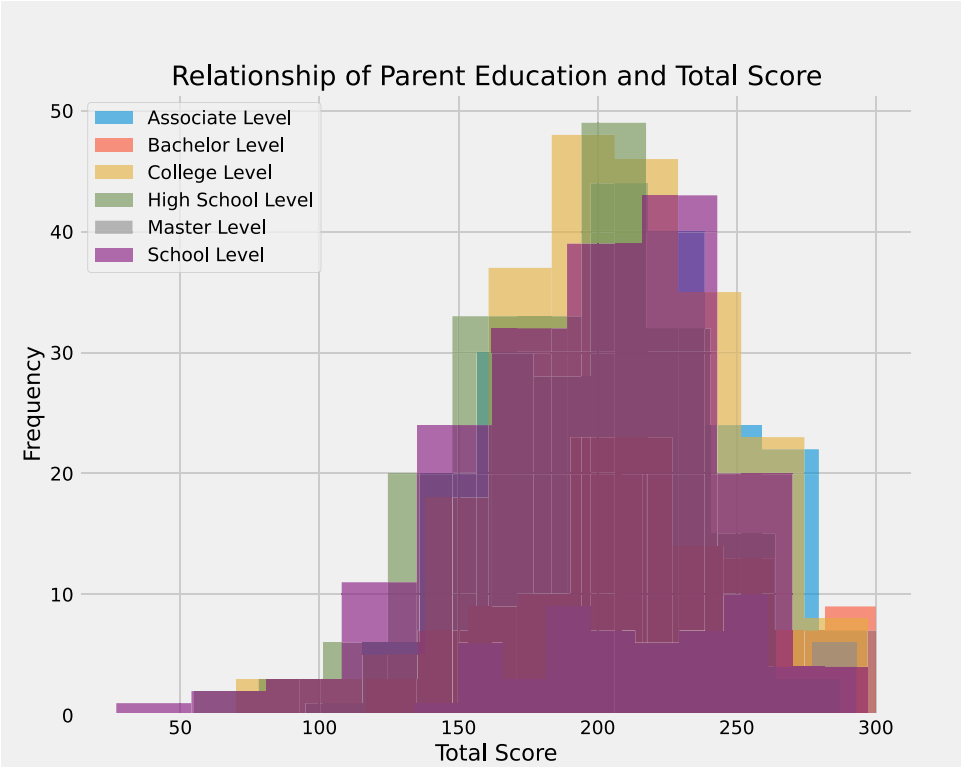

Fig. 10 Relationship of parents and overall score

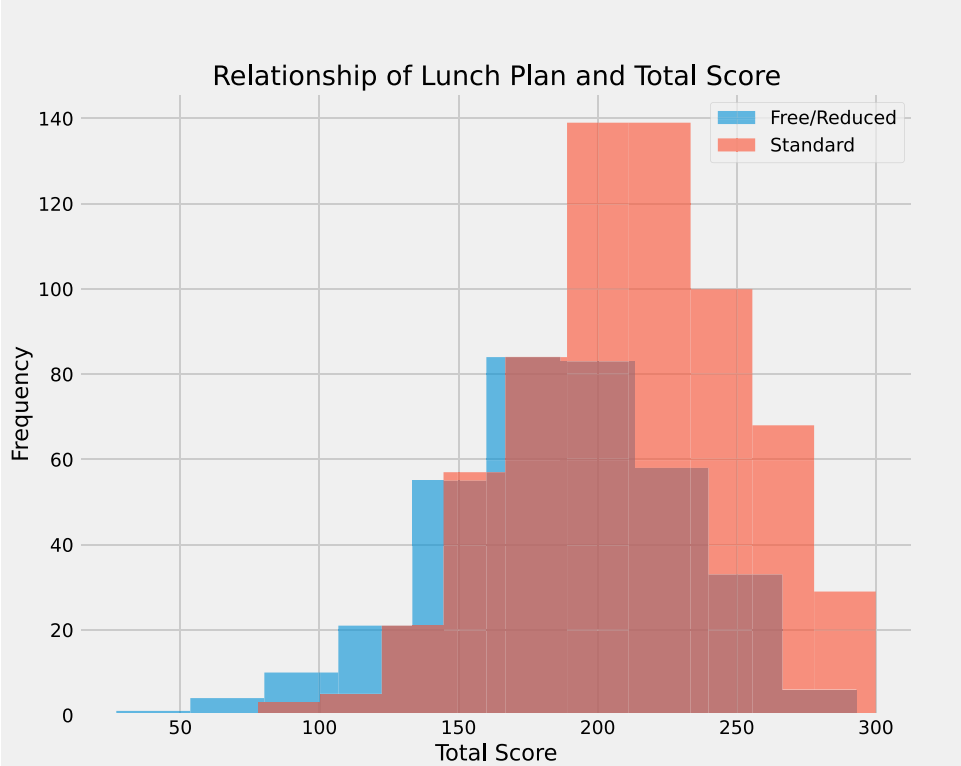

Fig. 11 Lunch plan and total score

there is a discernible effect of lunch choices on overall student performance. If significant patterns emerge, they may provide insights into the role of nutrition and meal quality in academic achievement.

A histograms presented in Fig. 11 clearly indicates that students who eat the standard lunch get higher total scores than those who get the lunch at reduced price or for free. This observation makes the researcher hypothesize that that there is relationship between the kind of lunch and the overall performance of the students with the standard lunch indicating better performance.

In the final stage of visualization, following feature engineering, Fig. 12 illustrates the relationship between students' preparation status (prepared or unprepared) and their corresponding scores. This graphical representation aims to provide insight into how course preparation affects student performance.

The correlation between students' level of preparation in their course and their scores is illustrated in Fig. 12. Interestingly, while only a few students indicated having taken a preparatory course, the majority fall under the "none" category suggesting that a significant portion of students had not engaged in any formal preparation. This observation provides insight into the number of students who did not continue their preparation and hints at the potential effect this may have had on their academic performance.

A notable aspect of this study is that the researcher ensures a comprehensive analysis by considering all five preparation categories holistically. This approach is essential, as each category may encompass factors that uniquely influence student outcomes. By including all categories in the analysis, the study maintains a broad and balanced perspective on the dataset, thereby strengthening the foundation for making conclusions and predictions about student performance. This integrative method proves especially useful when multiple variables contribute to the outcomes of interest in analytical research.

To facilitate regression analysis, mathematics, reading, writing, and total scores were included in the previously created numerical dataset as part of the final step of feature engineering. The primary objective of this integration is to generate a complete dataset encompassing all essential numerical attributes, ensuring the dataset is well-prepared for effective regression modeling and predictive analysis.

To enhance the knowledge about the lunch choice and students performance, the author has presented histogram in Fig. 9. This visualization represents the arithmetical average sum of all the scores made by the respondents and is also composed to illustrate the dispersion of the scores in relation to the type of lunch plans that were proposed to the

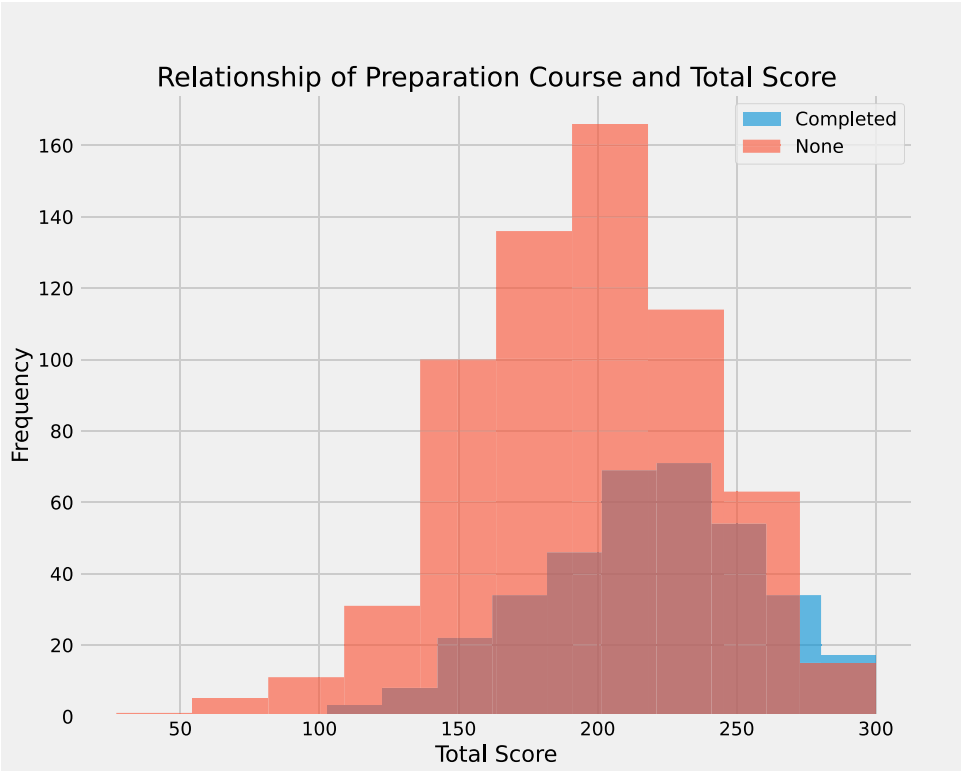

Fig. 12 Test preparation and total score

participants. When interpreting this histogram, one can make judgment whether the time preferred by the students for lunch has an influence on their performance.

