Predicting Student Achievement Using Socioeconomic and School-Level Factors

Thosporn Sangsawang[1,*], Liu Yang[2]

[1]Educational Technology and Communications Division, Faculty of Technical Education, Rajamangala University of Technology Thanyaburi, Thailand

[2]Vocational Education Division, Faculty of Technical Education, Rajamangala University of Technology Thanyaburi, Thailand

ABSTRACT

This study compares the performance of two machine learning models, Random Forest (RF) and XGBoost, in predicting student achievement based on socioeconomic and school-level factors. Both models demonstrated exceptional performance, with XGBoost slightly outperforming RF across key metrics, including accuracy, precision, and recall. The analysis revealed that socioeconomic factors, such as family income and parental education levels, as well as school characteristics, were the most significant predictors of student success. These findings align with the broader literature, reinforcing the influence of external factors on educational outcomes. The implications of this study suggest that educational policies should focus on addressing socioeconomic disparities to improve student performance. Schools serving disadvantaged communities would benefit from increased access to resources, academic support, and parental engagement programs. The importance of these factors in shaping student outcomes points to the need for equitable resource distribution, with targeted interventions aimed at closing the achievement gap. For future research, the application of deep learning models and hybrid approaches could enhance predictive accuracy and provide further insights into student performance. Additionally, incorporating more granular data at the student level, such as individual academic progress and behavioral metrics, may improve model precision. Exploring other datasets with varying socioeconomic contexts could also extend the generalizability of these findings. In conclusion, while RF and XGBoost both performed well, the findings underscore the need for continuous improvements in model development. Integrating more advanced features, such as longitudinal tracking of socioeconomic variables, would offer a more dynamic understanding of how these factors evolve and impact educational success. These insights could guide more effective interventions, ultimately fostering a more equitable educational environment.

Keywords Student Achievements, Socioeconomic Factors, Machine Learning, Random Forest (RF), Xgboost

Submitted 18 January 2025
Accepted 22 February 2025
Published 3 March 2025

Corresponding author
Thosporn Sangsawang,
sthosporn@rmutt.ac.th

Additional Information and Declarations can be found on page 32

Introduction

Student achievement is a critical measure of the success and quality of educational systems worldwide. It is a barometer for the effectiveness of instructional methodologies and the overall school learning environment. Governments, educational institutions, and communities rely on student achievement data to shape policies, allocate resources, and assess the impact of curriculum standards. The importance of student success extends beyond individual learning outcomes, influencing national educational standards and contributing to societal development. Consequently, identifying the factors contributing to or hindering academic performance remains a central focus of educational research.

How to cite this article: T. Sangsawang and L. Yang., "Predicting Student Achievement Using Socioeconomic and School-Level Factors ," *Artif. Intell. Learn.*, vol. 1, no. 1, pp. 20-34, 2025.

A growing body of research highlights that student achievement is not solely determined by innate abilities or the quality of instruction. Socioeconomic factors, school-level characteristics, and parental involvement also play crucial roles in shaping student outcomes. Students from lower socioeconomic backgrounds often face barriers that can impede academic success, including limited access to resources and support systems. Similarly, the school environment profoundly affects students' academic progress, ranging from teacher quality to available learning resources. Understanding their individual and collective impact on student achievement has become an essential pursuit in educational data mining as these factors intertwine. This paper focuses on predicting student performance by analyzing both socioeconomic and school-level variables using machine learning models.

The application of data mining techniques in educational environments has expanded rapidly, driven by the desire to improve student outcomes and optimize instructional processes. Educational Data Mining (EDM) uses data mining tools to analyze vast datasets generated by student activities and institutional operations to uncover patterns that can predict academic performance and support tailored educational interventions. With the proliferation of digital tools in education, institutions now have access to a wealth of data that offers insights into student behaviors, academic performance, and learning needs. This transformation has enabled educators to move beyond traditional teaching methods, using data-driven approaches to enhance learning outcomes and identify students who may be at risk of falling behind.

In recent years, the predictive capabilities of machine learning have expanded significantly, impacting fields from educational data mining to complex financial systems. Studies [1] and [2] have highlighted the potential of identifying key factors influencing student achievement and natural disaster risks, emphasizing the importance of targeted, data-driven interventions. Additionally, comparative analyses between algorithms such as Support Vector Machines (SVM) and Random Forests (RF) illustrate the effectiveness of ensemble learning for applications beyond education, as seen in [3] and [4] demonstrating these models' adaptability to diverse datasets. In financial contexts, the role of clustering and density-based approaches in [5] and [6] has revealed critical insights into identifying patterns within complex datasets, supporting the case for adaptable, robust methodologies in predictive tasks. Clustering techniques are also relevant in geospatial and market analyses, as demonstrated in [7] and [8] which illustrate how environmental and contextual factors can be effectively integrated into predictive models.

EDM serves various functions, but one of its most powerful applications is the ability to predict student performance. Classification algorithms, such as decision trees and regression models, have been widely used to assess academic data and identify key factors influencing student success. Studies by [9] demonstrate that data mining can be employed to recognize patterns in student behavior, allowing educators to identify students who may need additional support. Similarly, [10] show how the C4.5 algorithm can predict student achievement with high accuracy, helping institutions implement timely interventions. These techniques enable educators to personalize learning experiences, offering at-risk students the resources and guidance they need to improve their academic outcomes.

