Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15953
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparative Academic Performance Prediction in Primary Schools Using Linear Regression and Random Forest

Agustinus Sembiring[1)*], Handri Santoso[2)]
[1)2)]Information and Technology, Universitas Pradita, Indonesia
[1)]sembiringagustinus2@gmail.com, [2)]handri.santoso@pradita.ac.id

Submitted :2 Mar 2026 | **Accepted** : 12 Mar 2026 | **Published** : April 2, 2026

Abstract: Predicting academic performance is an important aspect of data-driven decision making in education, particularly in primary schools where early identification of learning difficulties is crucial. This study compares the performance of Linear Regression and Random Forest Regression models for predicting students' academic performance using an Educational Data Mining approach. The experiment uses the Students Performance Dataset from Kaggle, consisting of 1000 student records with eight predictor variables, including demographic and learning-related attributes. The target variable is the average score derived from mathematics, reading, and writing results. Model development and evaluation are conducted using Python in Google Colaboratory. Performance is assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), while Random Forest is further optimized using GridSearchCV with 5-fold cross-validation. The results show that Linear Regression achieves the best performance ($R^2 = 0.162$, RMSE = 13.40, MAE = 10.49), outperforming both the default Random Forest ($R^2 \approx 0.000$) and the tuned Random Forest ($R^2 \approx 0.112$). Although the explained variance is relatively low, this finding indicates that simple demographic features provide limited predictive power for academic performance. A case study using a local dataset from a private primary school involving 132 sixth-grade students further confirms that Linear Regression performs more effectively than Random Forest for small and simple educational datasets. These results highlight the importance of aligning model selection with dataset characteristics in educational data mining.

Keywords: Academic Performance Prediction, Educational Data Mining (EDM), Linear Regression, Machine Learning, Random Forest Regression

INTRODUCTION

Primary education plays a crucial role as the foundation for students' academic development, as fundamental cognitive and social skills are formed at this stage. Students acquire essential competencies in primary school that significantly influence their success in later educational levels. Early identification of learning difficulties is important because students who experience persistent academic problems in primary education often continue to face challenges at later educational stages. In recent years, the use of Educational Data Mining (EDM) and machine learning techniques has become increasingly important for supporting data-driven decision making in education and for analyzing factors that influence academic performance (Ling et al., 2024; Lyu & Xu, 2025; Nugraha et al., 2025). These approaches enable researchers and educators to identify patterns within educational data and support more effective learning interventions.

Numerous studies have applied machine learning techniques to predict students' academic performance using various datasets and algorithms (Abro et al., 2025; Bussaman et al., 2024; Deleña et al., 2025). However, previous research indicates that the effectiveness of predictive models is strongly influenced by dataset characteristics, including data size, feature relevance, and variability (Qureshi & Lokhande, 2024). While complex machine learning models often demonstrate strong performance in large and feature-rich datasets, their effectiveness may decrease when applied to small or simple educational datasets. This challenge is particularly relevant in primary education contexts, where available datasets are often limited and consist of relatively simple variables.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15953
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Despite the growing use of machine learning in education, comparative studies evaluating the performance of different prediction models in primary education contexts remain limited. Therefore, this study aims to compare the performance of Linear Regression and Random Forest Regression models for predicting students' academic performance using educational datasets with relatively simple characteristics. The contribution of this research lies in providing empirical evidence regarding the suitability of linear and non-linear machine learning models when applied to small-scale educational datasets, as well as highlighting the importance of aligning model selection with dataset characteristics in educational data mining.

LITERATURE REVIEW

The rapid development of information technology has significantly influenced the education sector, particularly in the use of data-driven approaches to analyze student learning outcomes. Educational Data Mining (EDM) has emerged as an interdisciplinary research field that focuses on developing computational methods to analyze data generated in educational environments and to support data-driven decision making in learning processes. EDM applies various data mining and machine learning techniques to discover patterns related to students' learning behavior, performance, and educational outcomes (Romero & Ventura, 2010, 2020). Through these analytical approaches, researchers and educators can better understand factors influencing academic success and design more effective educational interventions (Bulut et al., 2025; Kostopoulos et al., 2026). Through the application of machine learning algorithms, educational data mining enables the development of predictive models capable of identifying factors that influence student performance and learning outcomes (Deleña et al., 2025; Thaher & Jayousi, 2020).