The step of cleaning and preprocessing of data is very important prior to the task of training and testing of the machine learning algorithm on the data set. As it was pointed out in the previous section, after feature engineering and selection, the next process is to build and train the final machine learning models. This stage is critical not only in the prediction process but also in making various conclusions from the developed model.

For the purpose of making the evaluation of the models less biased and as objective as possible, the dataset is split into training and testing datasets. This division makes it possible to evaluate the model's ability to generalize to other data points that are yet to be seen. The training set is used in the training of the model and in the optimization of the model while the test set provide an outside measure of how well the model is predicting. This is crucial in machine learning to avoid that instead of the model learning how to predict from a new set of data, it almost predicts from the trained data only. Updating of the model can be done using non-test data hence there is always an opportunity for enhancing the existing models continually. Thus, the model gains better generality rather than overlearning on the example of separate data series of the certain parties. This serves to improve the resilience of the regression, classification, as well as other machine learning techniques to give efficient and accurate results.

4 Results and discussion

To analyze and predict student performance at the secondary and higher education levels, various regression models are employed. These models included CatBoost Regression, Gradient Boosting Regression, Random Forest Regression, XGBoost Regression, AdaBoost Regression, Lasso Regression, Decision Trees Regression and K-Nearest Neighbors Regression. Each of these models was assessed based on its predictive accuracy and error metrics, ensuring a comprehensive evaluation of their effectiveness.

Based on these models, the analysis has been made using mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE). Each of these metrics to an extent measures the ability of each model to predict student performance and the errors made in the predictions. Therefore, the analysis of the results from this evaluation provides a good understanding of various regression techniques strengths and weaknesses in order to determine which one is suitable for educational data analysis. The following sections delve deeper into the outcomes and comparative performance of these models.

To test the possibility of the model, the gathered dataset was split into two: the training set and the test set in the proportion of 80:20. This split is conventional in the methods of machine learning where big data is divided with a view of testing the performance of the model and general ability of the model in handling new data not encompassed by big data. Table 3 below shows the result that proves that Cat-Boost regression model has the highest accuracy as competitively compared with other regression models and got the lowest error rate in the prediction performance.

The general comparison of these models in the analysis brings out the advantages and disadvantages of each method in detail and shows how Cat-Boost is useful in modelling patterns in the dataset as well as obtain high precision in producing performance of students.

4.1 Cat-boosting regression model

The primary objective of the CatBoost Regression model is to estimate student performance based on selected input features such as lunch preference, parental education level, and course readiness. The results indicate that the proposed CatBoost Regression model demonstrates high efficiency, achieving an accuracy of approximately 87.46% on the test dataset, as presented in Table 3. This level of accuracy highlights the model's effectiveness in predicting student outcomes with high precision. One of the key advantages of CatBoost is its ability to handle categorical variables efficiently while minimizing overfitting, making it a suitable choice for educational data analysis. Its strong performance reinforces the importance of integrating demographic and academic features to enhance predictive accuracy in student performance analysis.

Correlation between input features and student performance found by CatBoost Regression model are presented on Table 3. This table is an invaluable resource for users trying to understand how the model is able to systematically draw such conclusions based on certain input variables. This table maps those features to predicted performance outcomes to be more transparent and interpretable so educators and stakeholders can understand the factors that contribute to

student success. Compared with directly optimizing the accuracy of the model, this interpretability gives you deeper understanding of the contribution from different attributes, including parental background, lunch preference, and course readiness, with respect to the prediction of academic performance. This transparency is important to understand the studied findings, for deciding on appropriate intervention options as well as to analyze student success patterns described by the model using real data.

After evaluating the Cat-Boost model, three common regression metrics are used:

- Mean Absolute Error (MAE): An MAE (Mean Absolute Error) closer to zero is preferred, as it indicates better model performance. This is because the MAE measures the average of the absolute differences between the predicted and actual values, with lower values signifying that the model's predictions are more accurate and closer to the true outcomes. A smaller MAE reflects a model's higher precision in approximating the actual data, making it an important metric for assessing the effectiveness of a regression model.
- Mean Squared Error (MSE): In this case, the Mean Squared Error (MSE) decreases, suggesting that a lower MSE value is preferable. This is because MSE gives greater weight to larger errors due to the squaring of differences between predicted and actual values. Consequently, the model is penalized more for significant deviations, which helps to emphasize the importance of reducing large errors in predictions. A lower MSE indicates that the model is not only making accurate predictions but is also minimizing large discrepancies between predicted and actual values, thus improving overall model performance.
- Root Mean Squared Error (RMSE): Root Mean Squared Error (RMSE) for the purpose of this study is used in reverse where the lower the RMSE the better approximated the predicted values are to the actual values. It gives the average amount of variation of the residuals (errors made in the prediction process) using the standard error of estimate formula. This means that the equation's closeness to the actual values is more compact and is stronger, thus the predictions from the model is reliable. In particular, it is useful for estimating the average absolute deviation of the model from the actual values in the same measurement scale as the initial data.

The graphical analysis in Figs. 13 and 14 shows the predicted model values matched closely with actual student performance scores thus verifying its data alignment accuracy. The execution of these figures exhibits how well the model predicts actual outcomes which leads to a substantial confirmation of its predictive capabilities. A calculated correlation coefficient value of 0.876 validates the model's effectiveness and provides strong evidence of the predictive power between forecasted scores and actual scores. Student performance prediction accuracy by the model is demonstrated through its strong relationship with actual results which makes it a beneficial application for academic assessment and intervention planning.

4.2 Gradient Boosting model

Gradient Boosting is a strong technique for learning from a set of decision trees that are striving to have the minimum error, for instance mean squared error or cross-entropy for continuing to learn in stages. In particular, this method should

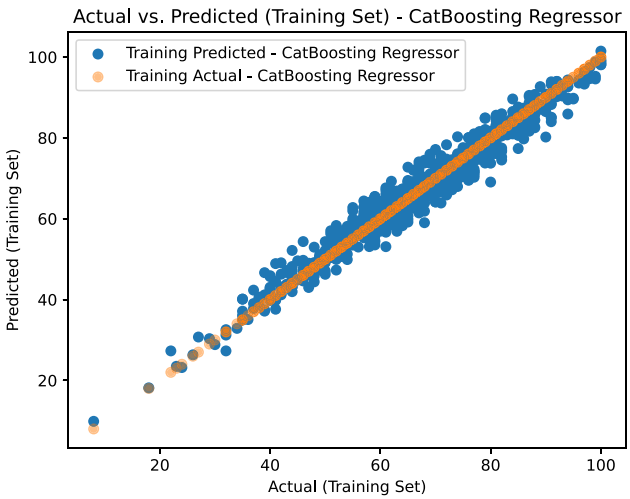

Fig. 13 Training Data Model Predictions by Cat-Boosting

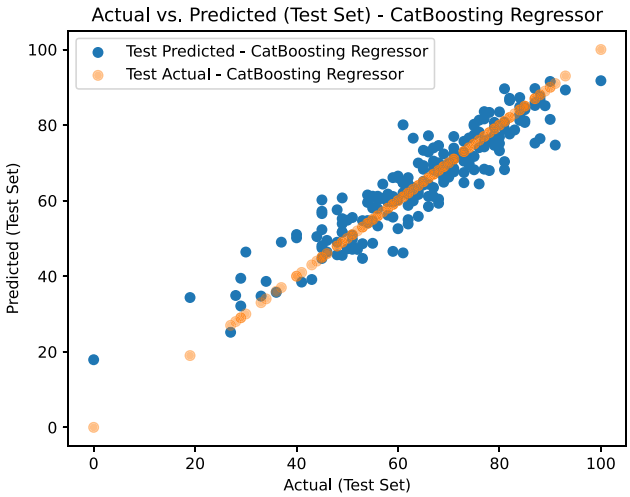

Fig. 14 Testing Data Model Predictions by Cat-Boosting

be used to increase the accuracy of models that require solving complex regression problems. In this study, the Gradient Boosting model depicted high accuracy using cross-validation scores of about 89% and 87.29% (Table 3) using the testing set. These findings are quite encouraging, with the exception that overfitting might be an issue with these models as there was very high accuracy attained. In order to address this, cross validation strategies used and lot of concern was given for the hyper-parameters tuning so that the model is well generalized and can predict the outcome of unseen data accurately.