The predictive power of data mining extends beyond identifying at-risk students and supports institutional decision-making. As shown by [11], decision trees can forecast final grades and GPA, providing a clear view of how various factors—such as attendance, participation, and socioeconomic background—affect academic success. This insight allows educators and administrators to take proactive measures to improve the educational environment. Moreover, EDM fosters the development of adaptive learning technologies that cater to individual student needs, as discussed by [12]. As educational institutions continue to embrace data-driven strategies, EDM stands at the forefront of efforts to enhance student performance and drive institutional improvements across various educational contexts.

The influence of socioeconomic factors and school characteristics on educational outcomes is an area of extensive research, emphasizing the intertwined effects of students' backgrounds and their institutional environments. Socioeconomic status (SES) encompasses variables such as family income, parental education, and occupation, all of which are pivotal in determining a student's access to educational resources and opportunities. These elements of SES collectively shape students' academic trajectories, affecting the availability of learning materials and the broader support networks crucial for educational success.

Numerous studies highlight the barriers faced by students from lower socioeconomic backgrounds. Research by [13] reveals a significant negative correlation between parental socioeconomic status and academic performance, illustrating how limited financial resources and lower parental educational attainment can restrict access to crucial educational tools and environments. Similarly, [14] argue that socioeconomic status affects students' psychological well-being, which, in turn, is intricately linked to their academic achievements. However, the complexity of this relationship is highlighted by [15], [16], whose findings suggest that in some contexts, socioeconomic factors may not directly influence academic outcomes, indicating the need for more nuanced investigations into how SES interacts with other variables in educational settings.

In addition to individual socioeconomic factors, school characteristics such as funding, teacher quality, and infrastructure also significantly influence educational outcomes. Schools in affluent communities typically benefit from greater financial support, smaller class sizes, and more experienced teachers, all of which contribute to better academic performance among students. Conversely, schools in lower-income areas often struggle with limited resources, overcrowded classrooms, and higher student-to-teacher ratios, which can negatively affect learning outcomes [16]. The combination of socioeconomic challenges at home and under-resourced school environments can compound disadvantaged students' difficulties, further widening the achievement gap.

The intersection of socioeconomic factors and school characteristics creates a multifaceted framework for understanding educational inequities. Research [17] emphasize that health disparities linked to socioeconomic status also impact educational outcomes, as students from lower SES backgrounds may face health challenges that hinder their academic progress. This underscores the necessity of addressing both home and school environments to create equitable educational opportunities. Tackling educational disparities, therefore, requires interventions that consider the broader socioeconomic context in which students live and the resources available to the schools they attend.

The application of data mining techniques in the field of EDM has gained prominence due to its potential to transform educational processes and outcomes. EDM leverages data mining methods to analyze educational data, uncovering patterns that inform teaching practices and interventions. Techniques such as classification, clustering, and association rule mining have been extensively applied, each offering unique insights into student behavior, performance, and educational dynamics. These methods enable educators and institutions to predict student success, identify at-risk students, and personalize learning experiences, leading to more targeted and effective educational interventions.

The application of machine learning algorithms in predicting student performance has become a significant area of focus within EDM. As educational institutions increasingly rely on data to inform policy decisions, the ability to accurately predict student outcomes based on various factors such as socioeconomic status, school characteristics, and individual student behaviors becomes essential. Publicly available educational datasets provide researchers with valuable resources to develop and test machine learning models. However, the field still demands further exploration into more sophisticated algorithms to improve prediction accuracy and address the complexities of educational data.

Numerous machine learning algorithms have already been applied in this field, yielding promising results. Traditional methods such as Random Forest, XGBoost, and decision trees have been employed to predict various educational outcomes. For example, [18] demonstrated that Artificial Neural Networks (ANN) achieved high prediction accuracy in modeling student performance. However, [19] argued that while these algorithms are effective, they often fail to capture the full complexity of educational data, leaving room for integrating more advanced techniques. The application of SVM and KNN in predicting student achievement shows significant potential, but the diversity of student data necessitates the examination of more complex and hybrid models.

While the field of EDM has advanced significantly, the comparative analysis of machine learning algorithms, particularly Random Forest and XGBoost, in predicting student achievement remains underexplored. The literature has extensively focused on various algorithms for predicting student performance, such as decision trees, neural networks, and ensemble methods. Yet, there is a limited direct comparison between Random Forest and XGBoost in the context of educational data. This gap highlights the need for a deeper understanding of how these two distinct algorithms perform in educational settings, particularly when analyzing the effects of socioeconomic and school-level factors on student achievement.

This research seeks to address this gap by conducting a comparative analysis of Random Forest and XGBoost, specifically in predicting student achievement using publicly available educational datasets. By focusing on the predictive capabilities of these algorithms in relation to socioeconomic and school-level factors, this study aims to provide valuable insights into their relative strengths and weaknesses. The findings from this analysis are expected to contribute to the growing body of knowledge in EDM, offering educators and researchers practical guidance on selecting the most appropriate algorithm for educational performance predictions.

This paper aims to evaluate the predictive power of these two machine learning algorithms by using publicly available educational datasets that incorporate diverse socioeconomic and school-level factors. These datasets allow for a robust analysis of how effectively each algorithm can model the relationships between these factors and student achievement. Understanding the strengths and limitations of each in the educational context will offer important insights for researchers and practitioners looking to apply these techniques in real-world settings.

