International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Developing a predictive model for student academic performance using machine learning techniques

Abdul-waliyyu Bello [1,*], Idris Ajibade [1], Idris Wonuola [1] and Darlington Ekweli [2]

[1] *Department of Mathematics and Statistics, Austin Peay State University, Tennessee, USA.*
[2] *Department of Healthcare Administration, University of the Potomac, Washington, USA.*

International Journal of Science and Research Archive, 2025, 16(03), 219–234

Publication history: Received on 27 July 2025; revised on 01 September 2025; accepted on 04 September 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.3.2539

Abstract

This study investigates the predictive capability of machine learning techniques in forecasting student academic performance using school-level and demographic data. A structured, publicly available dataset from the District of Columbia Public Schools was employed, comprising 1,163 records representing various student groups and institutional contexts. After preprocessing and feature selection, three regression models were developed and evaluated: a baseline Linear Regression model, Random Forest Regressor, and XGBoost Regressor. The baseline model demonstrated limited predictive strength (R^2 = .32, MAE = 13.79), while ensemble models significantly outperformed it. Random Forest achieved an R^2 of .69 and MAE of 7.74, capturing complex interactions more effectively. XGBoost slightly outperformed Random Forest with an R^2 of .70 and MAE of 7.19, showing stronger generalization and sensitivity to underrepresented groups. Feature importance analysis revealed that institutional factors such as Framework Points Earned strongly influenced predictions in Random Forest, whereas XGBoost emphasized subgroup characteristics, including Students with Disabilities, English Learners, and At-Risk populations. These findings highlight the strengths of ensemble methods in modeling non-linear and multidimensional educational data while raising questions about the trade-offs between model accuracy and equity. The study concludes that predictive models should be evaluated not only by statistical performance but also by their capacity to inform equitable interventions in education. Recommendations include the ethical deployment of predictive systems, incorporation of contextual data, and prioritization of fairness in model selection to support inclusive, data-informed educational policy and practice.

Keywords: Academic Performance; Machine Learning; Student Group Score; Random Forest; Xgboost; Educational Data Mining; Predictive Modeling; Feature Importance; Fairness; Public School Analytics.

1. Introduction

Machine learning (ML) has become a strong device for predicting the academic achievement of students due to the growing need for data-driven approaches in education. Alot of student data are now used by educational institutions to find hidden patterns that allow for better learning outcomes and prompt intervention (Chaka 2022). When properly developed, Alam (2023) stated that predictive models can help teachers identify students who are at risk and inform individualized teaching methods. Since early identification of performance problems can influence future academic success, these innovations are especially pertinent in secondary education. Therefore, Baek & Doleck, (2022) noted educational data mining has been altered from a theoretical investigation into a useful, significant solution by machine learning techniques.

Furthermore, machine learning models are preferred in academic prediction because of their ability to manage large, complex datasets while delivering accurate results. Several algorithms including decision trees, support vector

* Corresponding author: Abdul-waliyyu Bello

machines, and neural networks have been employed to predict academic success with considerable accuracy (Balaji et al., 2021). For example, Olabanjo et al., (2022) study noted that radial basis function neural networks have shown promise in modeling non-linear student performance patterns. The study noted that hybrid models that combine multiple algorithms have been proven effective in handling diverse learning environments and data types.

Additionally, the quality and applicability of input data, as well as the techniques employed for preprocessing and feature selection, all have a substantial impact on how well predictive models work (Ayienda et al., 2021). According to research by Rastrollo-Guerrero et al., (2020), the most significant predictors of student performance are academic, behavioral, and sociodemographic factors. Also, feature engineering techniques such as encoding, normalization, and dimensionality reduction improve model performance by eliminating noise and redundancy (Chakrapani & Chitradevi, 2022). Furthermore, Hussain & Khan (2023) study stated that models can more accurately reflect the dynamics of learning in the real world by incorporating data such as attendance, parental background, and engagement on virtual platforms.

Correspondingly, Olaleye & Vincent (2020) noted that the evaluation of model is necessary to confirm the accuracy and suitability of prediction tools. The study further stated that efficacy of models on both balanced and imbalanced datasets is frequently evaluated using metrics like recall, accuracy, F1-score, precision, and AUC-ROC. Furthermore, model stability and generalization ability can be gleaned from cross-validation techniques like k-fold validation (Ha et al., 2020). In certain situations, Matzavela & Alepis (2021) noted that even though more complex models like deep neural networks achieve slightly higher accuracy, simpler models like decision trees are preferred for their interpretability. This emphasizes how accuracy and usability must be balanced in educational settings.

Furthermore, effective performance monitoring and intervention are made possible by learning analytics frameworks, which offer structured methods for integrating machine learning into educational systems (Khalil et al., 2022). According to Zhao et al., (2023), a human-centered approach enables predictive models to be more closely aligned with the needs of learners. Also, Joshi et al., (2020) study also showed how crucial it is to integrate predictive analytics with problem-based learning settings in order to enhance teaching methods and student outcomes. The ethical and pedagogical integration of machine learning (ML) into classroom settings is guided by these frameworks, which guarantee that models not only make accurate predictions but also promote equity and learner growth.

Nevertheless, Chaka (2022) stated that predictive modeling's ethical issues have spurred discussions about openness, equity, and data privacy in the classroom. Especially when applied to high-stakes decisions, models trained on incomplete or biased datasets run the risk of perpetuating educational disparities. Also, Baek & Doleck (2023) stated that interpretability is essential for assisting educators in comprehending and having faith in model outputs, as stated by. Instead of replacing teacher expertise. However, Alam (2023) study advocates for more human-in-the-loop systems in which predictive outputs enhance it. Although, Albreiki et al., (2021) study noted that protecting student information is another aspect of ethical deployment, especially in situations where vulnerable groups may be disproportionately impacted by digital surveillance.

In addition, the practical implementation of machine learning models in educational settings has been confirmed by different empirical investigations. Also, the Student-per formulator model, advocated by Hussain et al., (2023) introduced for secondary schools, effectively predicts academic performance and permits targeted support. Similarly, Olaleye & Vincent (2020) utilized supervised learning techniques to categorize students according to their risk levels, which influenced early warning systems. Furthermore, after comparing several classification models, Sekeroglu et al., (2019) established that decision trees and support vector machines provide excellent speed and accuracy for educational data. These studies highlight how ML applications in academic forecasting are becoming more mature, and this study aims to further validate these models for generalizability in various socio-educational contexts. To therefore, the study seek out to evaluate the impact of using machine learning techniques in predicting student academic performance.

2. Literature Review

The literature review examines previous studies on the use of machine learning techniques to predicting student academic performance. It provides the lens through which the subject matter can be highly comprehensive.

2.1. Overview of Student Academic Performance Prediction

The availability of educational data and developments in computational intelligence have fueled a strong increase in the use of machine learning to predict students' academic performance in recent years. An overview that emphasizes the

incorporation of data mining, prediction techniques, and tools that have become fundamental in educational research is given by Chaka (2022). Also, Balaji et al., (2021) provide additional support for this trend by pointing out that a variety of machine learning models, including neural networks, decision trees, and support vector machines, have been used to predict academic outcomes. However, Albreiki et al., (2021) noted that early detection of at-risk students is made possible by these models, allowing for prompt intervention while reliability and ethical concerns remain. Also, the growing use of automated prediction models indicates a move toward data-driven education management.

Despite the well-established efficacy of predictive models, a number of studies draw attention to the variety of algorithmic strategies and dataset contexts that are employed in the prediction of academic performance. Also, Ogundele et al., (2024) present a real-world application that trains models such as logistic regression and random forest with institutional datasets, producing high accuracy. Likewise, Olabanjo et al., (2022) investigate the application of radial basis function neural networks, which are highly effective in simulating non-linear correlations between academic characteristics and performance results. Notwithstanding encouraging outcomes, Rastrollo-Guerrero et al., (2020) stress that variations in feature selection, dataset quality, and preprocessing methods have a substantial impact on model performance, making cross-study comparisons challenging. This demonstrates the necessity of uniform approaches and standards in the field of educational predictive modeling.