The regression metrics of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) give additional support to the identification of Gradient Boosting model's effectiveness in determining the intricate data patterns. These measures show the extent to which predicted values meet ground truth, demonstrating the accuracy of the model in question. It can clearly be observed in figures 15 and 16 that there is high level of correlation between the predicted and actual values in the test data set. Such representations make use of the model's ability to identify hidden patterns in the data and therefore validate the relevance of the model in predicting the performance of students.

4.3 Random Forest Regression model

The Random Forest model, an ensemble learning technique that aggregates predictions from multiple decision trees to enhance performance are employed to predict student performance. This model achieved a test accuracy of 86.21% (Table 3), showcasing its robustness in capturing intricate patterns and relationships within the dataset. The inherent ability of Random Forest to mitigate overfitting, by averaging results across numerous trees, further underscores its reliability in predictive tasks.

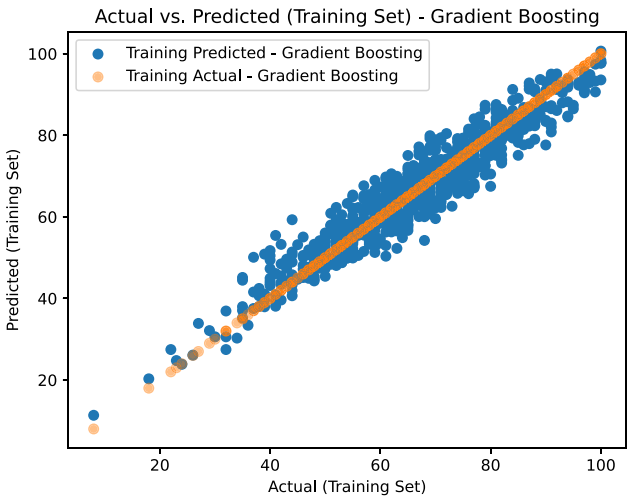

Fig. 15 Training model prediction by gradient boosting

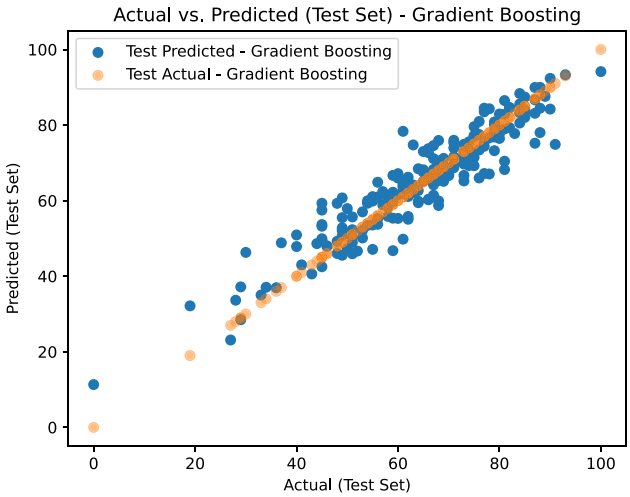

Fig. 16 Testing model prediction by gradient boosting

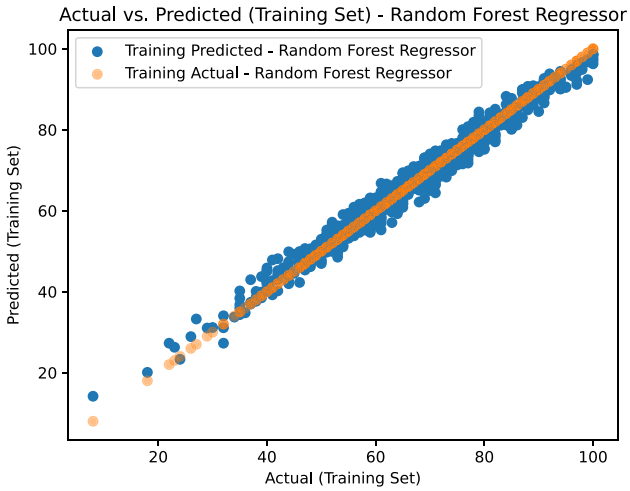

Fig. 17 Training Data Model Prediction by Random Forest

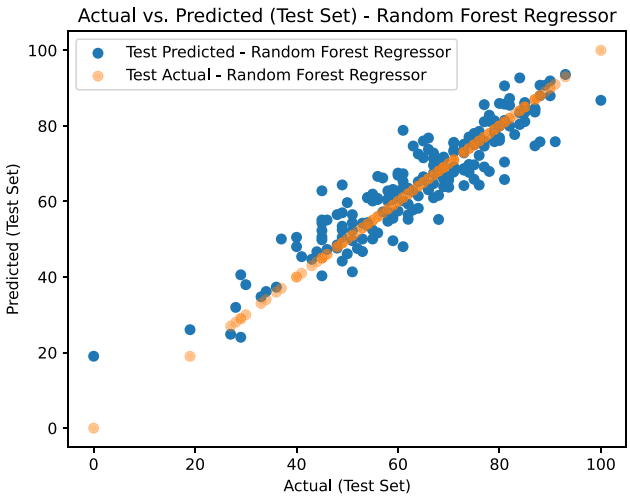

Fig. 18 Testing Data Model Prediction by Random Forest

Evidence supporting the proposed model is clearly illustrated in Figs. 17 and 18. These visualizations demonstrate the close alignment between the predicted and actual values, reinforcing the model's reliability. The near-linear relationship in these plots indicates that the Random Forest model effectively generalizes across the dataset, showcasing its capability to analyze complex educational data. Such visual representations are essential in validating the model's accuracy and its broader applicability in forecasting student performance indicators.

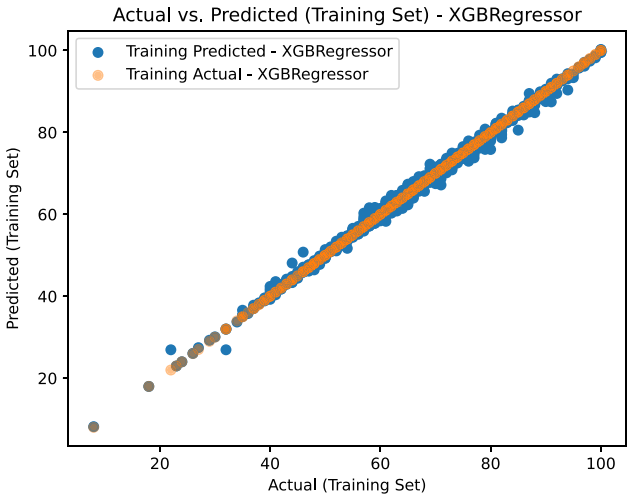

Fig. 19 Training Data Model Prediction by XG-Boost

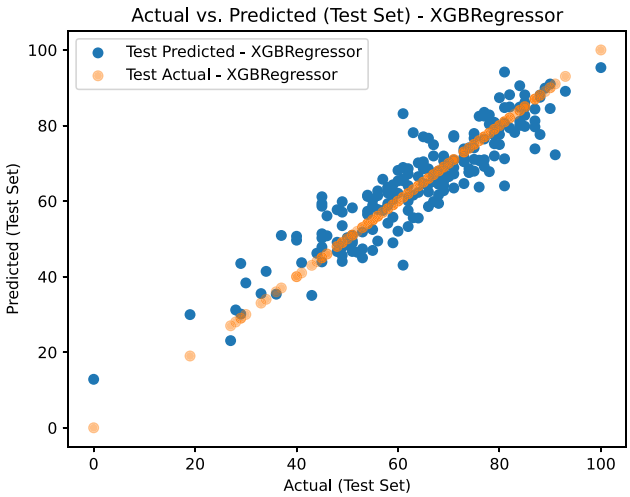

Fig. 20 Testing Data Model Prediction by XG-Boost

4.4 XG-Boost Regression model

Through ensemble learning XG-Boost facilitates robust supervised regression model construction by optimizing a loss function effectively with gradient boosting algorithms. The model reached an outstanding accuracy rate which reached 94.57% both on training data and 86.08% on test data (Table 3). The model demonstrates strong capability to detect intricate associations between variables by using regularization methods to prevent overfitting.