This study evaluates their respective performances by comparing Random Forest and XGBoost on key metrics such as accuracy, precision, recall, and F1 score. The findings are expected to contribute to the broader field of EDM by highlighting the conditions under which each algorithm performs best, thereby guiding the selection of appropriate models for future research and practical applications in education. This analysis not only fills a gap in the existing literature but also offers practical recommendations for improving predictive modeling in educational systems.

Literature Review

Educational Data Mining (EDM)

EDM is a field dedicated to applying data mining methodologies to educational data to uncover patterns and insights that can be used to enhance the learning process and inform institutional decision-making. It draws upon various computational techniques, including classification, regression, clustering, and association rule mining, to analyze data originating from educational environments like schools and universities [20], [21]. EDM's main objective is to improve understanding of student learning behaviors, predict academic performance, and suggest ways to optimize educational outcomes [22], [23]. Its application spans a broad spectrum, from identifying at-risk students to assessing the effectiveness of different instructional strategies, making it an indispensable tool in modern education [24].

The significance of EDM lies in its ability to manage and interpret the vast volumes of data generated by educational systems. As institutions continue to gather increasing amounts of information on student performance, attendance, and engagement, EDM provides educators and administrators with advanced tools to extract actionable insights. For example, predictive models developed within the realm of EDM can forecast key outcomes like graduation rates and academic success, offering opportunities for timely intervention and personalized student support. This data-driven approach ensures that decisions made at both the classroom and institutional levels are informed by robust, evidence-based insights, contributing to more efficient and effective educational practices [25].

Data Mining Techniques Used in Educational Settings

Data mining techniques have been extensively used in educational settings to predict student performance, with algorithms like decision trees and random forests being some of the most applied methods. These approaches help educational institutions anticipate student challenges and implement necessary interventions early on, ultimately improving educational outcomes.

For instance, Karalar et al. utilized a hybrid model combining Extra Trees, Random Forest, and Logistic Regression to predict students at risk of academic failure during the pandemic in distance learning contexts. The model's high specificity, reaching 90.34%, demonstrated the efficiency of ensemble methods in identifying at-risk students accurately [26]. Similarly, Sghir et al. conducted a systematic review of predictive learning analytics, highlighting how Random Forests, when employed in ensemble models, have consistently demonstrated high predictive accuracy in assessing academic performance over the past decade [27]. This growing body of evidence supports the utility of data mining techniques in the field of educational data mining.

Further research by Akçapınar et al. demonstrated the effectiveness of learning analytics in predicting student performance through early-warning systems. By focusing on engagement metrics, such as participation in online forums, they revealed how these behavioral indicators could predict academic success [28]. Additionally, Chen and Upah's study emphasized the practical application of predictive analytics in academic advising, where decision trees provided valuable insights that guided students toward more successful academic trajectories [29]. These studies underscore the relevance of data mining techniques in developing predictive tools that enable personalized support for students.

Method

The research method for this study consists of several steps to ensure a comprehensive and accurate analysis. The flowchart in Figure 1 outlines the detailed steps of the research method.

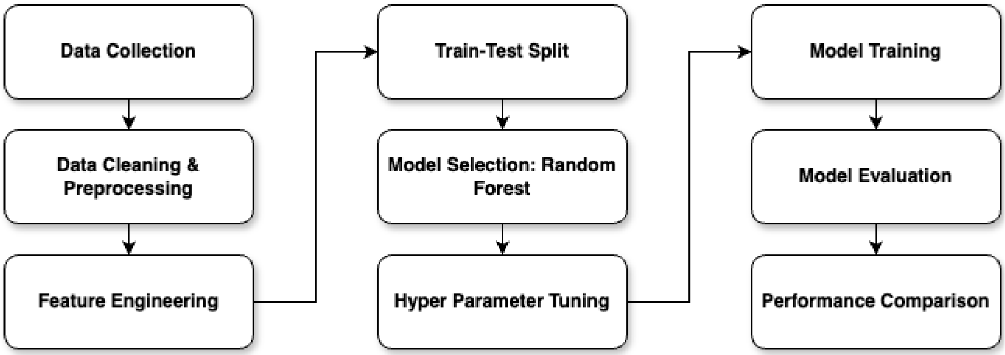

Figure 1 Research Method Flowchart

Dataset Description

The dataset used in this study comprises 6,607 records and 60 columns, representing a range of variables related to school demographics, student performance, and socioeconomic factors. The dataset provides detailed information on the schools, such as School Type, Province, School Level, and Enrolment, which contribute to an understanding of the educational context within which students operate. Each row in the dataset corresponds to a specific school or educational institution, capturing key metrics and characteristics that are relevant for predicting student achievement.

Key features used in the analysis include several socioeconomic factors and school-level characteristics. The Percentage of School-Aged Children Who Live in Low-Income Households and the Percentage of Students Whose Parents

Have No Degree, Diploma, or Certificate serve as proxies for socioeconomic status, offering insight into the economic and educational background of the student population. Additionally, Percentage of Students Receiving Special Education Services and Percentage of Students Identified as Gifted reflect the diversity of learning needs within schools. These features are crucial in understanding how varying educational backgrounds and services impact overall student performance.