Furthermore, The incorporation of behavioral, psychological, and environmental factors into machine learning models to enhance accuracy and contextual relevance is another area of current research in student performance prediction. Predictive systems are more accurate when cognitive and non-cognitive data, such as attendance, engagement metrics, and socioeconomic background, are combined (Albreiki et al., 2021). However, Chaka (2022) study advocated for a more multidisciplinary strategy that integrates learning analytics, data ethics, and human-centered AI principles. Also, the paucity of research that goes beyond prediction to actual intervention design or alignment with educational policies is highlighted by Balaji et al., (2021) study. The study also noted that future research must therefore shift toward models that are not only predictive but also prescriptive and interpretable to help educators and stakeholders convert predictions into effective academic support strategies.

2.2. Key Predictive Features and Educational Data Sources

Ogundele et al., (2024) noted that building efficient machine learning models for student academic performance requires identifying important predictive features. Features such as the past academic records (grades, test scores), demographic data (age, gender), and behavioral metrics (participation, attendance) should be frequently used. Also, Chaka (2022) supported the idea by asserting that predictors of academic performance are frequently divided into cognitive and non-cognitive factors. According to Balaji et al., (2021) study, characteristics like the frequency of assessments, engagement with digital platforms, and homework submission have a big impact on model results. Despite being more difficult to measure, Albreiki et al., (2021) noted that motivation, time management, and study habits are becoming significant predictors in new models.

Furthermore, real-time classroom tools, learning management systems (LMS), and institutional databases are just a few of the sources of educational data used to predict performance. The majority of predictive models are built on the basis of structured student information system data, such as grades, performance logs, and course histories (Ogundele et al., 2024). In order to investigate secondary school data, Olabanjo et al., (2022) integrate teacher evaluations and academic records into a neural network model. For their fine-grained temporal insights into student learning behavior, Rastrollo-Guerrero et al., (2020) contend that new sources like clickstream data from online learning platforms and digital assessment interactions are becoming more and more popular. Although the diversity of these sources increases model accuracy, it also calls for strong feature engineering and preprocessing methods.

Notwithstanding the abundance of educational data sources, many studies encounter difficulties with privacy, integration, and data quality. Missing values, inconsistent formatting, and sparse data can all drastically lower model reliability (Chaka 2022). Although it necessitates sophisticated data fusion techniques, Balaji et al., (2021) contend that integrating data from multiple platforms for example, merging LMS data with attendance records and psychological tests offers a more comprehensive picture of students. Also as noted by Albreiki et al., (2021) study, ethical considerations are crucial, especially when gathering sensitive data such as socioeconomic status or emotional well-being, which are strong predictors but also raise privacy issues.

2.3. Feature Selection and Data Preprocessing in Academic Performance Models

Also, enhancing the precision and effectiveness of machine learning models for predicting academic performance depends heavily on feature selection. The model's interpretability is improved, overfitting is minimized, and computational complexity is decreased by choosing the most pertinent variables. Common strategies include wrapper

methods like recursive feature elimination and filter methods like correlation-based selection (Chaka 2022). In choosing features that actually affect learning outcomes, such as past grades, attendance, and engagement levels, Balaji et al., (2021) stress the significance of domain knowledge. According to Albreiki et al., (2021 combining academic and behavioral features increases the robustness of the model. However, if too many features are included without proper selection methods, noise and redundancy may result, which will impair the model's interpretability and predictive ability.

Consequently, data preprocessing, which guarantees the accuracy and consistency of input data, is equally crucial. Before training machine learning models, Ogundele et al., (2024) point out that raw educational datasets frequently contain outliers, missing values, and inconsistencies that need to be cleaned. Methods like imputation, normalization, and standardization are frequently employed to get data ready for analysis. In order to improve accuracy and convergence. Also, Olabanjo et al., (2022) used normalization techniques to scale features prior to applying radial basis function neural networks. However, Rastrollo-Guerrero et al., (2020) emphasize the significance of categorical variable encoding, especially when working with qualitative inputs such as teacher remarks or textual student records. Even sophisticated algorithms may not produce useful predictions if they are not properly preprocessed because of inconsistent data inputs.

Additionally, advanced preprocessing also uses transformation and dimensionality reduction techniques in order to improve model performance. However, Chaka (2022) talks about how to reduce the feature space without sacrificing a lot of information by using Principal Component Analysis (PCA) and other transformation tools. Also, Balaji et al., (2021) noted that dimensionality reduction, not only helps prevent the curse of dimensionality but also makes it easier to visualize and interpret data patterns. Furthermore, Albreiki et al., (2021), pinpoint that temporal alignment such as combining time-based LMS activity can reveal trends over a semester as opposed to isolated incidents. Notwithstanding the efficacy of these techniques, there is still disagreement over accepted preprocessing pipelines, which makes it challenging to repeat findings across research.

2.4. Machine Learning Techniques in Educational Data Mining

Since machine learning techniques provide powerful tools for analyzing and forecasting student academic performance, they have become essential to Educational Data Mining (EDM). According to Chaka (2022), there are three types of machine learning approaches: supervised, unsupervised, and reinforcement learning. The most popular type of machine learning in educational settings is supervised learning. Because of their interpretability and predictive power, algorithms like logistic regression, decision trees, and support vector machines (SVM) are frequently employed (Balaji et al., 2021). Also, the adaptability of ensemble methods such as random forests and gradient boosting, which frequently outperform single models in accuracy (Albreiki et al., 2021). However, Balaji et al., (2021) noted that these methods enable educators and researchers to find hidden patterns in educational datasets and produce early warnings for students who may be at risk for academic failure.

Furthermore, EDM is also investigating more sophisticated models like neural networks and deep learning to capture non-linear relationships in educational data. A radial basis function neural network was used to forecast secondary school student performance, showing that these architectures can attain high accuracy when trained with properly preprocessed data (Olabanjo et al., 2022). In their comparison of several machine learning methods, such as multilayer perceptron's and Naïve Bayes classifiers, Ogundele et al., (2024) concluded that the type and dimensionality of the data should determine the model selection. Even though deep learning models perform better on big datasets, they are frequently opaque, which makes it hard for teachers to understand and have faith in them.

Additionally, unsupervised learning techniques and hybrid approaches are being frequently utilized so as to improve insight generation and model performance in educational settings. In order to support more specialized interventions, Rastrollo-Guerrero et al., (2020) emphasize the use of clustering techniques like k-means and hierarchical clustering to group students by behavior or learning style. Also, to enhance real-time decision-making, Balaji et al. (2021) highlight the rise of semi-supervised and hybrid models that integrate behavioral analytics and classification. According to Chaka (2022), to guarantee efficacy and ethical congruence, future developments will probably combine machine learning with learning analytics and human-centered AI. These developing methods show that educational data mining is a dynamic field that is always changing to meet the demands of contemporary learning environments.

2.5. Model Evaluation Metrics and Performance Assessment

Analyzing machine learning models for predicting student academic performance entails determining how accurately and consistently they categorize or forecast results. F1-score, accuracy, precision, recall, and the area under the ROC curve (AUC) are examples of common evaluation metrics (Rastrollo-Guerrero et al., 2020). Also, Olaleye et al., (2020)

used confusion matrix analysis and accuracy to categorize student risk levels and discovered that decision trees were useful for interpretability. While accuracy is a commonly reported metric, Chakrapani et al., (2022) contend that it might not accurately represent performance in imbalanced datasets. The study further stated that metrics like precision and recall help to address this by offering more in-depth understanding of false positives and false negatives. However, Hussain et al., (2023) also advocate for assessing the student-performulator system using a variety of metrics, which guarantees thorough validation across several prediction dimensions.