Previous studies have applied various machine learning techniques to predict academic performance using different types of educational datasets. These studies demonstrate that the performance of predictive models is strongly influenced by dataset characteristics, including data size, feature relevance, and variability (Jabir et al., 2025). In many cases, complex machine learning algorithms such as Random Forest or ensemble methods achieve high predictive accuracy when applied to large datasets with diverse features. However, several studies also report that simpler regression-based models can produce competitive or even superior performance when the dataset contains limited features or exhibits predominantly linear relationships (Hegde et al., 2023).

Linear Regression is widely used as a baseline predictive model due to its simplicity, interpretability, and effectiveness in modeling linear relationships between variables. In contrast, Random Forest is a non-linear ensemble learning algorithm that combines multiple decision trees to improve predictive accuracy and model stability. Random Forest often performs well when datasets contain complex and non-linear relationships among variables. Nevertheless, its performance may decline when applied to small-scale datasets with limited features, where simpler models may provide more stable predictions.

Despite the growing body of research in educational data mining, most existing studies focus on higher education contexts and the use of complex machine learning models. Comparative studies that evaluate the performance of simple and non-linear models in primary education contexts remain relatively limited (Poh & Khor, 2024). Therefore, further investigation is required to understand how different predictive models perform when applied to educational datasets with simple characteristics, particularly in primary school environments where data availability is often limited.

Table 1. Summary of Previous Studies on Academic Performance Prediction

Study	Dataset	Method	Key Findings
Deleña et al. (2025)	Higher education dataset	Machine Learning models	ML models effectively predict student retention
Hegde et al. (2023)	Educational dataset	PCA + ML	Simpler models perform well with limited features
Jabir et al. (2025)	E-learning dataset	ML framework	Model performance depends on dataset characteristics
Poh & Khor (2024)	Online learning dataset	Predictive analytics	Complex models perform well on large datasets

Based on the literature review, this study evaluates the predictive performance of Linear Regression and Random Forest models for predicting academic performance using educational datasets with relatively simple characteristics.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15953
e-ISSN : 2541-2019
p-ISSN : 2541-044X

METHOD

To achieve the objectives of this study, a structured and systematic methodological framework is established. This research adopts a quantitative approach based on educational data mining, encompassing data collection, data understanding and preprocessing, the development of predictive models using Linear Regression and Random Forest Regression, model performance evaluation using appropriate regression metrics, and interpretative analysis of the results. These stages are designed to ensure the validity and reliability of the findings in the context of predicting students' academic performance. In line with previous studies that employ educational data mining and machine learning techniques for academic performance analysis and prediction, this study focuses on developing and comparing the performance of Linear Regression and Random Forest Regression models to obtain a more comprehensive understanding of the effectiveness of data-driven academic performance prediction approaches (Begum & Padmannavar, 2023; Soares et al., 2022). The methodology is designed to ensure a fair and replicable comparison of models. The overall research methodology is presented in the form of a flowchart illustrating the stages of the study, from data processing and model development to performance evaluation and result analysis.