Figures 19 and 20 also support the argument on the viability of the model as a predictor of student performance. These figures indicate the closeness of the actual and predicted values to each other, thus proving the authenticity of the forecast done by the model. The fact that these data points are so proximal suggests that XGBoost is capable of dealing with detailed educational data, which is more evidence in favor of the feasibility of using the technique for student performance prediction. Thus, the correspondingly high degree of matching the actual test results and the forecasting based on the presented model proves its usefulness for educational data analysis to predict academic performance, having high accuracy and coherence.

4.5 Ada-Boost Regression model

Ada-Boost is an ensemble learning algorithm that improves prediction by combining the results of multiple weak classifiers, to form a stronger one. For Ada-Boost, the accuracy on the training data was found to be about 85.25

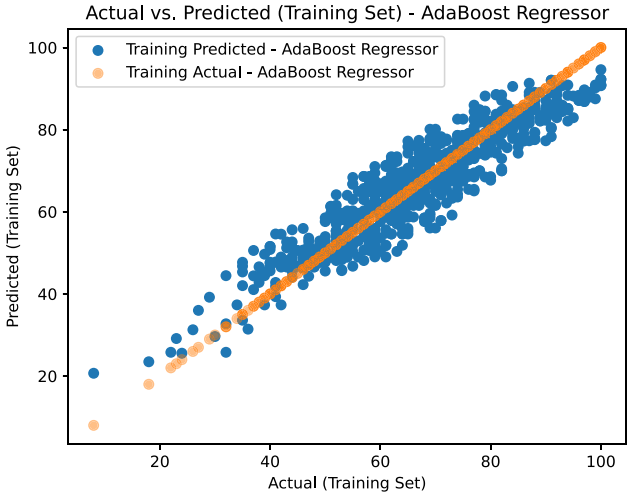

Fig. 21 Training Data Model Prediction by Ada Boosting

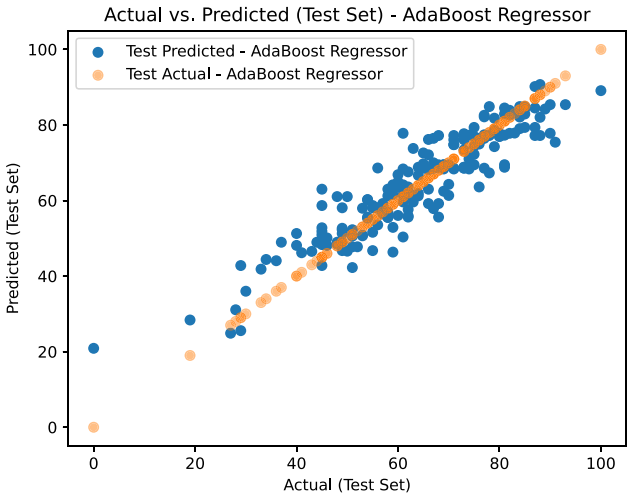

Fig. 22 Testing Data Model Prediction by Ada Boosting

percent with a test data accuracy of about 85.24 percent (Table 3). These results are quite reasonable and demonstrate the efficacy of the model while staying reasonable in terms of model detail and not over-complicating the process.

The accuracy of predicted values has been further depicted through bar diagrams in figures 21 and 22 through which it has been illustrated that the model has both learnt and predicted almost accurately as the actual values meant for training as well as testing sets. These graphical illustrations extend credence to the generality of the model to derive accurate predictions over other instances since the tests are built to represent all instances.

4.6 LASSO Regression model

Lasso Regression is widely used because the technique emphasizes on simplicity or selection of features; the model is good at handling with Multi-collinearity since large coefficients are penalized and some of them are shrunk to zero. This characteristic makes it possible for the model to stay interpretable as well as not compromised by numerous correlated predictors.

In the present work, Lasso Regression found to have a training accuracy of 80.71% and test accuracy of 82.53% (Table 3). These outcomes suggest its excellent ability to generalize across data sets while possessing low bias and high variance.

Figures 23 and 24 are conventional methods of analyzing the possibility of the actual test data being captured by the model by comparing the obtained test values with the predicted values. These visualizations ensure versatility in the chosen model providing essentially operational, accurate, and maintainably complex forecast results. Therefore, Lasso Regression can be considered a helpful contribution to the list of other predictive models used in this study because,

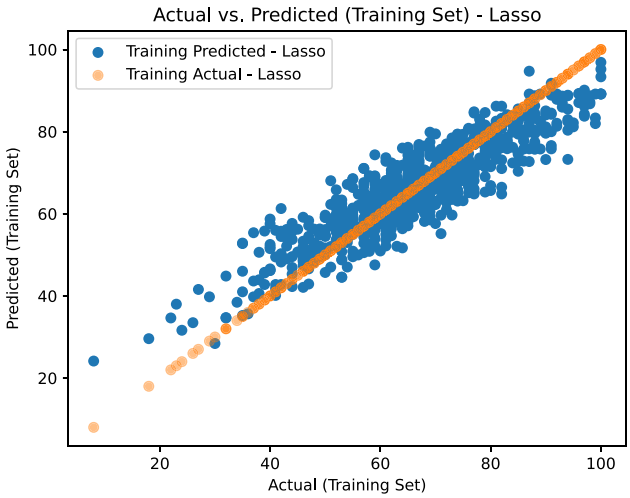

Fig. 23 Training Data Model Prediction by LASSO Regression

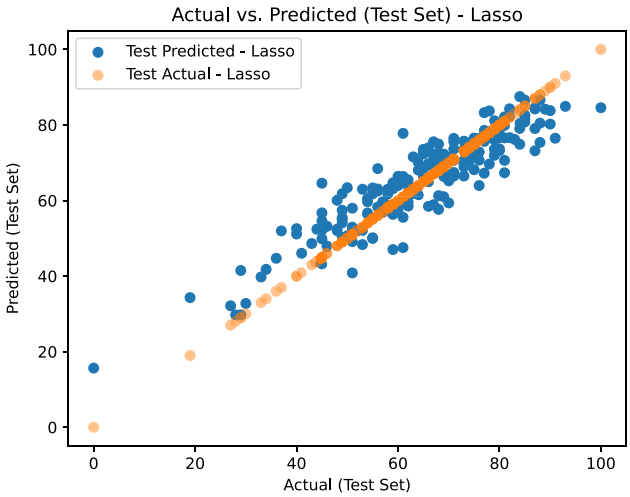

Fig. 24 Testing Data Model Prediction by LASSO Regression

despite its reasonable accuracy, it requires less complex calculations. This makes Lasso Regression practical to apply in education data analysis since there is always a balance between a model's complexity and efficiency.

4.7 Decision Tree Regression model

The Decision Tree Regression model provides an effective strategy to forecast student performance through its simple yet impactful methodology. The model creates a tree structure for systematic data division into smaller subsets while using input variables as partition criteria. By implementing hierarchical partitioning across the model it can identify unique data patterns thus proving the model as an effective tool for student outcome evaluation. The model shows value for educational data analysis because it provides both understandable results together with its ability to examine both numeric and categorical inputs.

This research used Decision Tree Regression model which successfully captured patterns in the dataset during training sessions. The model achieves an 82% accuracy level on the test dataset (Table 3) indicating possible overfitting problems. When models display outstanding results on known data they neglect to effectively recognize unknown and fresh information. The model's reliability under real-world applications is limited by overfitting and this deficiency requires implementation of techniques such as pruning or ensemble methods with cross-validation to improve generalization.

The accuracy of the proposed model is presented graphically in Figs. 25 and 26. Presenting these results in a visual way shows the great fit between the forecast and the actual values and also proves the models ability to find important patterns and correlations in the data. This is however done at the cost of generalization of the Decision Tree Regression whereby it is therefore recommended to use with other techniques of regularization before using it for understanding and making prediction on student performance.