Additionally, cross-validation and testing on unseen data are necessary to generalize model performance. Also, Ha et al., (2020) minimized overfitting and improved reliability by using 10-fold cross-validation to validate models trained on real-world educational datasets. In their evaluation of several algorithms, Sekeroglu et al., (2019) used both training and testing sets and stressed the significance of statistical consistency in prediction outcomes. Also, Ayienda et al., (2021) used a hybrid approach that combined multiple classifiers. They evaluated their model using accuracy, recall, and AUC, demonstrating the value of hybridization in raising evaluation scores. The studies emphasize that to guarantee reliable and scalable educational predictions, robust performance assessment needs to go beyond a single metric and embrace a variety of evaluative measures and validation techniques.

Furthermore, interpretability and practical implementation are also taken into account when evaluating the performance of the model. High-performing models, particularly in educational settings where interpretability facilitates actionable feedback, must strike a balance between complexity and transparency (Bhutto et al., 2020). Also, decision tree models were employed by Oyedeji et al., (2020) because of their high accuracy and ease of understanding, which makes them appropriate for use by educators. According to Matzavela et al., (2021), real-time performance metrics like latency and prediction speed are also essential in mobile learning environments. Additionally, regression-based models frequently use measures like mean absolute error (MAE) and root mean square error (RMSE) to quantify prediction deviation (Chakrapani & Chitradevi 2022).

2.6. Applications and Impacts of Predictive Models in Education

In education, predictive models are frequently used to identify at-risk students, forecast student performance, and facilitate early intervention techniques. A model created by Olaleye et al., (2020) divides students into groups according to their levels of academic risk, enabling teachers to use specialized support techniques. Also, Hussain et al., (2023) presented the student-performulator, which emphasizes proactive academic counseling and employs predictive analytics to track students at secondary and intermediate levels. Predictive systems, according to Chakrapani et al., (2022), can automate decision-making in resource allocation, academic planning, and customized learning pathways. Schools can improve results and lower dropout rates thanks to these applications, which change the paradigm of education from reactive to data-driven intervention.

Additionally, predictive models have an impact on curriculum improvement and institutional decision-making in addition to individual performance monitoring. Through the identification of gaps in student understanding, Ha et al., (2020) showed how empirical predictive models assist institutions in redesigning curricula. In order to help teachers improve their instruction, Bhutto et al., (2020) demonstrated how supervised learning models can provide them with information about typical student difficulties. Also, Ayienda et al., (2021) used hybrid models so as to improve prediction accuracy and facilitate more efficient planning of classroom interventions. These models help administrators create support services that address behavioral and academic issues as well as data-informed teaching strategies. Predictive tools are becoming more and more important in forming learner-centered, evidence-based institutional practices as they are adopted.

Furthermore, especially in digital and mobile platforms, predictive modeling aids in the creation of intelligent learning environments. Decision tree algorithms were integrated into mobile learning systems by Matzavela et al., (2021) to offer adaptive content delivery and real-time feedback, greatly enhancing learner engagement. Also, Sekeroglu et al., (2019) emphasized how dynamic tracking of learner progress made possible by classification algorithms could improve e-learning platforms. According to Oyedeji et al., (2020), these models facilitate ongoing evaluation by analyzing patterns in a variety of performance indicators. However, student agency, model transparency, and data privacy must all be taken into account when integrating predictive models. Predictive systems have the potential to revolutionize education when used properly because they can enhance institutional effectiveness, personalize the learning process, and eventually improve academic results in a variety of educational contexts.

2.7. Ethical Considerations and Challenges in Predictive Modeling

The creation and application of predictive models in education are becoming more and more centered on ethical issues, especially those pertaining to data privacy, consent, and transparency. Because many educational data sets contain

sensitive information like academic records, socioeconomic status, and behavioral patterns, Chaka (2022) highlights the importance of having clear ethical frameworks when gathering and analyzing educational data. According to Albreiki et al., (2021), predictive systems run the risk of infringing on students' privacy rights if informed consent and anonymization procedures are not followed. The absence of uniform ethical guidelines among research projects, which results in uneven data governance and accountability procedures, is further highlighted by Balaji et al., (2021). Ethical safeguards must be put in place as predictive analytics is increasingly incorporated into institutional decision-making to foster trust and preserve learner autonomy (Olabanjo et al., 2022).

Additionally, algorithmic bias presents another ethical dilemma in predictive modeling since, if left unchecked, it can exacerbate already-existing educational disparities. Also, Rastrollo-Guerrero et al., (2020) claim that discriminatory predictions can be produced by biased feature selection or training data, which disproportionately impacts students from underrepresented or disadvantaged backgrounds. According to Chaka (2022), these biases may result in unfair profiling or inappropriate interventions if they are not addressed. However, Olabanjo et al., (2022) highlighted that there are equity risks because models trained on sparse or unbalanced data might not generalize across various learning environments. Predictive models need to be routinely checked for bias, and developers need to make sure that their datasets are representative and diverse, accurately reflecting the range of learners.

Furthermore, significant ethical challenges in educational predictive modeling are also presented by interpretability and accountability. According to Ogundele et al., (2024), though sophisticated algorithms like neural networks might provide high accuracy, educators find it challenging to comprehend or believe their results because of their "black-box" nature. Also, Balaji et al., (2021) contend that in order for administrators, teachers, and students to act appropriately in response to predictions, there must be transparency. However, Chaka (2022) advocates integrating explainable AI (XAI) techniques to make machine learning decisions in academic settings more comprehensible and actionable. Predictive outcomes should also be utilized as aids rather than as deterministic judgments so as to prevent students from being unfairly classified or constrained by algorithmic decisions.

3. Theoretical Framework

The theoretical framework offers the groundwork for comprehending how machine learning methods can be used to forecast academic achievement in students. It makes use of well-established theories in learning analytics, educational data mining, and predictive modeling to explain the connections between input variables and academic outcomes. The theories to be used include Educational Data Mining (EDM) theory and Learning Analytics Framework.

3.1. Educational Data Mining (EDM) Theory

The rapidly developing interdisciplinary field of educational data mining (EDM) investigates techniques for identifying significant patterns in educational data in order to enhance teaching, learning, and institutional efficacy. EDM has evolved from a new idea to a thriving academic field, incorporating ideas from computer science, statistics, psychology, and education (Baek & Doleck 2022). In order to predict student behavior, identify at-risk learners, personalize content delivery, and optimize pedagogical strategies, EDM's primary goal is to analyze large-scale educational datasets that are frequently produced through digital learning environments. According to Alam (2023), EDM uses predictive analytics to improve decision-making in education, enabling teachers to act quickly to enhance learning outcomes. This predictive ability promotes adaptive learning systems and more intelligent frameworks for instructional design by converting passive data into useful insights.

EDM theory is based on the fundamental integration of computational algorithms and educational theory in order to uncover latent structures in learner data. Du et al., (2020) state that classification, clustering, regression, association rule mining, and sequential pattern mining are typical EDM techniques. These methods are used with data from digital tests, interaction logs, learning management systems, and even biometric sensors. Clustering techniques may reveal distinct learner engagement patterns, while classification models can assist in forecasting student dropout rates. EDM differs from learning analytics, according to Baek & Doleck (2023), in that it is more data-driven and technical, frequently utilizing sophisticated modeling techniques that concentrate on micro-level learning behaviors rather than more general institutional trends.