Academic performance metrics are also integral to the dataset. The study focuses on key indicators such as the Percentage of Grade 6 Students Achieving the Provincial Standard in Reading, Writing, and Mathematics. These scores provide a snapshot of students' academic abilities across core subjects. Moreover, the Change in Achievement Over Three Years for each of these subjects offers a longitudinal view of academic progress, enabling a more nuanced analysis of factors affecting student achievement over time. These academic measures, when combined with socioeconomic factors, allow for a comprehensive analysis of the relationship between external influences and educational outcomes.

Preprocessing steps included handling missing values and standardizing the dataset for analysis. Columns with missing data were imputed using the mean for numerical variables and the mode for categorical variables. This ensured that the dataset was complete and consistent. Additionally, numerical features such as Latitude and Longitude were scaled to standardize the input, preventing variables with larger numerical ranges from disproportionately influencing the machine learning models. This preprocessing prepared the dataset for accurate and effective analysis using machine learning algorithms.

Exploratory Data Analysis (EDA)

To address the issue of missing data, imputation techniques were applied to ensure the dataset was complete for analysis. Columns with missing values, such as the Percentage of School-Aged Children Who Live in Low-Income Households and Est Absent Students, had missing rates of 4.12% and 7.81%, respectively. These values were imputed using the column’s mean for continuous variables and the mode for categorical variables. For instance, key columns such as Percentage of Grade 3 Students Achieving the Provincial Standard in Reading, which had approximately 26% missing data, were similarly treated using mean imputation. This approach allowed us to minimize data loss and maintain consistency across the dataset.

Key descriptive statistics for important features were generated to gain insight into the central tendencies and spread of the data. For example, Est Absent Students had a mean of 54.79, with a standard deviation of 53.45, indicating a wide variance in absenteeism rates across schools. The median absenteeism was 44, suggesting that half of the schools reported absenteeism below this level. For the Percentage of School-Aged Children Who Live in Low-Income Households, the mean percentage was approximately 23.5%, highlighting socioeconomic disparities that could potentially impact academic performance.

Visual exploration of the data was conducted through histograms and boxplots, providing a clear view of the distribution of key variables. For instance, a histogram of Family Income Levels indicated a right-skewed distribution, with a significant portion of students coming from lower-income households. Boxplots of Student Performance metrics, such as Grade 3 Mathematics Achievement, revealed potential outliers and a significant range in academic achievement across schools. Additionally, visualizing Parental Education Levels helped illustrate the varying degrees of educational attainment among students' parents, which is a critical predictor of academic success.

A correlation matrix was generated and visualized using a heatmap to understand the relationships between key socioeconomic factors and student performance. For instance, a moderate negative correlation (-0.45) was observed between Percentage of School-Aged Children Who Live in Low-Income Households and Grade 6 Reading Achievement, suggesting that students from lower-income households tend to perform worse in reading. Similarly, Parental Education Levels exhibited a positive correlation (0.62) with Student Performance, emphasizing the role of parental influence in student success. This correlation matrix provided valuable insights into how external socioeconomic factors might be driving academic outcomes, helping to guide further analysis.

Feature Selection

Feature selection was a critical step in preparing the dataset for model training, ensuring that only the most relevant predictors were included in the analysis. The dataset contained numerous variables, but specific attention was given to those with a direct impact on student achievement, particularly socioeconomic and school-level factors. Features such as Est Absent Students, Percentage of School-Aged Children Who Live in Low-Income Households, Percentage of Students Whose Parents Have No Degree, Diploma or Certificate, and Percentage of Students Receiving Special Education Services were identified as significant predictors of student performance. These variables were chosen based on their known correlations with academic success, as indicated in prior research.

To ensure the integrity of the analysis, only numeric features were retained, and categorical variables were either excluded or converted into numeric form where necessary. Missing values in numeric columns were handled through mean imputation, ensuring that the dataset was complete and suitable for further analysis. This preprocessing step was crucial, as missing data could potentially skew the results or lead to inaccuracies in the models. By filling missing data with the mean, we maintained the overall statistical balance without introducing bias from outliers.

A key part of the feature selection process involved examining the correlations between variables. A heatmap was generated to visualize these relationships, showing, for instance, a moderate negative correlation between Percentage of School-Aged Children Who Live in Low-Income Households and student performance, as measured by Percentage of Grade 6 Students Achieving the Provincial Standard in Reading. This correlation indicated that students from lower-income households tended to have lower academic achievement, a finding consistent with educational research. Additionally, a positive correlation was observed between Parental Education Levels and student success, reinforcing the importance of family background in shaping academic outcomes.

After feature selection and correlation analysis, a classification process was applied to the target variable, Percentage of Grade 6 Students Achieving the Provincial Standard in Reading. Students were classified into three categories: High Achievers (80% or above), Moderate Achievers (50%–79%), and Low

Achievers (below 50%). This multi-class classification allowed for a more nuanced analysis of student performance and provided a clearer understanding of how various socioeconomic and school-level factors influenced different levels of academic achievement.

Data Normalization

To prepare the dataset for model training, data normalization was applied, primarily using Min-Max Scaling. This approach was chosen to bring all feature values into a standardized range between 0 and 1, which is particularly beneficial for models sensitive to feature scale variations, such as Random Forest (RF) and XGBoost. By scaling each feature to a similar range, the model's ability to distinguish patterns in the data improves, especially for algorithms like XGBoost, which relies on gradient-based methods that benefit from consistent feature distributions. The normalization process also helps reduce potential biases introduced by features with larger value ranges, thereby enhancing overall model accuracy.