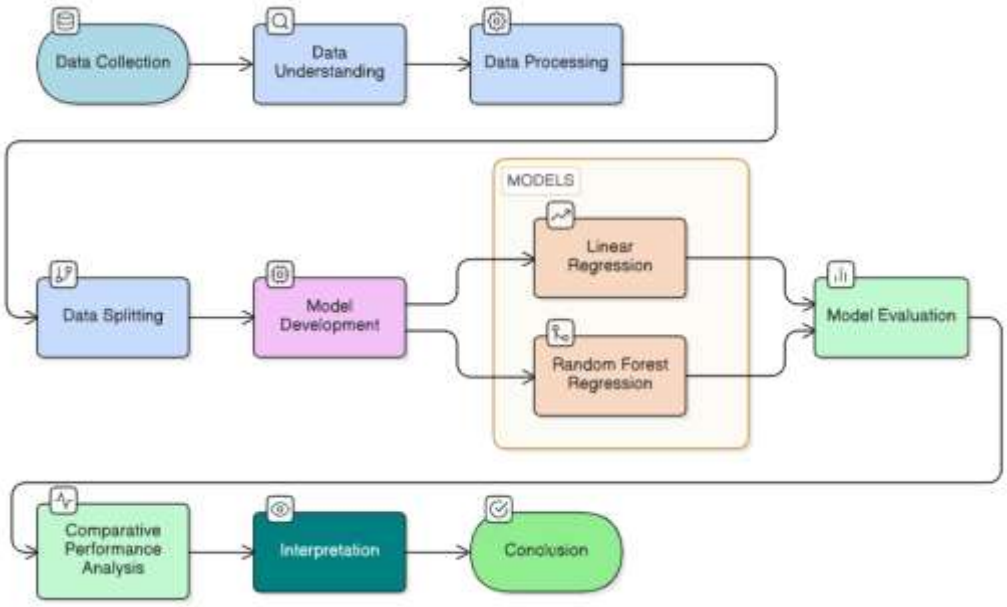

Fig. 1 Flowchart of Research Methodology

The use of authentic datasets in scientific research is considered important for enhancing the validity and comparability of research findings (Eriksson et al., 2021), The data used in this study include academic and non-academic variables relevant to students' learning performance, which are subsequently analyzed using two machine learning approaches Linear Regression and Random Forest Regression to compare model performance in predicting academic achievement. This comparative approach is consistent with previous studies that emphasize the importance of evaluating education-based predictive models using different regression algorithms (Ali et al., 2025).

Dataset and Data Source

This study uses the Students Performance Dataset obtained from Kaggle https://www.kaggle.com/datasets/spscientist/students-performance-in-exams, which contains 1000 student records and nine attributes, including demographic variables and academic scores. The dataset includes features such as gender, race/ethnicity, parental level of education, lunch type, and participation in a test preparation course, as well as academic scores in mathematics, reading, and writing.

In this study, the target variable is constructed as the average score, calculated from the mathematics, reading, and writing scores to represent overall academic performance. The dataset was selected because it provides structured educational attributes commonly used in educational data mining research and allows reproducible experimentation.

In addition to the public dataset, a local dataset from a private primary school consisting of 132 sixth-grade students was also used to evaluate model robustness on smaller educational datasets.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15953
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Data Processing

The data preprocessing stage is conducted to ensure the quality and consistency of the dataset before it is used for predictive model development. This process includes data cleaning to address missing values and inconsistencies, transforming categorical variables into numerical representations, and normalizing numerical features to ensure comparable scales. All preprocessing and analysis procedures are performed using a Python-based programming environment on Google Colaboratory, ensuring that the data are prepared for the development and evaluation of machine learning models.

Since several predictor variables are categorical, a preprocessing step was conducted using one-hot encoding to transform categorical attributes into numerical representations suitable for machine learning algorithms. This transformation enables Linear Regression and Random Forest models to process categorical features effectively without introducing ordinal bias.

Multicollinearity Test

To ensure the validity of the Linear Regression model, a multicollinearity test was conducted using the Variance Inflation Factor (VIF). Multicollinearity occurs when predictor variables exhibit strong linear relationships with each other, which may distort regression coefficient estimates. A commonly accepted threshold is VIF < 10, indicating that multicollinearity is not problematic.

Regression Assumption Testing

Residual analysis was conducted to evaluate the assumptions of the linear regression model. The distribution of residuals was examined to assess normality, while scatter plots of predicted values against residuals were used to assess homoscedasticity.

Feature Importance Analysis

Feature importance analysis was conducted using the Random Forest model to identify the relative contribution of predictor variables in predicting students' academic performance.

Model Development and Evaluation

This study develops predictive models using two machine learning approaches: Linear Regression and Random Forest Regression to predict students' academic performance based on demographic and educational attributes. Linear Regression is employed as a baseline model to analyze the linear relationship between predictor variables and the target variable (average score), as it provides interpretable results and is commonly used in educational data analysis. Random Forest Regression is used as a comparative non-linear ensemble method that combines multiple decision trees to improve predictive capability and reduce overfitting. To improve the performance of the Random Forest model, hyperparameter tuning is conducted using GridSearchCV with 5-fold cross-validation, allowing the model to identify optimal parameter configurations for the dataset.

The performance of the developed models is evaluated using three regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). These metrics measure prediction accuracy and the proportion of variance in academic performance explained by the model. To ensure model validity, additional analyses are conducted, including multicollinearity testing using the Variance Inflation Factor (VIF) and regression assumption testing through residual analysis to assess normality and homoscedasticity. Furthermore, feature importance analysis is performed using the Random Forest model to identify the relative contribution of predictor variables. To assess model robustness, a cross-dataset validation is also conducted by applying the same modeling framework to a secondary dataset obtained from a private primary school, enabling comparison of model performance across datasets with different sizes and characteristics.