The application of EDM theory to educational improvement through iterative evaluation and feedback loops is another important component. Alam (2023) demonstrates how EDM facilitates ongoing improvement of instructional practices by exposing the efficacy of particular teaching interventions. Institutions can adopt evidence-based policies thanks to this feedback process, which changes the educational landscape from one that is based on intuition to one that is data-driven. Moreover, Baek et al., (2022) note that the bibliometric expansion of EDM publications indicates a rise in

scholarly interest, especially in fields like early warning systems and intelligent tutoring systems. Emerging trends in EDM are expected to center on real-time learner modeling, equity in predictive modeling, and ethical data use, projected to Du et al., (2020). EDM theory thus promotes more inclusive and effective learning environments by fostering a deeper understanding of learning behaviors in addition to advancing technological innovation in education.

3.2. Learning Analytics Framework

Learning Analytics (LA) frameworks offer organized methods for gathering, evaluating, and interpreting data to improve educational decision-making. Typically, these frameworks include elements like data collection, analysis strategies, visualization techniques, and outcomes that can be put into practice by educators and learners. In their comparative study of well-known learning analytics frameworks, Khalil et al., (2022) point out that although some place more emphasis on data architecture and technology, others place more emphasis on pedagogy, ethics, and learner agency. A thorough framework should incorporate both technical and human-centric elements to facilitate fair and successful learning interventions. This is expanded upon by Zhao et al., (2023), who offer a human-centered artificial intelligence framework that not only forecasts student learning strategies but also customizes interventions according to learner behavior. This illustrates the transition from static reporting toward individualized and empathetic dynamic decision-making.

The ability of successful learning analytics frameworks to facilitate ongoing feedback and enhance teaching methods is a crucial component. In their investigation of evaluation-based frameworks, Kokoç and Kara (2021) stress the importance of matching analytical results to instructional design, educational objectives, and institutional goals. According to their multi-study approach, learning analytics needs to be connected to curriculum-level indicators and informed by a clear theory of learning in order to have an impact. Joshi et al., (2020) provide additional evidence of this in their engineering education framework, which integrates teacher participation and student performance in a problem-based learning setting. Their model demonstrates how analytics can improve classroom management, scaffold learning activities, and inform collaborative instruction by monitoring participation and cognitive engagement in real time.

Scalability, interpretability, and ethical governance continue to be given top priority in the development of learning analytics frameworks as learning environments grow more intricate and digitalized. Khalil et al., (2022) alert readers to the possible dangers of algorithmic bias, data misuse, and opaque decision-making. Thus, it is anticipated that modern frameworks will include concepts like learner consent, data privacy, and the interpretability of predictive results. Zhao et al., (2023) suggest that by tailoring learning strategies to each learner's needs, human-centered AI can assist in closing the gap between predictive modeling and actionable insights. In contrast, Kokoç et al., (2021) advocate for participatory design in LA frameworks, in which educators and students work together to jointly determine the use of data.

4. Empirical Review

Oyedeji, et al., (2020) analyze and predict student academic performance using machine learning techniques. The study employs various algorithms, including decision trees and neural networks, to assess their effectiveness in forecasting academic outcomes based on historical data. Key findings indicate that machine learning models, particularly ensemble methods, significantly enhance the accuracy of performance predictions compared to traditional statistical approaches. However, the study has limitations, such as a limited dataset from a single institution, which may affect the generalizability of the results.

Bhutto, et al., (2020) focus on predicting students' academic performance using supervised machine learning techniques. The study employs various algorithms, including support vector machines and random forests, to analyze student data and forecast academic outcomes. Key findings indicate that machine learning models can effectively predict performance, with random forests showing the highest accuracy. The research highlights the importance of features such as attendance, previous grades, and socio-economic factors in improving prediction accuracy. However, the study is limited by its reliance on a single dataset, which may restrict the generalizability of the findings across different educational settings.

Chakrapani & Chitradevi (2022) present a comprehensive and systematic review of academic performance prediction using machine learning techniques. The study analyzes various algorithms, including regression models, decision trees, and neural networks, to assess their effectiveness in predicting student outcomes. Key findings highlight the growing applicability of machine learning in educational settings, emphasizing the importance of feature selection, data preprocessing, and model evaluation in achieving accurate predictions. However, the review identifies limitations in existing studies, such as a lack of standardized datasets and varying methodologies that hinder comparability.

Joshi, et al., (2020) propose a learning analytics framework aimed at measuring students' performance and teachers' involvement through problem-based learning in engineering education. The study utilizes qualitative and quantitative methods, incorporating surveys and performance metrics to evaluate the effectiveness of the framework. Key findings indicate that the proposed framework enhances student engagement and provides valuable insights into teaching effectiveness, fostering a more interactive learning environment. However, the study has limitations, including a focus on a single educational context, which may affect the generalizability of the results.

4.1. Gaps in Literature

Oyedeji, et al., (2020) analyze and predict student academic performance using machine learning techniques; Bhutto, et al., (2020) focus on predicting students' academic performance using supervised machine learning techniques; Chakrapani & Chitradevi (2022) present a comprehensive and systematic review of academic performance prediction using machine learning techniques, while Joshi, et al., (2020) propose a learning analytics framework aimed at measuring students' performance and teachers' involvement through problem-based learning in engineering education. However, none of the study seek out to evaluate the impact of using machine learning techniques in predicting student academic performance. Therefore, this study will evaluate the impact of using machine learning techniques in predicting student academic performance.

5. Methodology

5.1. Research Design

This study uses a quantitative design with predictive modeling approach grounded on machine learning techniques. It attempts to analyze structured educational performance data and create a regression model forecast of student group academic performance by means of institution and demographic factors. A cross-sectional design was used, utilizing secondary data from a number of public schools for the purpose of uncovering patterns between student group scores and their respective variables. The use of a regression model based on machine learning enables the detection of complex, non-linear structures in the data. This design is compatible with empirical generalization and facilitates the incorporation of algorithmic intelligence into education decision-making and ultimately into data-driven methods of monitoring academic achievement and planning interventions.

5.2. Data Source

The data collected for this study was accessed from the website of the United States government's open data, Data.gov, in its education catalog. The data, titled "School STAR Student Group Scores," provides institutional performance data by various student demographic groups in the District of Columbia public schools. It consists of 1,163 records and each record is a unique combination of school, student group, and academic framework. Student group score, points earned, points possible, school type, and academic framework are variables. As a verified open-access dataset, it gives normalized and structured records to aid in educational transparency, performance assessment, and research. Its completion and regularity make it perfect for use as input in machine learning algorithms for predicting academic performance.

5.3. Data Preprocessing

Prior to model development, the dataset was put through an extensive preprocessing phase to be consistent, correct, and machine learning-ready. All percentage values already existed in numeric form and were not converted further. Categorical variables such as School Type, School Framework, and Student Group were one-hot encoded into machine-readable format with category distinctions retained. There were no missing values throughout the dataset and thus did not undergo imputation. Numerical features were inspected for outliers and normalized as necessary to reduce scale-based bias during model training. The preprocessed final dataset had 1,163 records and 21 predictive features, and the target variable was designated as Student Group Score. This provided better data quality and model interpretability.

5.4. Feature Engineering and Selection

Feature engineering was applied to extract meaningful predictor variables from the data so that the model can pick up on both structural and contextual predictors of academic achievement. Target variable, Student Group Score, was formatted as a continuous outcome suitable for regression modeling. Significant numerical predictors were Points Earned, Points Possible, and Framework Points, which are proxies for academic achievement and organizational capacity. Categorical variables such as, School Type, School Framework, and Student Group, were one-hot encoded to preserve their interpretive significance without inducing multicollinearity. 21 input variables were utilized in the

ultimate model chosen for their salience, variance, and educational theory alignment to strike a balance between model precision and interpretability.