Features such as Est Absent Students and Percentage of School-Aged Children Who Live in Low-Income Households were normalized to ensure that their scale aligned with other predictive variables. This normalization not only facilitated smoother model training but also allowed for more effective hyperparameter tuning, as the models could interpret each feature's contribution proportionally. Scaling was done separately on the training and test sets to avoid any data leakage, maintaining the validity of the evaluation process.

Model Training

The Random Forest model was trained using a grid search to optimize key hyperparameters. Random Forest, known for its robustness and ability to handle large feature sets, builds multiple decision trees and averages their predictions to enhance model stability and accuracy. For this study, hyperparameters such as `n_estimators` (set to 100 and 200), `max_depth` (10 and 20), `min_samples_split` (2 and 5), and `min_samples_leaf` (1 and 2) were tuned through 5-fold cross-validation, which provided a reliable assessment of model performance across different subsets of the training data. The final Random Forest model, selected based on its cross-validated accuracy, demonstrated high precision and recall, as seen in the classification report, underscoring its suitability for this educational dataset.

The XGBoost model was configured to enhance accuracy and capture complex patterns through gradient boosting. Known for its efficiency in handling large datasets with many features, XGBoost combines several weak learners to reduce bias and variance within the model. Key hyperparameters included `n_estimators` (100 and 200), `learning_rate` (0.01 and 0.1), `max_depth` (6 and 10), and `subsample` (0.8 and 1.0), optimized through grid search with 5-fold cross-validation. This configuration allowed the model to generalize well without overfitting. The final XGBoost model achieved impressive results, with an accuracy comparable to Random Forest and strong precision, recall, and F1-scores, particularly in the higher-achieving class categories.

Overall, the data normalization and model training processes collectively contributed to developing predictive models capable of effectively analyzing student achievement based on socioeconomic and school-level factors. The final models were saved using `joblib` for potential future use, enabling consistent and efficient deployment in practical educational data mining applications.

Result and Discussion

The performance of the Random Forest (RF) and XGBoost models was evaluated based on metrics including precision, recall, F1-score, and accuracy. For RF, the classification report indicates high precision and recall across all categories, with an accuracy of 0.9965. Specifically, the model demonstrated a weighted average F1-score of 1.00, signifying robust performance in predicting the three categories of student achievement. The RF model's performance highlights its effectiveness in identifying patterns within the dataset, handling imbalanced classes efficiently, and producing minimal misclassifications.

XGBoost similarly showed strong performance metrics, with an accuracy of 0.9981 and a weighted F1-score of 1.00. The precision and recall scores were consistent across the target classes, indicating that XGBoost successfully differentiated among the various achievement levels. In terms of accuracy, XGBoost slightly outperformed RF, suggesting that its gradient-boosting mechanism contributed to slightly higher precision in classification. This performance affirms XGBoost's utility in educational data mining tasks, where complex patterns in socioeconomic and school-level factors need to be accounted for in predicting student achievement.

To further illustrate the models' performance, a comparative visualization of the accuracy, precision, recall, and F1-scores for both RF and XGBoost was created. Bar plots show the high alignment of performance metrics between the two models, with minor improvements in accuracy and recall observed for XGBoost across most categories. Additionally, confusion matrices visually depict the true positives and false positives across achievement categories, highlighting that both models exhibited minimal error rates, particularly in distinguishing high-achieving students from others.

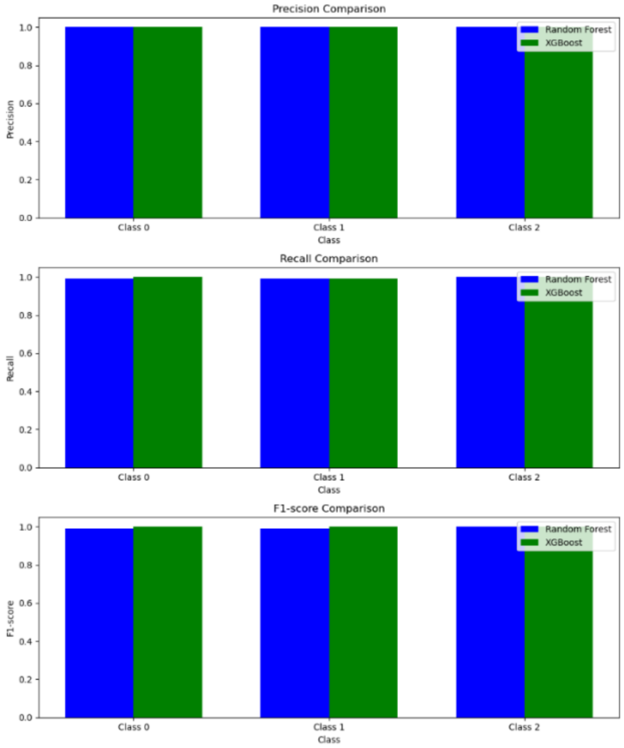

Figure 2 Performance Metric Comparison

Figure 2 presents a comparison between the performance of the Random Forest (RF) and XGBoost algorithms across three key metrics: precision, recall, and F1-score for each class (Class 0, Class 1, Class 2). Both models show almost identical performance across all metrics and classes. The precision, recall, and F1-scores for Class 0, Class 1, and Class 2 are all close to 1.00 for both algorithms. This indicates that both models exhibit excellent classification capabilities with minimal difference between their performances, as reflected by the similar height of the bars in each plot.