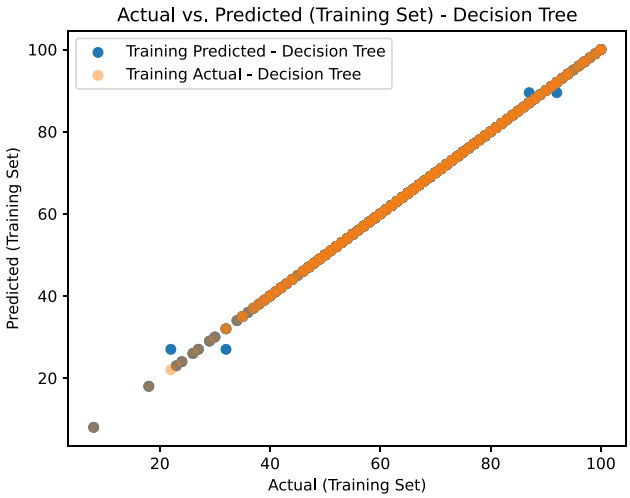

Fig. 25 Training Data Model Predictions by Decision Tree

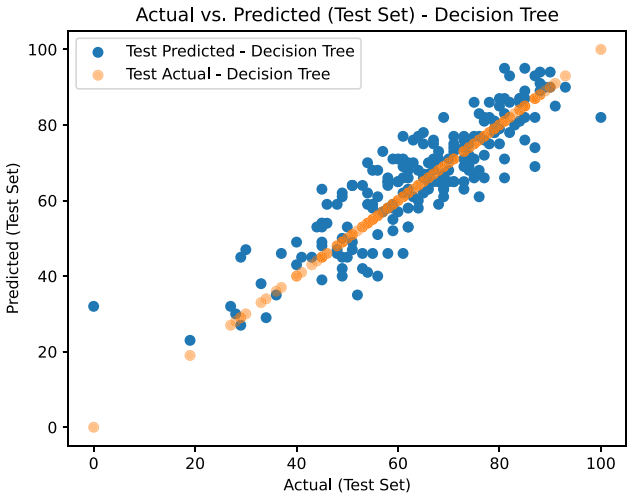

Fig. 26 Testing Data Model Predictions by Decision Tree

4.8 K-Neighbours Regression model

K Nearest Neighbors (KNN) Regression model which is famously known for its simplicity and efficiency uses distance measure to estimate results. In this study, using KNN Regression, gains a training accuracy of 84.52% and testing accuracy of 79.07% (Table 3). The results presented in this paper indicate that the proposed model can be used for analyzing the data and that its efficiency is somewhat lower compared to other similar models.

Figures 27 and 28 are used to present the graphical representation of model values and actual values separately. Such tight coupling illustrated in these figures confirms that the model yields predictive results based on the neighboring values. Mostly, the KNN Regression model is found to be useful for unveiling local patterns and trends of the dataset despite the fact that building the model may take a very long time when tested on a big dataset.

4.9 Comparison of all eight Boosting Regression Models

Figure 29 presents a detailed comparative analysis of all eight regression models, showcasing their performance metrics across key indicators such as accuracy, mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE). This side-by-side comparison offers a clear overview of how each model fares in predicting student performance, allowing for an in-depth evaluation of the strengths and weaknesses of each approach. By highlighting the differences in accuracy and error metrics, Fig. 29 provides valuable insights into the relative effectiveness of each model in handling the dataset.

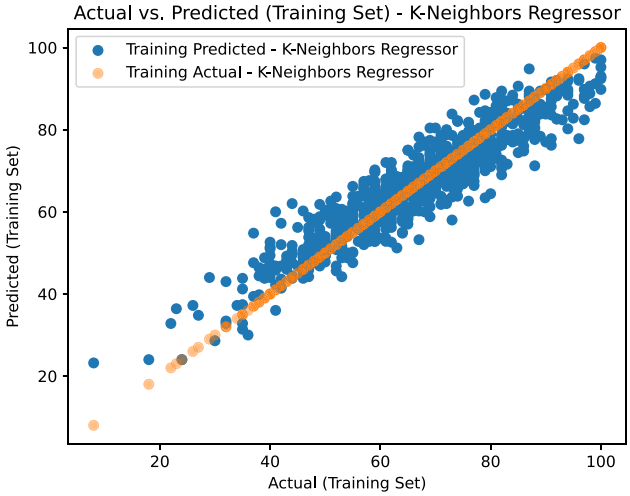

Fig. 27 Training data model prediction by K-Neighbours

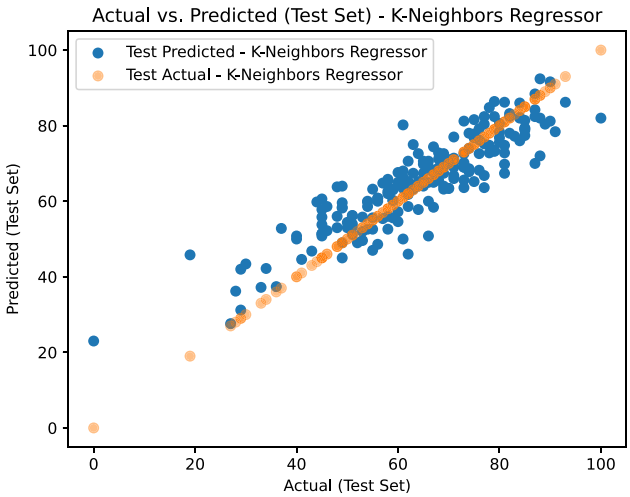

Fig. 28 Testing data model prediction by K-Neighbours

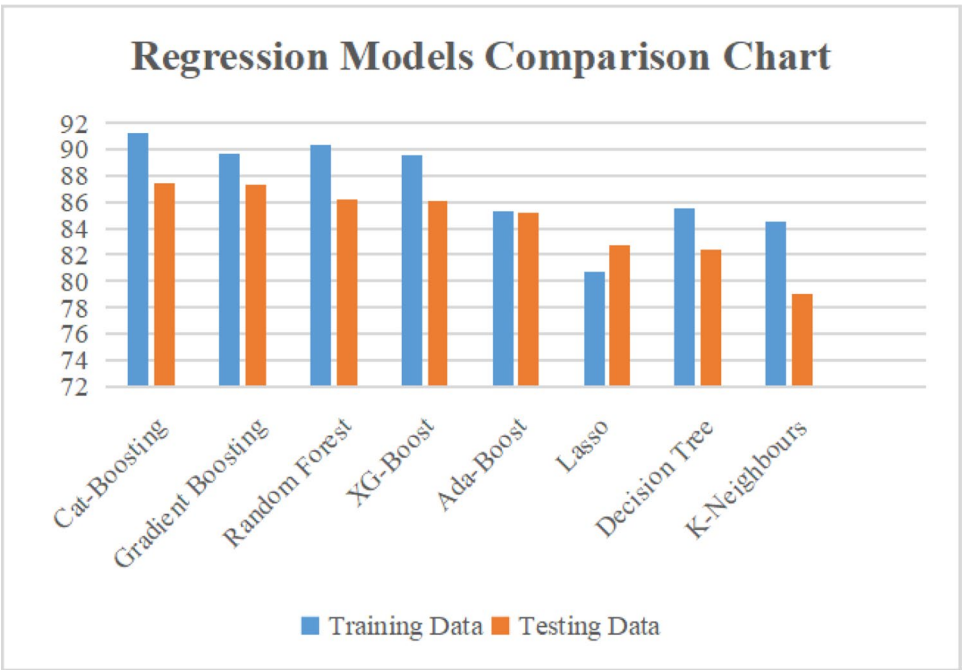

Fig. 29 Comparison of All eight Regression Models

In addition, Figs. 30 and 31 provide a more detailed analysis of the performance evaluation of the proposed approach since it plots the ERROR curve of the least square measure for both training and testing datasets. These plots aid in evaluating the performance of each model in extrapolation as such the future prediction of new data. Figures 30