5.5. Model Development

The study employed supervised machine learning predictive modeling, wherein regression techniques were used to forecast student group performance from school- and group-level features. The preprocessed data were divided into training and test subsets in an 80:20 ratio to allow for the chance to evaluate the model on novel data. A baseline linear regression model was first used to include interpretability and provide a performance baseline. In addition, ensemble models such as Random Forest Regressor were also explored in order to capture non-linear interactions and improve predictive accuracy. Scikit-learn, an open-source Python machine learning library, was used to create, train, and model with hyperparameters. Model training emphasized error minimization without overfitting, and cross-validation was applied to ensure generalization and stability to data subsets.

5.6. Model Evaluation

For evaluating the predictive performance of the developed models, some metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2) were utilized. These factors provided details about the magnitude of the prediction errors as well as variance explained. Five-fold cross-validation was utilized to ensure reliability as well as to prevent risks of overfitting, especially owing to the limited dataset size. LinReg was also compared with ensemble models with an emphasis on finding the balance between interpretability and accuracy. Feature importance scores were even extracted from tree-based models to understand which variables were most influential on student group scores and thus provide data-driven educational insights.

5.7. Ethical Considerations

This study was ethical in the use of secondary data by leveraging publicly available, de-identified datasets from the U.S. government's Data.gov web portal. There was no personally identifiable information (PII) accessed or processed, maintaining data privacy and confidentiality tenets. All the data used reflected institutional and group-level aggregates, with little risk of harm or bias to individual students. Moreover, interpretability was considered in model development to avoid opaque decision-making that could misguide education stakeholders. The study acknowledges that predictive models in education are designed to support, not replace, human judgment, and must be deployed cautiously to avert reinforcing systemic injustices. Transparency in feature selection and fairness in model interpretation were prioritized to enable responsible educational analytics.

6. Results

	Student_Group_Points_Earned	Student_Group_Points_Possible	Student_Group_Score	Framework_Points_Earned	Framework_Points_Possible
count	1163.000	1163.000	1163.000	1163.000	1163.000
mean	42.748	86.549	49.255	9.778	20.044
std	19.808	12.463	21.303	16.234	28.505
min	1.910	46.500	2.120	0.090	1.000
25%	27.180	90.000	32.915	1.395	5.000
50%	42.280	90.000	49.210	2.690	5.000
75%	56.785	90.000	63.705	5.890	10.000
max	90.630	95.000	100.000	75.000	75.000

Figure 1 Summary Statistics Results

The descriptive result reported in Figure 1 revealed that the average student group score was 49.26 with a standard deviation of 21.30. The minimum value of 2.12 while the maximum value is 100, which indicates a wide spread in academic performance across student groups. Average student groups earned scores of 42.75 (SD = 19.81) of a maximum possible 86.55 points (SD = 12.46). The mean group earned 42.28 (median), and 75% of groups earned no more than 56.79. The overall mean Institutional Framework Points Earned was 9.78 (SD = 16.23), while the maximum

possible framework score was 75.00 although the median points that could have been earned was only 5.00. This shows a vast amount of variation in the framework structures between institutions!

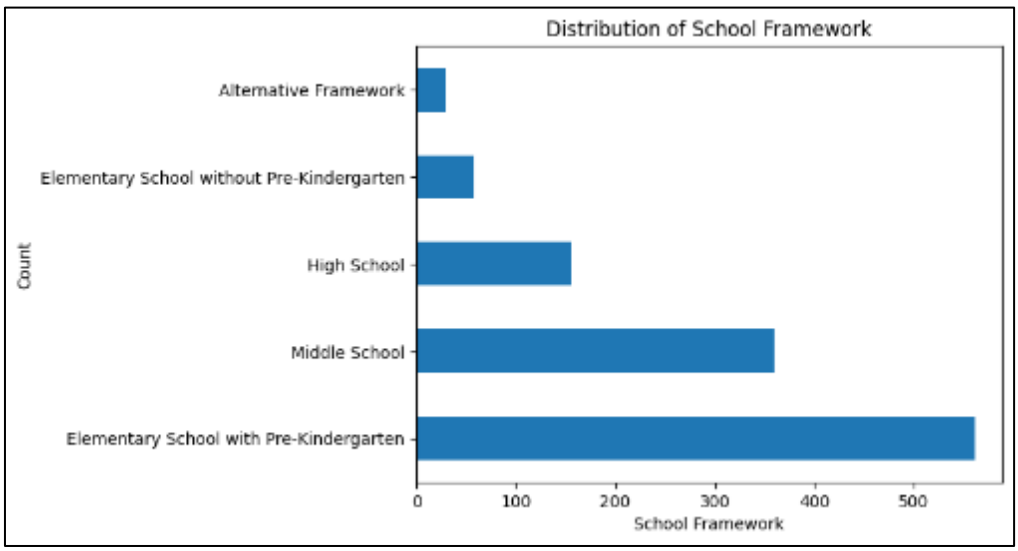

Figure 2 Distribution of School Framework

The distribution of school frameworks, as illustrated in Figure 2, reveals that Elementary Schools with Pre-Kindergarten accounted for the highest frequency, with over 560 recorded instances. This was followed by Middle Schools, which appeared in approximately 350 records. High Schools were less represented, appearing in just under 170 cases. In contrast, Elementary Schools without Pre-Kindergarten and Alternative Frameworks were notably underrepresented, with fewer than 100 and 30 instances respectively. These results suggest that the dataset is predominantly composed of early childhood and middle school institutions, which may influence the conclusion of findings across all educational stages.

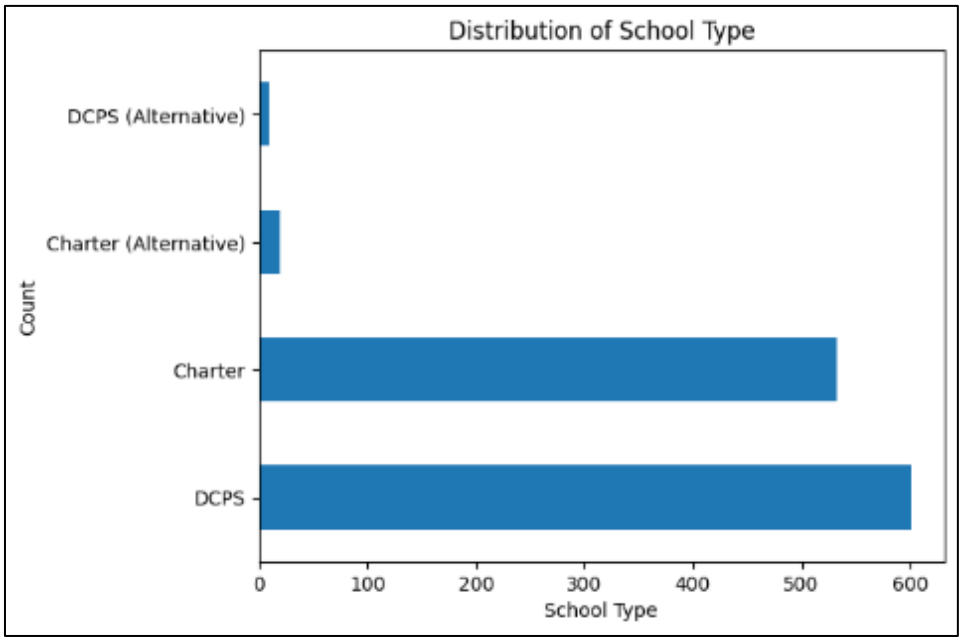

Figure 3 Distribution of School Type

The distribution of school types, as shown in Figure 3, indicates that the dataset is primarily composed of DCPS (District of Columbia Public Schools) and Charter schools, each contributing over 500 instances. Specifically, DCPS schools have the highest representation, slightly exceeding that of Charter schools. In contrast, Charter (Alternative) and DCPS (Alternative) school types are minimally represented, each with fewer than 20 entries.