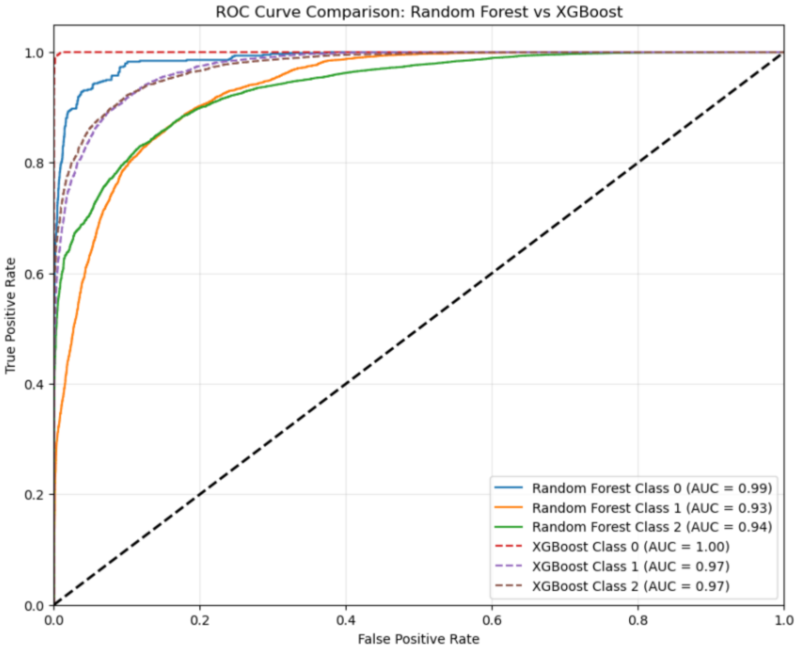

Figure 3 ROC Curve Comparison

Figure 3 present comparison between Random Forest and XGBoost provides further insight into their classification performance. The ROC curves show that both models performed well across all three classes, with XGBoost slightly outperforming Random Forest in terms of the Area Under the Curve (AUC) values. Specifically, XGBoost achieved an AUC of 1.00 for Class 0 and 0.97 for both Class 1 and Class 2, while Random Forest attained an AUC of 0.99 for Class 0, 0.93 for Class 1, and 0.94 for Class 2. These high AUC scores for both models indicate their strong ability to distinguish between the different classes, with XGBoost demonstrating a marginally superior performance, particularly for Class 1 and Class 2.

While both RF and XGBoost achieved high performance, XGBoost demonstrated a slight edge over RF in terms of accuracy and precision, which can be attributed to its ability to minimize errors by iteratively adjusting weights in its boosting process. The RF model, however, presented a similar overall F1-score and performed comparably across the classification metrics, making it a highly reliable alternative. XGBoost's computational efficiency in handling feature importance and its ability to manage complex, nonlinear relationships in the data make it advantageous in datasets with intricate inter-variable dynamics, as seen in educational contexts.

The results of the study demonstrated that both the Random Forest (RF) and XGBoost models exhibited exceptional performance in predicting student achievement based on socioeconomic and school-level factors. XGBoost slightly outperformed RF across most metrics, particularly in terms of AUC scores, where XGBoost achieved 1.00 for Class 0 and 0.97 for Class 1 and Class 2, compared to RF's 0.99, 0.93, and 0.94 for the same classes. Although both models showed near-perfect classification performance in terms of precision, recall, and F1-score, the marginal advantage of XGBoost can be attributed to its ability to handle complex relationships and non-linear interactions between the features more efficiently than RF.

One of the reasons XGBoost consistently performed slightly better than Random Forest could be due to its use of gradient boosting, which optimizes the model by minimizing the error iteratively. XGBoost is known for its capacity to handle high-dimensional data with ease and its ability to reduce bias and variance more effectively. In contrast, while Random Forest is robust and powerful for classification tasks, it may be less flexible in capturing intricate patterns in the data, particularly when dealing with imbalanced classes or subtle variations in the features.

Despite the promising results, this study encountered several limitations. One challenge was handling the large number of missing values in key features such as parental education and income levels, which could have influenced the model's performance. Imputing missing data using the mean for continuous variables may have led to the loss of nuanced patterns within the dataset. Furthermore, the dataset used in this study may not fully capture the dynamic nature of socioeconomic factors over time, limiting the generalizability of the findings to different educational settings or populations. Additionally, while both models achieved high accuracy, their practical application in real-world educational interventions requires further validation to ensure their robustness across varied datasets.

The findings of this study emphasize the strong influence of socioeconomic factors on student performance. Features such as family income, parental education levels, and school-level factors were significant predictors of student achievement, aligning with existing literature that highlights socioeconomic disparities in education. These results suggest that targeted educational interventions, such as additional academic support for students from low-income households or schools with limited resources, could help mitigate the impact of socioeconomic inequalities. Policies that focus on improving access to educational resources, tutoring programs, and parental involvement in education may play a crucial role in enhancing student performance and reducing achievement gaps.