RESULT

This study aims to predict students' academic performance represented by the average score calculated from mathematics, reading, and writing scores. Two machine learning approaches are compared: Linear Regression as the baseline model and Random Forest Regression as a non-linear ensemble model. Model performance is evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), the coefficient of determination (R^2), and relative error percentage. The models are trained using an 80:20 train–test split. In addition to predictive accuracy, diagnostic analyses including multicollinearity testing, regression assumption testing, and feature importance analysis are conducted to ensure model validity and interpretability. To provide additional insight into prediction accuracy, relative error percentages are calculated based on the ratio between MAE and the mean value of the target variable. The results indicate relative errors of approximately 15.6% for Linear Regression, 17.0% for the default Random Forest, and 15.9% for the tuned Random Forest model.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15953
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2. Performance Comparison of Regression Models

Model	MAE	RMSE	R^2	Relative Error (%)
Linear Regression	10.49	13.40	0.162	15.6
Random Forest (Default)	11.38	14.64	0.000	17.0
Random Forest (Tuned)	10.70	13.80	0.112	15.9

Based on Table 2, Linear Regression achieves the highest predictive performance with an R^2 value of 0.162, indicating that the model explains approximately 16.2% of the variance in students' academic performance. Although this value appears relatively low, it reflects the limited predictive capability of the demographic variables included in the dataset. Academic performance is influenced by many complex factors such as cognitive ability, learning motivation, and socio-economic background, which are not represented in the available features. The default Random Forest model shows very poor performance with an R^2 value close to zero, suggesting that the model fails to capture meaningful patterns in the dataset. After hyperparameter tuning using GridSearchCV with 5-fold cross-validation, the Random Forest model improves to an R^2 value of 0.112, but still remains inferior to Linear Regression. This result suggests that the relationships within the dataset are predominantly linear and that increasing model complexity does not necessarily improve predictive performance when the available features are limited.

Model Significance Test

To evaluate whether the regression model provides statistically meaningful predictions, a significance test is conducted using the F-test in the Linear Regression model. The statistical test examines whether the predictor variables collectively contribute to explaining variations in the target variable. The results indicate that the regression model is statistically significant ($p < 0.05$), suggesting that the predictor variables jointly contribute to the prediction of students' academic performance, although the overall explanatory power remains limited.

Multicollinearity Analysis

This study evaluates multicollinearity among predictor variables using the Variance Inflation Factor (VIF). Multicollinearity occurs when independent variables are highly correlated, potentially affecting the stability and interpretability of regression coefficients.

The VIF results indicate that all predictor variables have very low VIF values, ranging from approximately 0.003 to 0.021. These values are significantly below the commonly accepted threshold of 10, which indicates the absence of problematic multicollinearity. The results confirm that the predictor variables used in the dataset are sufficiently independent and suitable for regression modeling. This condition ensures that the Linear Regression model can estimate coefficients reliably without being affected by strong correlations among predictors.

Table 3. Multicollinearity Test Using VIF

Feature	VIF
gender_male	0.0038
race/ethnicity_group B	0.0097
race/ethnicity_group C	0.0074
race/ethnicity_group D	0.0087
race/ethnicity_group E	0.0115
parental education (bachelor)	0.0117
parental education (high school)	0.0087
parental education (master)	0.0206
parental education (some college)	0.0079
parental education (some high school)	0.0087
lunch_standard	0.0041
test preparation course_none	0.0040

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15953
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Regression Assumption Testing

To evaluate the validity of the Linear Regression model, residual analysis is conducted to examine the distribution and behavior of prediction errors. Figure 2 illustrates the distribution of residuals, which appear approximately symmetric and centered around zero. This pattern indicates that the residuals follow a near-normal distribution, suggesting that the normality assumption of the regression model is reasonably satisfied. Figure 3 presents the residual scatter plot between predicted values and residuals. The points are randomly dispersed around the horizontal zero line without forming systematic patterns. This indicates that the model satisfies the homoscedasticity assumption, meaning that the variance of residuals remains relatively constant across predicted values. These findings suggest that the regression model assumptions are adequately satisfied, supporting the reliability of the Linear Regression model used in this study.

In addition, the residual distribution does not exhibit strong skewness or extreme outliers, indicating that prediction errors are relatively balanced across observations. The absence of systematic structures in the residual plot further confirms that the linear model does not suffer from major specification errors. These results strengthen the reliability of the regression model and indicate that the Linear Regression approach is appropriate for modeling the relationships present in the dataset.