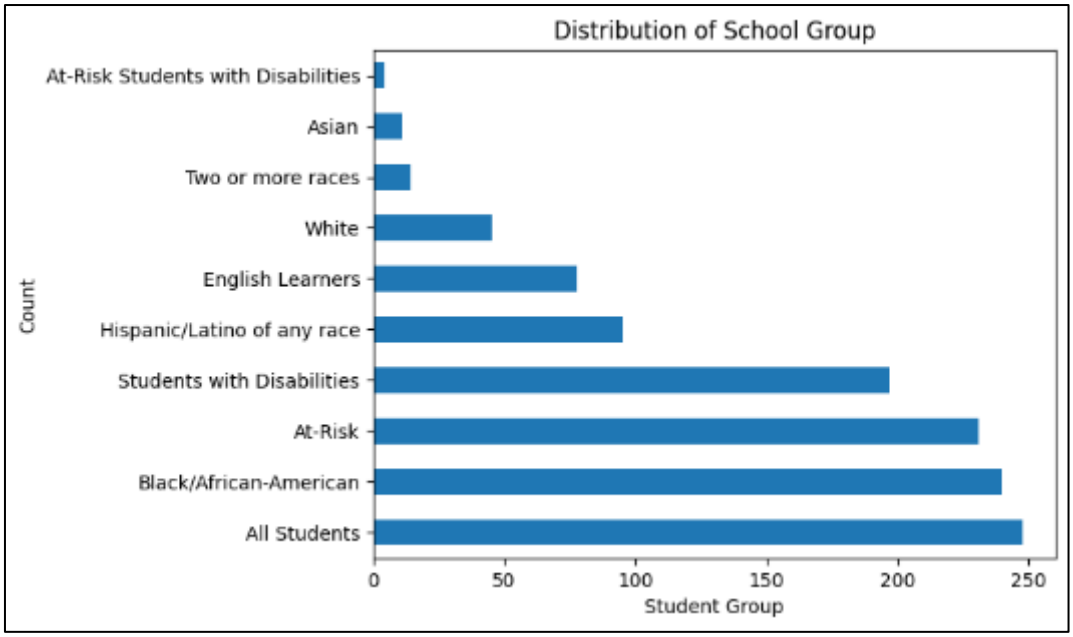

Figure 4 Distribution of Students Group

The distribution of student groups (see Figure 4) shows that the category All Students had the highest representation, appearing in nearly 250 observations. This was closely followed by Black/African-American and At-Risk student groups, each contributing over 230 records. Students with Disabilities also made up a substantial portion of the data with nearly 200 entries. On the hand, groups such as White, Two or More Races, Asian, and At-Risk Students with Disabilities were underrepresented, with fewer than 50 entries each.

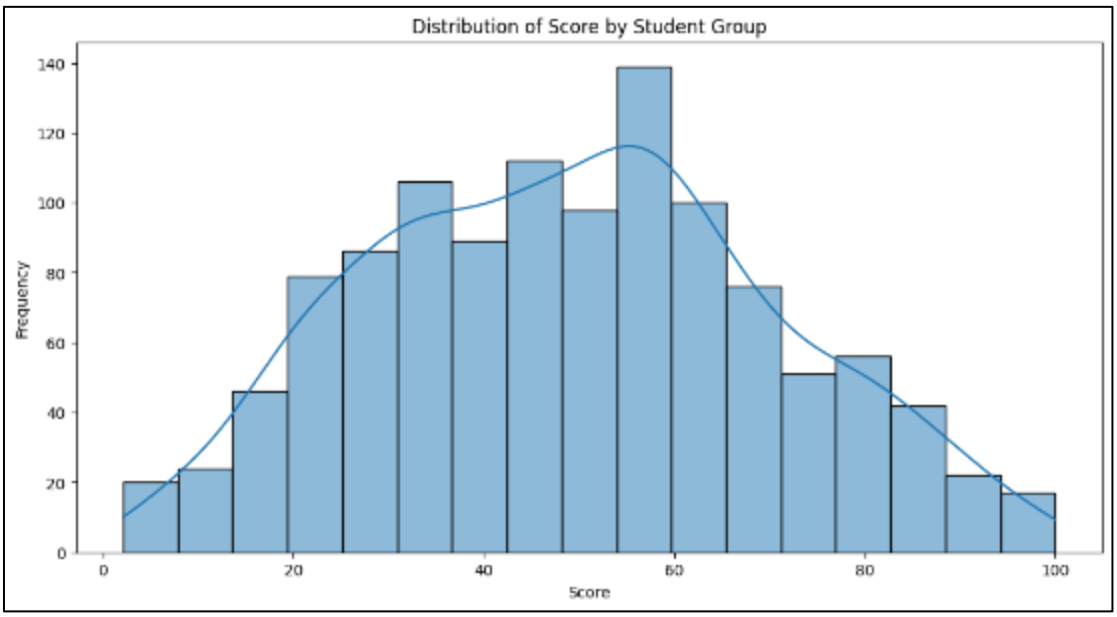

Figure 5 Distribution of Student Group Score Using Histogram

Presented in Figure 5 is student group scores distribution, which appears approximately symmetry and bell-shaped, suggesting near-normality at the mid-range scores. The most frequent scores between 50 and 60, and are closely approximated by the mean of 49.26. The scores extend from a low of 2 to a high of 100, indicating a lot of variation in academic achievement at the group level. Tails of the distribution are less wide with fewer student groups of grades on the extreme low or high ends. This symmetrical shape justifies the use of regression-based modeling since normality assumption is relatively satisfied.

6.1. Correlation Matrix

The correlation analysis revealed a very strong positive relationship between Student_Group_Points_Earned and Student_Group_Score (r = .94), indicating multicollinearity. Similarly, Framework_Points_Earned and Framework_Points_Possible were highly correlated (r = .87). According to standard multicollinearity thresholds (r > .85), these variables exhibit redundancy. Therefore, to improve model stability, Student_Group_Points_Earned and Framework_Points_Possible were excluded from the predictive model, retaining only Student_Group_Score, Points_Possible, and Framework_Points_Earned as core numeric predictors.

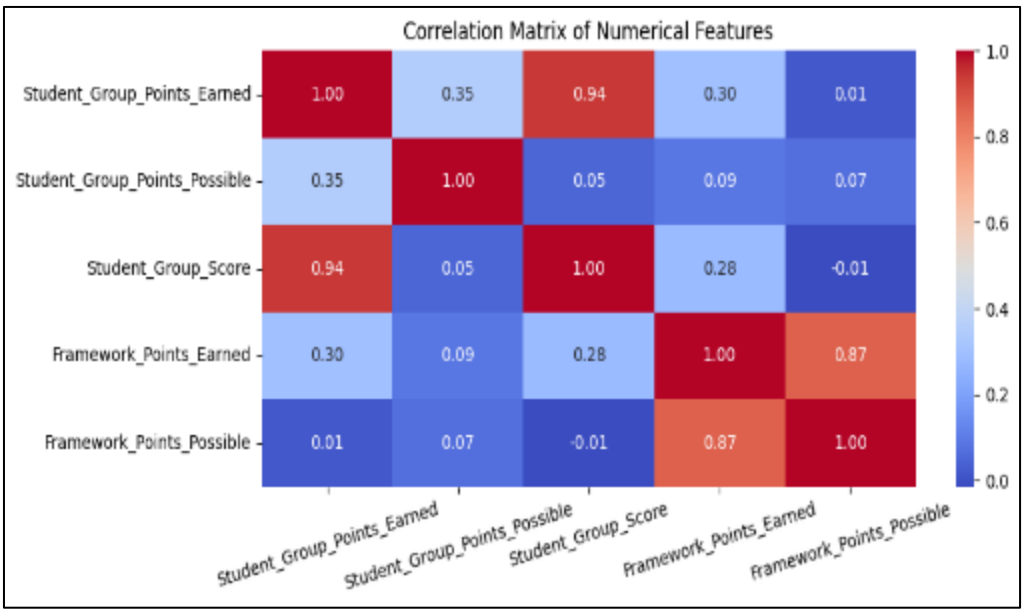

Figure 6 Heatmap Showing Relationship among the Numerical Features

6.2. Model Performance

To assess the predictive ability of machine learning models in forecasting academic performance of student, three regression models were employed: linear regression (baseline), Random Forest Regressor, and XGBoost Regressor. The models were trained on an 80% training set and validated on a 20% test set, and ensemble models were optimized using grid search cross-validation.