Conclusion

This study compared the performance of two machine learning models, Random Forest (RF) and XGBoost, in predicting student achievement based on socioeconomic and school-level factors. Both models exhibited excellent performance, with XGBoost slightly outperforming RF across all evaluation metrics, including precision, recall, and F1-score. The ROC curve comparison further highlighted XGBoost's superior ability to handle complex patterns in the data. Among the predictors, socioeconomic factors, such as family income and parental education, as well as school characteristics like school type and access to resources, emerged as the most significant determinants of student success.

The findings of this study underscore the importance of socioeconomic factors in shaping student outcomes. Students from lower-income households and schools with limited resources are more likely to underperform, reinforcing the need for equitable distribution of educational resources. Policymakers could use these insights to guide resource allocation, ensuring that schools serving disadvantaged communities receive the support needed to improve student outcomes. Furthermore, educational policies should prioritize interventions that promote parental involvement and provide additional support for students facing socioeconomic challenges, ultimately bridging the achievement gap.

Future research could explore the application of additional machine learning algorithms, such as deep learning models, to improve prediction accuracy. Testing these models on other educational datasets with varying socioeconomic contexts could also provide insights into the generalizability of the findings. Additionally, incorporating more granular student-level data, such as individual academic progress or behavioral metrics, may enhance the models' ability to predict student achievement more precisely. Further exploration of hybrid models or ensemble techniques may also offer new perspectives for refining predictive performance.

While both RF and XGBoost performed well, the models could benefit from incorporating a broader range of features, such as real-time data on student engagement or attendance. Future models could also consider longitudinal data to track changes in socioeconomic factors over time, providing a more dynamic and comprehensive view of how these factors influence academic success. Moreover, integrating more advanced feature selection techniques could help in identifying additional critical variables, allowing for a more targeted approach to improving educational interventions.

Declarations

Author Contributions

Conceptualization: T.S.; Methodology: T.S.; Software: T.S.; Validation: T.S.; Formal Analysis: T.S.; Investigation: T.S.; Resources: T.S.; Data Curation: T.S.; Writing Original Draft Preparation: T.S.; Writing Review and Editing: T.S.; Visualization: T.S.; The author have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

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

References

[1] M. Murnawan, S. Lestari, R. Samihardjo, and D. A. Dewi, "Sustainable Educational Data Mining Studies: Identifying Key Factors and Techniques for Predicting Student Academic Performance," *Journal of Applied Data Sciences*, vol. 5, no. 3, Art. no. 3, Sep. 2024, doi: 10.47738/jads.v5i3.347.

[2] H. T. Sukmana, Y. Durachman, A. Amri, and S. Supardi, "Comparative Analysis of

SVM and RF Algorithms for Tsunami Prediction: A Performance Evaluation Study," *Journal of Applied Data Sciences*, vol. 5, no. 1, Art. no. 1, Jan. 2024, doi: 10.47738/jads.v5i1.159.

[3] B. H. Hayadi and I. M. M. E. Emary, "Predicting Campaign ROI Using Decision Trees and Random Forests in Digital Marketing," *Journal of Digital Market and Digital Currency*, vol. 1, no. 1, Art. no. 1, May 2024, doi: 10.47738/jdmdc.v1i1.5.

[4] Hery and A. E. Widjaja, "Analysis of Apriori and FP-Growth Algorithms for Market Basket Insights: A Case Study of The Bread Basket Bakery Sales," *Journal of Digital Market and Digital Currency*, vol. 1, no. 1, Art. no. 1, May 2024, doi: 10.47738/jdmdc.v1i1.2.

[5] T. Hariguna and A. S. M. Al-Rawahna, "Unsupervised Anomaly Detection in Digital Currency Trading: A Clustering and Density-Based Approach Using Bitcoin Data," *Journal of Current Research in Blockchain*, vol. 1, no. 1, Art. no. 1, Jun. 2024, doi: 10.47738/jcrb.v1i1.12.

[6] Henderi and Q. Siddique, "Anomaly Detection in Blockchain Transactions within the Metaverse Using Anomaly Detection Techniques," *Journal of Current Research in Blockchain*, vol. 1, no. 2, Art. no. 2, Sep. 2024, doi: 10.47738/jcrb.v1i2.17.

[7] D. Sugianto and A. R. Hananto, "Geospatial Analysis of Virtual Property Prices Distributions and Clustering," *Int. J. Res. Metaverese*, vol. 1, no. 2, Art. no. 2, Sep. 2024, doi: 10.47738/ijrm.v1i2.10.

[8] T. Wahyuningsih and S. C. Chen, "Determinants of Virtual Property Prices in Decentraland an Empirical Analysis of Market Dynamics and Cryptocurrency Influence," *Int. J. Res. Metaverese*, vol. 1, no. 2, Art. no. 2, Sep. 2024, doi: 10.47738/ijrm.v1i2.12.

[9] P. Amutha and R. Priya, "A Survey on Educational Data Mining Techniques in Predicting Student's Academic Performance," *Int. J. Eng. Technol.*, vol. 7, no. 3.3, p. 634, 2018, doi: 10.14419/ijet.v7i2.33.14853.

[10] G. A. Putri, D. Maryono, and F. Liantoni, "Implementation of the C4.5 Algorithm to Predict Student Achievement at SMK Negeri 6 Surakarta," *Ijie Indones. J. Inform. Educ.*, vol. 4, no. 2, p. 51, 2020, doi: 10.20961/ijie.v4i2.47124.