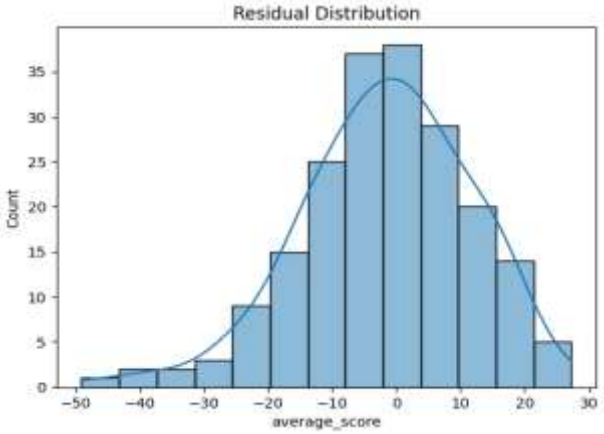

Fig. 2 Residual Distribution

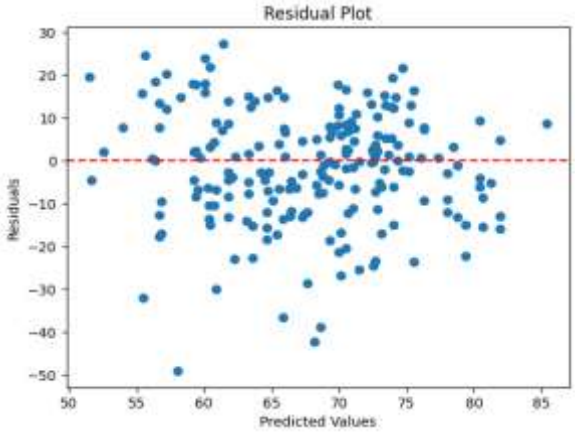

Fig. 3 Residual Plot

Feature Importance Analysis

To further understand the contribution of predictor variables in predicting students' academic performance, feature importance analysis is conducted using the Random Forest model. Figure 4 shows the relative importance of each predictor variable in the model. The results indicate that lunch type and test preparation course are among the most influential variables in predicting students' academic performance. Other demographic variables such as gender, race/ethnicity, and parental education level demonstrate lower contributions.

These findings suggest that factors related to learning preparation and educational support may play a more significant role in academic outcomes than demographic characteristics alone. However, the relatively moderate importance values across variables also indicate that the available predictors provide limited explanatory power for academic performance.

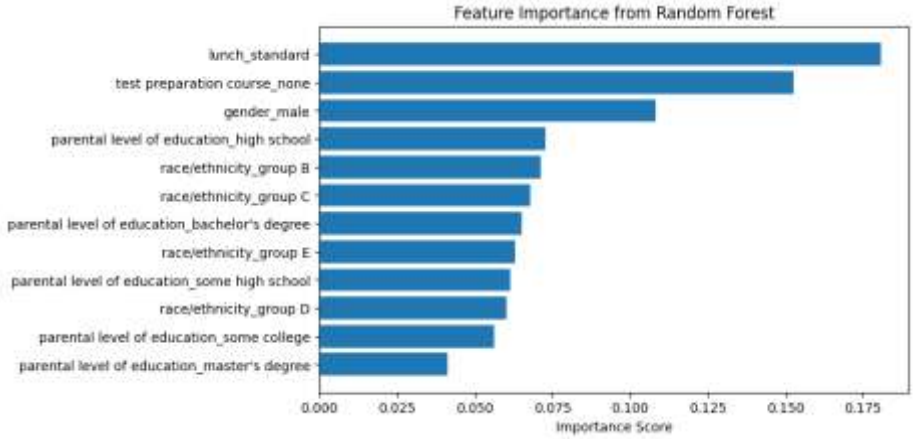

Fig. 4 Feature Importance of Predictor Variables

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15953
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Bias Variance Evaluation

Model performance is also interpreted from a bias–variance perspective. The relatively low R^2 values suggest that both models exhibit high bias, meaning that the models are limited in capturing the underlying complexity of academic performance prediction using the available features. However, the consistent performance of Linear Regression across experiments indicates that the model maintains stable variance and does not overfit the training data. In contrast, Random Forest, which is typically capable of modeling complex non-linear relationships, does not demonstrate superior performance in this study due to the limited number of features and the relatively simple structure of the dataset. This finding suggests that model complexity alone does not guarantee improved predictive accuracy when the available predictors provide limited explanatory information.