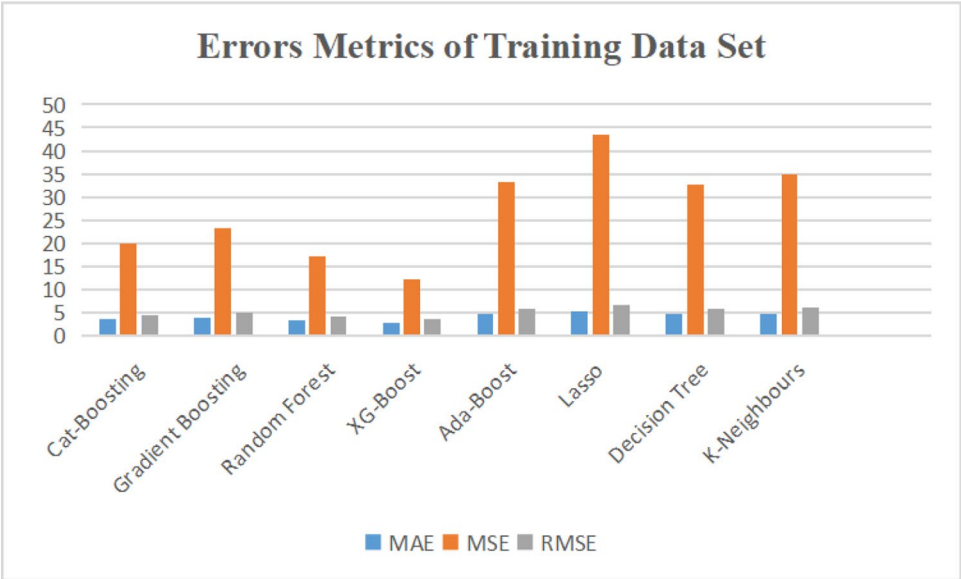

Fig. 30 Errors Metrics of Training Data of all Regression Models

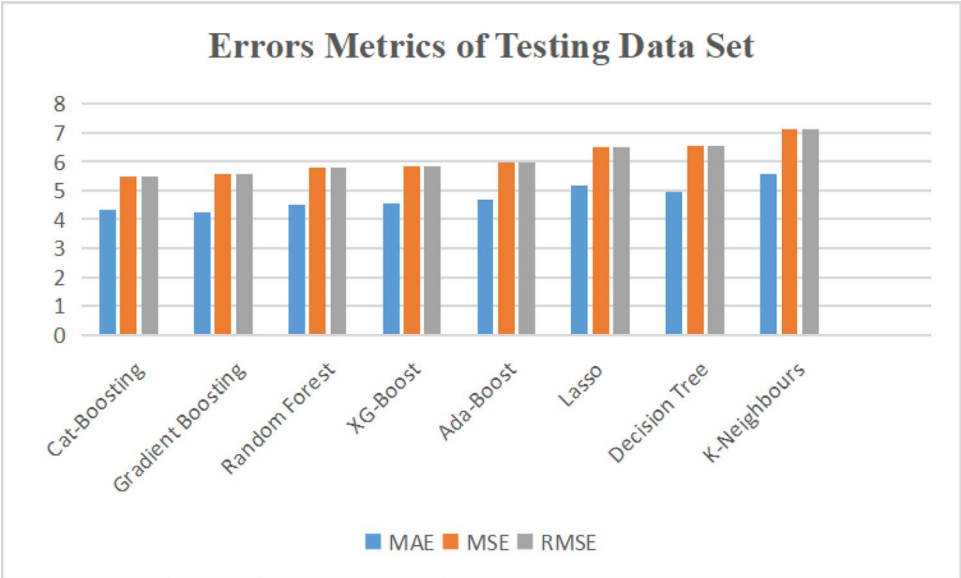

Fig. 31 Errors Metrics of Testing Data of all Regression Models

Table 3 Result Scores and errors metrics of all models

Serial No.	Regression Model	Train Data Accuracy	Test Data Accuracy	MAE	MSE	RMSE
1	Cat-Boosting	91.18	87.46	4.32	5.52	5.52
2	Gradient Boosting	89.67	87.29	4.25	5.56	5.56
3	Random Forest	92.38	86.21	4.5	5.79	5.79
4	XG-Boost	94.57	86.08	4.56	5.82	5.82
5	Ada-Boost	85.25	85.24	4.68	5.99	5.99
6	Lasso	80.71	82.53	5.15	6.51	6.51
7	Desion Tree	85.50	82.42	4.93	6.54	6.54
8	K-Nieghbours	84.52	79.07	5.56	7.13	7.13

and 31 allow moreover the setup of an accurate analysis of how each regression model performs depending on what specific situation it is applied to, while focusing on the models, to make sure they provide the most accurate depiction of the data trends and relationships while avoiding overfitting as much as possible. Together, these measures help to assess model performance to identify the best performing model in predicting student performance.

The findings of this study indicate that eight of the applied regression models including CatBoost Regression, Gradient Boosting Regression, Random Forest Regression, XGBoost Regression, AdaBoost Regression, Lasso Regression, Decision Trees Regression, K-Nearest Neighbors Regression are effective in developing student success prediction. All models show their effectiveness in establishing the correlation between the input variables and the students' achievement with high accuracy and using complicated data sets.

Nevertheless, the most superior model among these models is CatBoost Regression. Owns the highest test accuracy of 87.46% coupled with the lowest measures of errors making it the best model closer to reality. This is especially true in prediction problems where both high accuracy and low error-rate is desirable in generating valuable insights. The existence of Categorical Feature Handling and Resistance to Overfitting makes CatBoost stand out and powerful when it gets to the top. While other models such as Gradient Boosting Regression, Random Forest, and XGBoost also perform admirably, reaching accuracy rates of around 86–89%, they do not match the CatBoost model in both accuracy and error metrics. For instance, although Gradient Boosting achieves impressive accuracy and handles complex data well, it is still susceptible to minor over-fitting, as indicated by its slightly higher error rates compared to Cat-Boost. Similarly, Random Forest and XG-Boost, though effective, exhibit slightly higher error metrics (MSE, MAE, RMSE). Hence, although all models show their effectiveness to forecast the students performance, optimum and effective Cat-boost model offers opportunities with precise forecast and lesser prediction error. Taking into account its stability, stated scalability, and versatility concerning features' input, we introduce Cat-Boost Regression as the most suitable model to be used for the student success prediction in education data analysis. Its outstanding performance makes it a reliable tool for choosing educational institutions to use in detecting their learning risk learners and providing ways to regain academic success.

4.10 Feature importance analysis

Table 4 presents the top five most influential features in predicting student performance, as identified using the CatBoost model based on SHAP (SHapley Additive exPlanations) values. The most significant predictor is Parental Education Level, with an importance score of 0.23, indicating that students with more educated parents tend to achieve better academic outcomes. Test Preparation follows with an importance score of 0.19, highlighting the positive impact of structured preparation programs on student performance.

Additionally, Race and Lunch Type were identified as influential factors with importance scores of 0.15 and 0.12 respectively. These features may reflect underlying socioeconomic and demographic factors affecting students' educational experiences. Lastly, Math Score with an importance score of 0.11, suggests that mathematical proficiency is closely linked to overall academic success, reinforcing the interdependence of different subject areas.

The findings emphasize the need for targeted interventions, such as enhancing parental involvement, implementing structured test preparation initiatives, and ensuring equitable access to educational resources. By leveraging these insights, educators and policymakers can make informed decisions to support students more effectively and improve learning outcomes.

Table 4 Top five important features identified using the CatBoost model based on SHAP values

Feature	Importance score
Parental Education Level	0.23
Test Preparation	0.19
Race	0.15
Lunch Type	0.12
Math Score	0.11

4.11 SHAP analysis for explainability

Figure 32 presents the SHAP summary plot, which illustrates the contribution of different features to the predictive outcomes of the CatBoost model. This plot provides an interpretable visualization of feature importance by displaying the magnitude and direction of each feature's impact on student performance predictions. Each point in the plot represents a single data instance, with colors indicating feature values (e.g., high or low). Features positioned at the top contribute the most to the model's decision-making process.

The results indicate that Parental Education Level, Test Preparation, and Race are among the most influential features in predicting student performance, aligning with the feature importance rankings shown in Table 4. The spread of SHAP values for these features suggests that variations in parental education and test preparation status significantly affect predicted student outcomes. Additionally, categorical attributes, such as Lunch Type and Race, also exhibit notable effects on predictions. The SHAP analysis enhances the model's interpretability by explaining how specific attributes drive the predictive results, thereby offering actionable insights for educators and policymakers.

While the tree-based importance scores provide a global view of feature influence, SHAP enables a local and detailed explanation of predictions. Future work will incorporate more extensive SHAP analysis to refine intervention strategies and policy recommendations for educational institutions.

4.12 Cross-validation and overfitting analysis

To ensure model robustness and mitigate overfitting, a **5-fold cross-validation** strategy was employed. The CatBoost model achieved a mean **Mean Absolute Error (MAE)** of **25.1932** with a standard deviation of **3.3016** across the five folds. The relatively low standard deviation indicates that the model maintains consistent predictive performance across different subsets of the dataset, reinforcing its generalization ability. These results confirm that the model is not overly fitted to the training data, as its performance remains stable across various validation sets.