[11] M. A. Al-Barrak and M. Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study," *Int. J. Inf. Educ. Technol.*, vol. 6, no. 7, pp. 528–533, 2016, doi: 10.7763/ijiet.2016.v6.745.

[12] A. K. AL-Mashanji, A. H. Hamza, and L. H. Alhasnawy, "Computational Prediction Algorithms and Tools Used in Educational Data Mining: A Review," *J. Univ. Babylon Pure Appl. Sci.*, pp. 87–99, 2023, doi: 10.29196/jubpas.v31i1.4531.

[13] Q. Suleman, I. Hussain, and Z. Nisa, "Effects of Parental Socioeconomic Status on the Academic Achievement of Secondary School Students in District Karak (Pakistan)," *Int. J. Hum. Resour. Stud.*, vol. 2, no. 4, p. 14, 2014, doi: 10.5296/ijhrs.v2i4.2511.

[14] F. Alivernini, E. Cavicchiolo, S. Manganelli, A. Chirico, and F. Lucidi, "Students' Psychological Well-Being and Its Multilevel Relationship With Immigrant Background, Gender, Socioeconomic Status, Achievement, and Class Size," *Sch. Eff. Sch. Improv.*, vol. 31, no. 2, pp. 172–191, 2019, doi: 10.1080/09243453.2019.1642214.

[15] M. Fateel, S. Mukallid, and B. Arora, "The Interaction Between Socioeconomic Status and Preschool Education on Academic Achievement of Elementary School Students," *Int. Educ. Stud.*, vol. 14, no. 8, p. 60, 2021, doi: 10.5539/ies.v14n8p60.

[16] B. Bado and T. Tahir, "The Influence of Parents' Socio-Rconomic Status on Student Academic Achievement at Vocational Schools," pp. 215–222, 2023, doi: 10.2991/978-2-38476-084-8_29.

[17] K. Jiao, M. Xu, and M. Liu, "Health Status and Air Pollution Related Socioeconomic Concerns in Urban China," *Int. J. Equity Health*, vol. 17, no. 1, 2018, doi: 10.1186/s12939-018-0719-y.

[18] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *Ieee Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/access.2020.2986809.

[19] S. F. A. Aziz, “Students’ Performance Evaluation Using Machine Learning Algorithms,” *Coll. Basic Educ. Res. J.*, vol. 16, no. 3, pp. 977–986, 2020, doi: 10.33899/berj.2020.166006.

[20] C. Romero and S. Ventura, “Data Mining in Education,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, no. 1, pp. 12–27, 2012, doi: 10.1002/widm.1075.

[21] A. F. Meghji, N. A. Mahoto, M. A. Unar, and M. A. Shaikh, “Predicting Student Academic Performance Using Data Generated in Higher Educational Institutes,” *3c Tecnol. Innov. Apl. Pyme*, pp. 366–383, 2019, doi: 10.17993/3ctecno.2019.specialissue2.366-383.

[22] F. Alshareef, H. Alhakami, T. Alsubait, and A. Baz, “Educational Data Mining Applications and Techniques,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, 2020, doi: 10.14569/ijacsa.2020.0110494.

[23] S. Alturki, I. Hulpuș, and H. Stuckenschmidt, “Predicting Academic Outcomes: A Survey From 2007 Till 2018,” *Technol. Knowl. Learn.*, vol. 27, no. 1, pp. 275–307, 2020, doi: 10.1007/s10758-020-09476-0.

[24] D. Agha, “Clusters of Success: Unpacking Academic Trends With K-Means Clustering in Education,” *Vfast Trans. Softw. Eng.*, vol. 11, no. 4, pp. 15–31, 2023, doi: 10.21015/vtse.v11i4.1633.

[25] M. Arifin, W. Widowati, and . Farikhin, “Using Education Data Mining (EDM) and Tracer Study (TS) Data as Materials for Evaluating Higher Education Curriculum and Policies,” *Kne Soc. Sci.*, 2022, doi: 10.18502/kss.v7i14.11948.

[26] H. Karalar, C. Kapucu, and H. Gürüler, “Predicting Students at Risk of Academic Failure Using Ensemble Model During Pandemic in a Distance Learning System,” *Int. J. Educ. Technol. High. Educ.*, vol. 18, no. 1, 2021, doi: 10.1186/s41239-021-00300-y.

[27] N. Sghir, A. Adadi, and M. Lahmer, “Recent Advances in Predictive Learning Analytics: A Decade Systematic Review (2012–2022),” *Educ. Inf. Technol.*, vol. 28, no. 7, pp. 8299–8333, 2022, doi: 10.1007/s10639-022-11536-0.

[28] G. Akçapınar, A. Altun, and P. Aşkar, “Using Learning Analytics to Develop Early-Warning System for at-Risk Students,” *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, 2019, doi: 10.1186/s41239-019-0172-z.

[29] Y. Chen and S. Upah, “Data Analytics and STEM Student Success: The Impact of Predictive Analytics-Informed Academic Advising Among Undeclared First-Year Engineering Students,” *J. Coll. Stud. Retent. Res. Theory Pract.*, vol. 22, no. 3, pp. 497–521, 2018, doi: 10.1177/1521025118772307.