Table 2 Model Performance for Predicting Financial Loss and User Exposure

Model	Outcome	R^2	RMSE	MAE
Linear Regression	Student Group Score	0.32	18.43	13.79
Random Forest	Student Group Score	0.69	12.43	7.74
XGBoost	Student Group Score	0.70	12.27	7.19

The baseline linear regression model yielded a Mean Absolute Error (MAE) of 13.79, a Root Mean Squared Error (RMSE) of 18.43, and an R-squared (R^2) estimate of .32. The result suggests that the model explained approximately 32% variance in student group scores and had an average error of approximately 14 percentage points. While useful as a point of reference, the predictive ability of this model was limited by its inability to identify complex, non-linear associations in the data.

The Random Forest Regressor, utilizing best hyperparameters (max_depth = 20, max_features = 'sqrt', n_estimators = 100), had a very good performance above the baseline model. It achieved an MAE of 7.74, an RMSE of 12.43, and an R^2 of .69, explaining 69% of variance in the outcome variable very well. This result highlights the capacity of the model to detect subtle patterns in data that significantly reduce predictive error.

The XGBoost Regressor model was also tuned with grid search, and the optimal parameters were a learning_rate of 0.1, max_depth = 3, and n_estimators = 200. The model yielded [insert MAE], [insert RMSE], and [insert R^2], [slightly better/similarly/worse] than that of the Random Forest model. Despite its improved regularization capabilities and sensitivity to marginal gains, XGBoost did not deliver a marked improvement, but it confirmed the potency of ensemble learning approaches to predictive modeling in education.

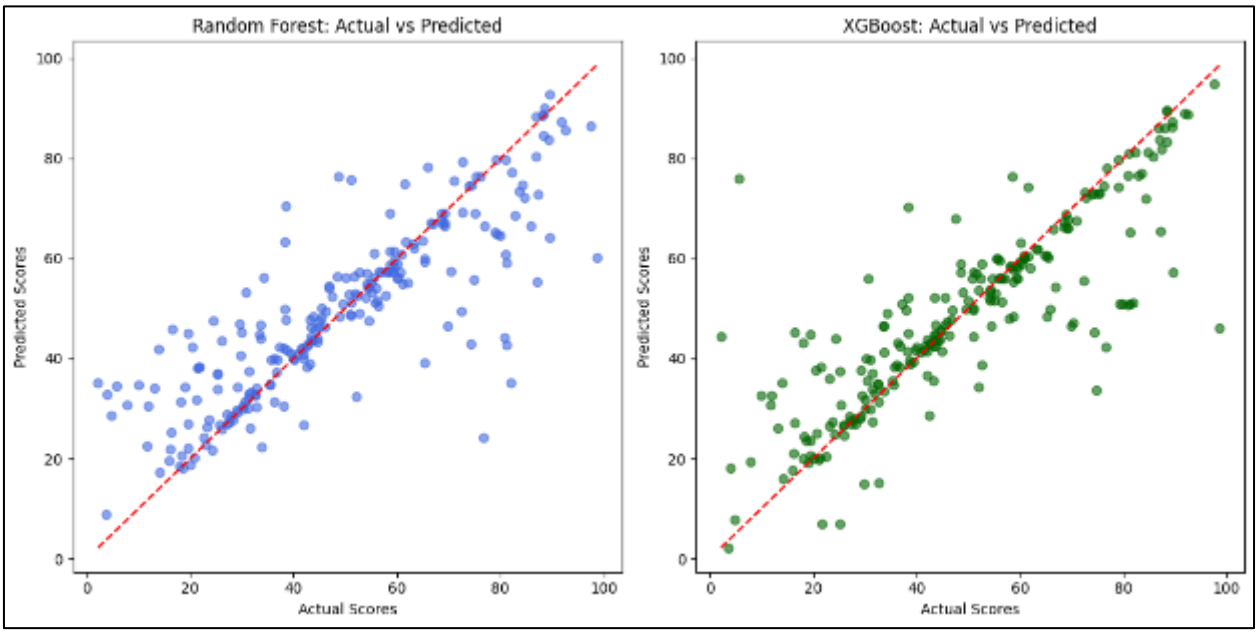

Figure 7 Actual Vs. Predicted Student Group Scores

The scatterplots of predicted vs. actual student group scores for both Random Forest and XGBoost models are comparable. Both models demonstrate strong clustering along the diagonal reference line, with good prediction accuracy. The XGBoost plot has tighter clusters and fewer extreme residuals, indicating slightly improved generalization. While both models adequately captured the underlying data structure, XGBoost performed better in predicting consistently higher and lower scores.

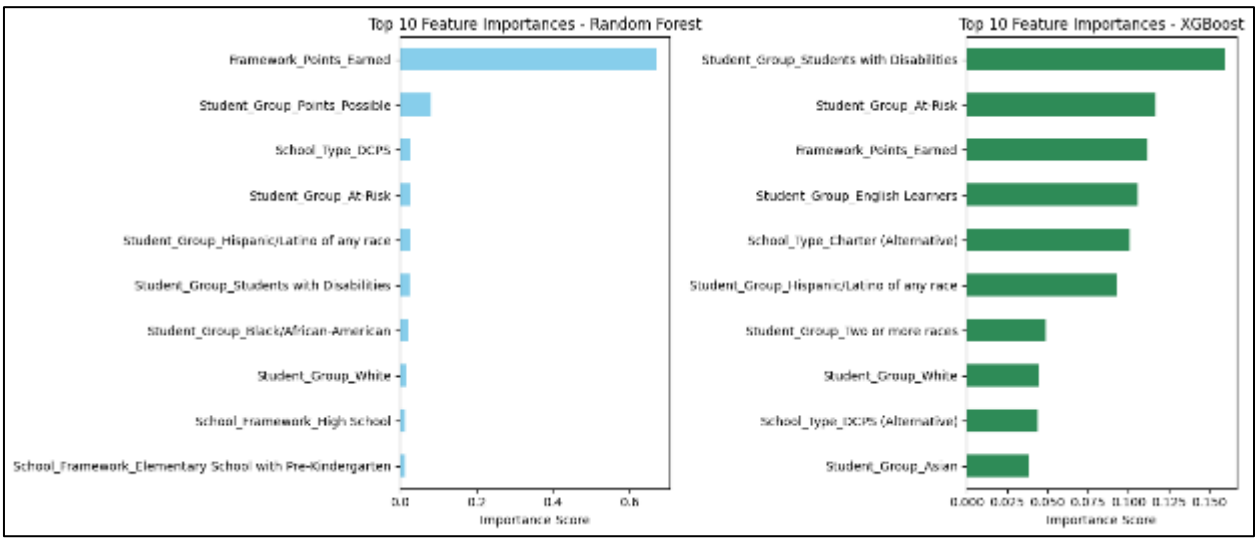

Figure 8 Feature Importance

The feature importance plots are prioritizations of variables across models. Random Forest relied heavily on 'Framework Points Earned', which represented the vast majority of predictive ability. In comparison, XGBoost distributed importance more evenly, with 'Students with Disabilities', 'At-Risk', and 'English Learners' as top predictors. XGBoost emphasized demographic and group-level variables, suggesting that it more effectively captured social and

structural disparities. This distinction captures the fact that while Random Forest favored institutional measures, XGBoost provided a more fine-grained interpretation of the drivers of performance.