These findings demonstrate that the CatBoost model maintains reliable predictive accuracy across various partitions of the dataset, reducing the likelihood of overfitting.

4.13 Discussion on results and policy implications

This section provides an in-depth explanation of the methodology used in this study and highlights the key findings. The research employs a machine learning approach, specifically CatBoost Regression, to determine whether student

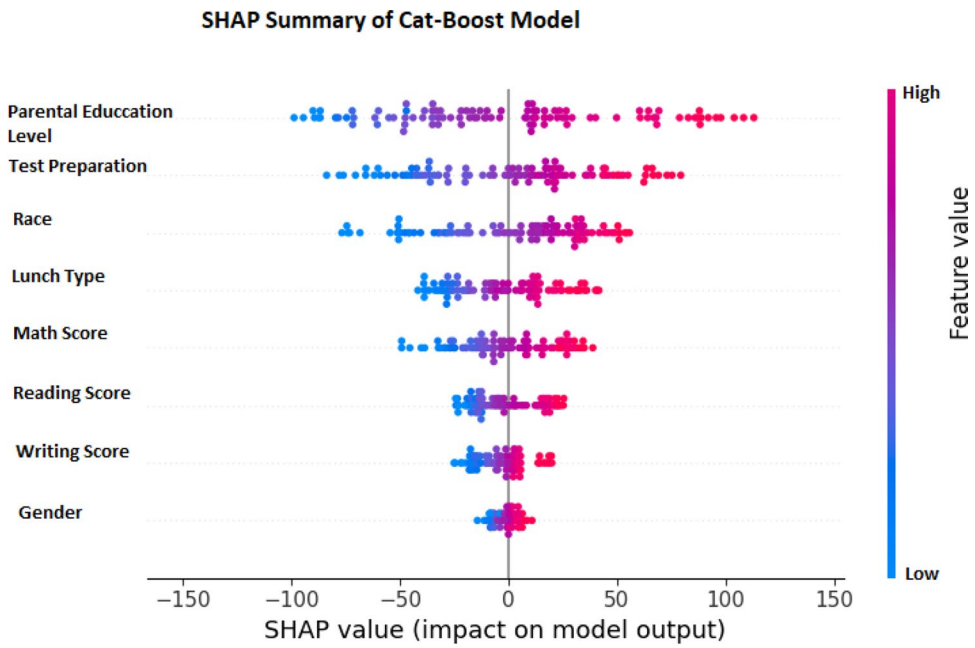

Fig. 32 SHAP summary plot illustrating the impact of different features on student performance predictions

performance can be accurately predicted based on their demographic and academic profiles. The findings underscore the importance of incorporating socio-demographic factors alongside academic performance in educational prediction models. By integrating these variables, the study offers deeper insights into the factors influencing student outcomes, emphasizing the necessity of leveraging machine learning techniques for enhanced educational analytics and informed decision-making. Here is a summary of the key steps and findings:

Data collection and methodology: The data for the study comprises variables, including gender, race, parents 'education level, lunch and exam. Each of these features would contribute to giving the different variables which affect student learning a colorful backdrop. In this research, student records database contains 1000 students and the data were cleaned in a way that removed missing data and redundant data to fit the modeling.

Key findings and model performance: The findings show that the accuracy of the proposed method, namely the CatBoost Regression model, equals about 87.46 % and surpasses other machine learning techniques in both accuracy and less error metrics (Table 3). This is particularly remarkable since other models fail to identify such associations within the data set. A subject score total named 'Total' was made in the model to complement the separate subject scores and the addition of this score was very essential for improving the model's forecasting capability. This feature engineering step appended an overall examination of the student's performance and enhanced the competency of the predictions.

Policy implications: The conclusion of this study has major implications for the policies governing educational units. The advantage of a model of this type is the possibility of early detecting those students who are likely to be performing poorly and therefore may require some extra attention and support. These findings may be used by educational institutions to find needs-based measures, including extra student support and curriculum modifications. In addition, the study discusses the relevance of analytics in realizing data-informed decisions for learning, as well as how score predictions can enhance tutor performance and student approaches to learning and teaching.

Future research directions:

While the study offers valuable contributions, there are several avenues for future research that can build upon the foundation laid in this work. First, expanding the dataset to include more diverse educational environments and additional socio-economic factors could further enhance the generalizability of the model. Additionally, incorporating other machine learning techniques, such as deep learning or ensemble methods, may improve prediction accuracy, particularly when dealing with more complex data. Investigating the interpretability of the models, especially for stakeholders like educators and policymakers, will be crucial to ensure that the predictions made by the model can be easily understood and used for decision-making.

Furthermore, this research could be developed to describe changes following interventions established based on expectations of the model. For example, the link between the identification of students at risk for poor academic performance and the effectiveness of conventional forms of targeted interventions is still an issue for future research. The coordination of the longitudinal data may offer important information on the dynamics of student outcomes over time, which would increase the temporal perspective of the predictive models.

Ultimately, this study demonstrates the immense potential of machine learning in the field of education, not only for predicting student performance but also for guiding decision-making processes that aim to improve educational equity and success across diverse student populations.

5 Concluding remarks

This study has significant implications for both data analysis and education. Applying machine learning algorithms and techniques to predict and analyze student performance based on various parameters proves to be highly relevant and beneficial. The study underscores its importance for educational institutions, offering valuable insights that can be used to evaluate and enhance student performance. By identifying key factors affecting performance, institutions can tailor their curricula and support systems to better meet students' needs and address specific performance gaps.

Implications for educational institutions: This study has the findings that can act as useful recommendations to the educational institutions. After grasping the key factors that are making students achieve very low, institutions of learning can adopt the same strategies. Educational institutions can then utilize the models built in this study to provide personalized interventions and learning programs focused on students that they identify as being at the highest risk of underperforming. Better tailoring support will also help improve students' academic outcomes.

Implications for school administration: The study has implications for school administrators and may be useful to practitioners since it identified the current student places as well as their preferences in areas close to the schools.

If these areas and factors which define the student success can be identified more effectively by schools then, the intervention models and policies that are in place for the underperforming students can be adjusted. The use of machine learning means that data can be analyzed to predict the achievement of certain objectives through an optimized learning plan to help the specific individual ends up improving their academic performance. The outcomes enable the lecturers and policy makers come up with the right decisions of enhancing the quality of education and the student performances. Risk assessment can help in making regular interventions which are essential in enhancing the performance of students.

Implications for parents: This research reveals different factors which affect how well children achieve in school for their parents to know. The obtained understanding enables parents to better support their children by recognizing both their skills and difficulties which might require additional help. Parental knowledge regarding academic performance indicators leads to better family participation in educational support and appropriate home guidance decisions.

The role of machine learning in education: The research shows educational institutions now depends more on machine learning and data analysis. Advanced data collection tools let educational organizations use machine learning to monitor student performance right when it happens and respond promptly to needs. The technology can change school equity practices by delivering teaching support specifically designed to help students master their subjects. Educational organizations use this technology to create better learning opportunities for students that generate optimal results in specific educational settings. Educators can design learning plans that match students' individual requirements by analyzing data which improves learning results.

Python implementation and accessibility: The results indicate that Python's data analysis and machine-learning techniques are both accessible and efficient for such tasks. As an open source programming language, Python allows other educators and researchers to re implement and or modify this research methodology. Broad applicability in analyzing student data enables to achieve this feature - transparency and reproducibility of the research process. These techniques can then be levered by educational institutions to analyze student data and to improve academic outcomes at scale.

Recommendations for future work: This research can be used as a foundation for future research that is focused on data analysis and education. This phase can be extended to future studies to analyze more robust datasets and other factors and predictive variables that are associated with the student performance, in conjunction with actual education contexts. In addition, the dataset can be expanded with incorporating the real educational data rather than the publicly available Kaggle datasets in order to increase the generalizability of the findings. Deep learning methods for analyzing such complex educational data, which are promising, could also be explored for future research. Additional capabilities will be added to expand the features included in the model, increase model explainability by utilizing SHAP, LIME, as well as incorporate cross validation techniques to increase model robustness. Furthermore, crossinstitutional testing can point to usefulness of the model for other student demographics and educational environment.

This study finally demonstrates the predictive and enhancement potential of machine learning and educational data mining for students' performance. It is clear through analyzing explorations of different machine learning models and techniques that data driven approaches can provide useful insights into student behavior, detect at risk learners and inform targeted interventions. Despite challenges that still exist, including data quality and model optimization, the use of advanced algorithms is likely to change the face of education practices, help improve retention rates, and enhance personalized learning experiences. Future research should then concentrate on the refinement of these models using more varied datasets, and to explore possible practical applications of these models in many students' educational environments.

Acknowledgements The Heriot-Watt University has supported the research.

Author contributions 1) Muhammad Nadeem Gul: Conceptualization, methodology design, supervision, manuscript writing, result interpretation, and final revisions. 2) Waseem Abbasi (CA): Data collection, preprocessing, statistical analysis, and literature review. 3) Muhammad Zeeshan Babar: Implementation of machine learning models, evaluation of model performance, and manuscript editing. 4) Abeer Aljohani: Experimental validation, visualization of results, and review of machine learning techniques. 5) Muhammad Arif: Implementation of machine learning models, hyperparameter optimization, refinement of experimental results, and manuscript writing contributions.

Funding The Heriot-Watt University provided funds for this research.

Data availability The dataset used in this study can be accessed at https://www.kaggle.com/datasets/spscientist/students-performance-in-exams/data

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors affirm that none of their known personal relationships or financial or Conflict of interest may have seemed to impact the work presented in this paper.

References

1. Baniata LH, Kang S, Alsharaiah MA, Baniata MH. Advanced deep learning model for predicting the academic performances of students in educational institutions. Appl Sci. 2024;14(5):1963.
2. Hussain S, Khan MQ. Student-performulator: predicting students' academic performance at secondary and intermediate level using machine learning. Ann Data Sci. 2023;10(3):637–55.
3. Siram J, Chikati Srinu DS, Tripathi A. Towards a framework for performance management and machine learning in a higher education institution. J Inf Educ Res. 2024 1;4(2).
4. Jin Z, Zainudin Z. E-learning outcomes of engineering college students prediction model based on machine learning technique. ICCCM J Soc Sci Hum. 2024;3(5):76–90.
5. Khan MI, Khan ZA, Imran A, Khan AH, Ahmed S. Student performance prediction in secondary school education using machine learning. In2022 8th International Conference on Information Technology Trends (ITT) 2022 25 (pp. 94-101). IEEE.
6. Issah I, Appiah O, Appiahene P, Inusah F. A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. Decision Anal J. 2023;1(7): 100204.
7. Kaggle. (n.d.). Students Performance in Exams. collected from https://www.kaggle.com/datasets/spscientist/students-performance-in-exams/data.
8. Nanavaty S, Khuteta A. A deep learning dive into online learning: predicting student success with interaction-based neural networks. Int J Intell Syst Appl Eng. 2024;12(1):102–7.
9. Ouyang F, Wu M, Zheng L, Zhang L, Jiao P. Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. Int J Educ Technol Higher Educ. 2023;20(1):4.
10. Yağcı M. Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environ. 2022;9(1):11 (**3**).
11. Arashpour M, Golafshani EM, Parthiban R, Lamborn J, Kashani A, Li H, Farzanehfar P. Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization. Comput Appl Eng Educ. 2023;31(1):83–99.
12. Ajibade SS, Dayupay J, Ngo-Hoang DL, Oyebode OJ, Sasan JM. Utilization of ensemble techniques for prediction of the academic performance of students. J Optoelectron Laser. 2022;41(6):48–54 (**5**).
13. Forero-Corba W, Bennasar FN. Techniques and applications of Machine Learning and Artificial Intelligence in education: a systematic review. RIED-Revista Iberoamericana de Educación a Distancia. 2024;27(1).
14. Hussain M, Zhu W, Zhang W, Abidi SM, Ali S. Using machine learning to predict student difficulties from learning session data. Artif Intell Rev. 2019;1(52):381–407.
15. Namoun A, Alshanqiti A. Predicting student performance using data mining and learning analytics techniques: a systematic literature review. Appl Sci. 2020;11(1):237 (**29**).
16. Waheed H, Hassan SU, Aljohani NR, Hardman J, Alelyani S, Nawaz R. Predicting academic performance of students from VLE big data using deep learning models. Comput Hum Behav. 2020;1(104):106189.
17. Salal YK, Abdullaev SM, Kumar M. Educational data mining: student performance prediction in academic. Int J Eng Adv Technol. 2019;8(4C):54–9.
18. Jacob ON, Abigeal I, Lydia AE. Impact of COVID-19 on the higher institutions development in Nigeria. Electron Res J Soc Sci Hum. 2020;2(2):126–35.
19. Acharya A, Sinha D. Early prediction of students performance using machine learning techniques. Int J Comput Appl. 2014;107(1):37–43.
20. Kaensar C, Wongnin W. Analysis and prediction of student performance based on Moodle log data using machine learning techniques. Int J Emerg Technol Learn. 2023;18(10):184–203 (**15**).
21. Holicza B, Kiss A. Predicting and comparing students' online and offline academic performance using machine learning algorithms. Behav Sci. 2023;13(4):289.

22. Rejeb A, Rejeb K, Appolloni A, Treiblmaier H, Iranmanesh M. Exploring the impact of ChatGPT on education: a web mining and machine learning approach. Int J Manag Educ. 2024;22(1):100932.
23. Ni L, Wang S, Zhang Z, Li X, Zheng X, Denny P, Liu J. Enhancing student performance prediction on learnersourced questions with sgnn-llm synergy. InProceedings of the AAAI Conference on Artificial Intelligence 2024(Vol. 38, No. 21, pp. 23232-23240).
24. Vijayalakshmi V, Venkatachalapathy K. Comparison of predicting student's performance using machine learning algorithms. Int J Intell Syst Appl. 2019;11(12):34.
25. Ghorbani R, Ghousi R. Comparing different resampling methods in predicting students' performance using machine learning techniques. IEEE Access. 2020;9(8):67899–911.
26. Asthana P, Mishra S, Gupta N, Derawi M, Kumar A. Prediction of Student's Performance With Learning Coefficients Using Regression Based Machine Learning Models. IEEE Access. 2023.
27. Ho IM, Cheong KY, Weldon A. Predicting student satisfaction of emergency remote learning in higher education during COVID-19 using machine learning techniques. PLoS ONE. 2021;16(4):e0249423 (**2**).
28. Li S, Liu T. Performance prediction for higher education students using deep learning. Complexity. 2021;2021(12):1.
29. Çakıt E, Dağdeviren M. Predicting the percentage of student placement: a comparative study of machine learning algorithms. Educ Inf Technol. 2022;27(1):997–1022.
30. Harif A, Kassimi MA. Predictive modeling of student performance using RFECV-RF for feature selection and machine learning techniques. Int J Adv Comput Sci Appl (IJACSA). 2024;15(7).
31. Badal YT, Sungkur RK. Predictive modelling and analytics of students' grades using machine learning algorithms. Educ Inf Technol. 2023;28(3):3027–57.
32. Rastrollo-Guerrero JL, Gómez-Pulido JA, Durán-Domínguez A. Analyzing and predicting students' performance by means of machine learning: a review. Appl Sci. 2020;10(3):1042
33. Kausar G, Saleem S, Subhan F, Suud MM, Alam M, Uddin MI. Prediction of gender-biased perceptions of learners and teachers using machine learning. Sustainability. 2023;15(7):6241
34. Chen W, Shen Z, Pan Y, Tan K, Wang C. Applying machine learning algorithm to optimize personalized education recommendation system. J Theory Pract Eng Sci. 2024;4(01):101–8 (**1**).
35. Song Y, Cutumisu M. Using machine learning to predict student science achievement based on science curriculum type in TIMSS 2019. Int J Sci Educ. 2024;14:1–45.
36. Agarwal A, Das RR, Das A, Development of an Essential Education Performance Prediction Tool Using Machine Learning. InAnalytics Global Conference,. 7. Cham: Springer Nature Switzerland; 2024. p. 217–26.
37. Dad I, He J, Noor W, Samad A, Ullah I, Ara S. Cross classification matrix to evaluate the performance of machine learning algorithms in predicting students performance of developing regions. SN Comput Sci. 2024;5(5):621.
38. Adu-Twum HT, Sarfo EA, Nartey E, Adesola Adetunji A, Ayannusi AO, Walugembe TA. Role of advanced data analytics in higher education: using machine learning models to predict student success. J Data Sci Artif Intell. 2024;3(1).
39. Parkavi R, Karthikeyan P, Sujitha S, Abdullah AS. Enhancing educational assessment: predicting and visualizing student performance using EDA and machine learning techniques. J Eng Educ Transform. 2024;37(Special Issue 2).