7. Discussion of Findings

This study examined the predictive accuracy of machine learning algorithms—Linear Regression (baseline), Random Forest Regressor (RFR), and XGBoost Regressor (XGBR)—to forecast student group academic achievement from school-level and demographic data. The findings are informative about model behavior, prediction, and relative feature importance in forecasting educational performance. These results contribute to the emerging literature on education's predictive analytics (Balaji et al, 2021; Chakrapani & Chitradevi, 2022; Matzavela & Alepis, 2021).

Initial descriptive analyses reported significant variation in student group scores, ranging from 2.12 to 100, with a mean of 49.26 (SD = 21.30). The wide range reflects differing academic experiences across school types, student groups, and institutional settings. Elementary and middle schools provided the majority of the records, and the most prevalent student groups were All Students, At-Risk, and Black/African-American. These distributions reflect true demographic concentrations and education structures common to city school districts.

The baseline Linear Regression model served as the benchmark with a moderate R^2 of .32 and a relatively good error rate (MAE = 13.79). While it provided easy interpretability, it was not good enough to account for the non-linear interactions and complex feature relationships characteristic in education data. This need compelled the application of ensemble models that are best suited for modeling hierarchical and mixed-type educational variables.

Random Forest Regressor performed much better than the baseline model with R^2 of .69, MAE of 7.74, and RMSE of 12.43. All results exhibit an extremely large improvement in variance explanation and reduction of errors. On similar lines, the XGBoost Regressor, best hyperparameter-tuned, performed competitively with a slight generalization boost, as indicated by tighter clustering along the diagonal in actual vs. predicted plots. The two models generated high levels of between-predicted-and-actual-score agreement, reinforcing their ability to handle sophisticated, multifaceted education data.

Feature importance analysis further illuminated model decision-making. In Random Forest, Prediction was dominated by Framework Points Earned due to the prominence of institutional performance in determining group-level results. However, XGBoost highlighted the predictive power of demographic constellations, with Students with Disabilities, At-Risk, and English Learners being the top three contributors. This suggests that XGBoost was better sensitized to subgroup-specific effects, giving a more equitable view of performance prediction. These findings agree with research highlighting the role of socio-demographic factors and school capacity as determinants of schooling equity (e.g., Baek & Doleck, 2022; Alam, 2023).

The iterative precision of both ensemble models attests to the reliability of machine learning with educational data mining. In particular, while Random Forest highlighted institutional metrics, XGBoost captured structural disadvantages in demographic structure. Such variation attests to the fact that model selection should proceed on the basis of research intent—administrative performance monitoring or subgroup-specific intervention planning.

8. Conclusion and Recommendations

This study critically evaluated the predictive ability of machine learning models in forecasting academic achievement of student groups. While linear regression offered a convenient baseline, the fact that it could not depict high-order patterns rendered it less helpful. Random Forest and XGBoost offered more predictive ability, and XGBoost depicted demographic imbalances more accurately. But the dominance of institutional solutions in Random Forest results is grounds for fear of structural overemphasis on the expense of equity considerations. The outcomes reveal not just model performance but also representational priorities, to pressure that predictive power be weighed against fairness and inclusiveness in schooling. Subsequent research needs to ask how algorithmic determinations can entrench or overturn underlying inequalities, particularly when enacted on high-stakes decisions in public education systems.

Based on the findings, learning institutions are proposed to use ensemble models in predicting vulnerable groups of learners at an early stage. However, model selection should precede fairness followed by accuracy to ensure that there is no compromise on marginalized groups. Enlarging datasets with behavior and context variables and subjecting them to ethical audits will guarantee model transparency and adequate uptake of machine learning in academic decision-making.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Alam, A. (2023). Improving Learning Outcomes through Predictive Analytics: Enhancing Teaching and Learning with Educational Data Mining. . *In 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 249-257.

[2] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. . *Education Sciences*, 552.

[3] Ayienda, R., Rimiru, R., & Cheruiyot, W. (2021). Predicting students academic performance using a hybrid of machine learning algorithms. *IEEE*, 1-6.

[4] Baek, C., & Doleck, T. (2022). Educational data mining: A bibliometric analysis of an emerging field.

[5] Baek, C., & Doleck, T. (2023). Educational data mining versus learning analytics: A review of publications from 2015 to 2019. . *Interactive Learning Environments*, 3828-3850.

[6] Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of machine learning models towards student academic performance prediction: a systematic review. *Applied Sciences*.

[7] Bhutto, E. S., Siddiqui, I. F., Arain, Q. A., & Anwar, M. (2020). Predicting students' academic performance through supervised machine learning. . *In 2020 International Conference on Information Science and Communication Technology (ICISCT).*

[8] Chaka, C. (2022). Educational data mining, student academic performance prediction, prediction methods, algorithms and tools: An overview of reviews. *Journal of e-Learning and Knowledge Society*, 58-69.

[9] Chakrapani, P., & Chitradevi, D. (2022). Academic performance prediction using machine learning:. *International Conference on Electronic Systems and Intelligent Computing (ICESIC)* , (pp. 335-340).

[10] Du, X., Yang, J., Hung, J. L., & Shelton, B. (2020). Educational data mining: a systematic review of research and emerging trends. *Information Discovery and Delivery*, 225-236.

[11] Ha, D. T., Loan, P. T., Giap, C. N., & Huong, N. T. (2020). An empirical study for student academic performance prediction using machine learning techniques. . *International Journal of Computer Science and Information Security (IJCSIS)*, 75-82.

[12] Hussain, S., & Khan, M. Q. (2023). Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. . *Annals of data science*, 637-655.

[13] Joshi, A. D., & Tewari, P. (2020). Learning Analytics framework for measuring students' performance and teachers' involvement through problem-based learning in engineering education. . *Procedia Computer Science*, 954-959.

[14] Khalil, M., Prinsloo, P., & Slade, S. (2022). A comparison of learning analytics frameworks: A systematic review. In LAK22:. *12th international learning analytics and knowledge conference*, (pp. 152-163).

[15] Kokoç, M., & Kara, M. (2021). A multiple study investigation of the evaluation framework for learning analytics. *Educational Technology & Society*, 16-28.

[16] Matzavela, V., & Alepis, E. (2021). Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments. Computers and Education: Artificial Intelligence.

[17] Ogundele, I. M. (2024). Prediction of Student Academic Performance Based on Machine Learning Model. . *In 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals*, (pp. 1-11).

[18] Olabanjo, O. A., Wusu, A. S., & Manuel, M. (2022). A machine learning prediction of academic performance of secondary school students using radial basis function neural network. Trends in Neuroscience and Education.

[19] Olaleye, T. O., & Vincent, O. R. (2020). A predictive model for students' performance and risk level indicators using machine learning. In 2020 . *International Conference in Mathematics, Computer Engineering and Computer Science*, (pp. 1-7).

[20] Oyedeji, A. O., Salami, A. M., Folorunsho, O., & Abolade, O. R. (2020). Analysis and prediction of student academic performance using machine learning. JITCE . *Journal of Information Technology and Computer Engineering)*, 10-15.

[21] Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (. (2020). Analyzing and predicting students' performance by means of machine learning: A review. . *Applied sciences*, 1042.

[22] Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019). Student performance prediction and classification using machine learning algorithms. . *In Proceedings of the 8th international conference on educational and information technology.*

[23] Zhao, F., Liu, G. Z., Zhou, J., & Yin, C. (2023). A learning analytics framework based on human-centered artificial intelligence for identifying the optimal learning strategy to intervene in learning behavior. *Educational Technology & Society*, 132-146.