Model Performance Comparison

Figures 5 and 6 illustrate the performance comparison between Linear Regression and Random Forest models. Linear Regression achieves the highest coefficient of determination with an R^2 value of 0.162, indicating that approximately 16.2% of the variance in students' academic performance can be explained by the predictor variables used in the model. In contrast, the default Random Forest model produces an R^2 value close to zero, suggesting that the model fails to capture meaningful relationships between the input variables and the target variable. After hyperparameter tuning using GridSearchCV with 5-fold cross-validation, the performance of Random Forest improves to $R^2 = 0.112$, accompanied by reductions in MAE and RMSE. Although the tuned Random Forest model shows improvement compared to its default configuration, its performance remains inferior to Linear Regression. This finding indicates that the relationships within the dataset are likely predominantly linear and that increasing model complexity does not necessarily lead to better predictive performance.

The relatively low R^2 values observed in this study suggest that the available demographic and contextual variables provide limited predictive power for academic performance. This result highlights the need for incorporating additional features such as learning behavior, cognitive ability, or socio-economic indicators to improve model performance in educational data mining tasks.

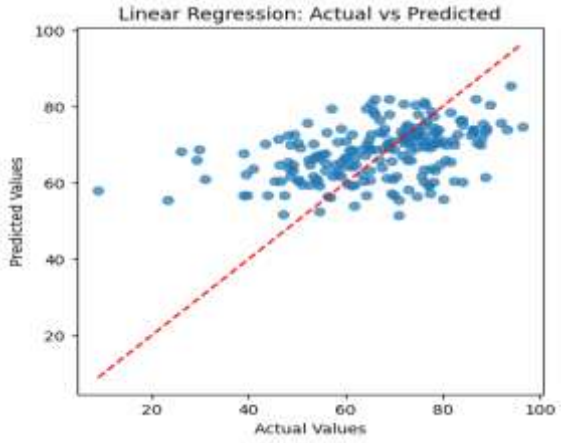

Fig. 5 Actual vs Predicted Linear Regression

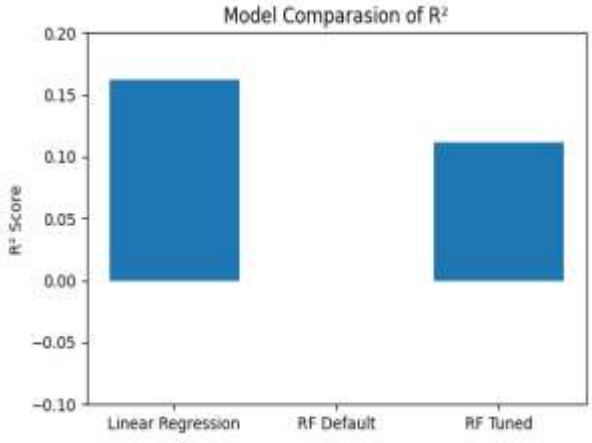

Fig. 6 Comparison of R^2 Values Across Models

DISCUSSIONS

The results demonstrate that Linear Regression provides the most consistent predictive performance compared to Random Forest Regression. The obtained coefficient of determination ($R^2 = 0.162$) indicates that the demographic and contextual variables included in the dataset explain approximately 16.2% of the variance in students' academic performance. Rather than indicating model failure, this relatively low explanatory power reflects the limited predictive capability of the available features. Academic performance is influenced by complex factors such as cognitive ability, learning motivation, learning behavior, and socio-economic background, many of which are not represented in the dataset used in this study.

The inferior performance of Random Forest in this study can be explained by several structural characteristics of the dataset. Random Forest is generally effective when datasets contain large numbers of observations, diverse features, and complex non-linear relationships among variables. However, the dataset used in this study contains a limited number of predictor variables that are primarily demographic in nature. Such variables often exhibit relatively simple or weak relationships with academic performance, which reduces the advantage of complex ensemble models. In this situation, simpler models such as Linear Regression may produce more stable predictions because they are less sensitive to noise and overfitting in low-dimensional datasets.

*name of corresponding author

The feature importance analysis further supports this interpretation. The results indicate that variables related to learning conditions, particularly lunch type and participation in test preparation courses, contribute more strongly to the prediction of academic performance than demographic variables such as gender or parental education level. Nevertheless, the overall importance values of these predictors remain relatively moderate, suggesting that the available features capture only a limited portion of the factors influencing students' academic outcomes. This finding reinforces the need to incorporate additional behavioral and cognitive variables in future educational data mining studies.The findings of this study are consistent with several previous studies that emphasize the importance of data characteristics in determining the effectiveness of predictive models in educational contexts. Studies by (Ling et al., 2024; Qureshi & Lokhande, 2024) reported that the performance of machine learning models for academic performance prediction is highly dependent on dataset size, feature relevance, and variability, rather than model complexity alone. Similarly, (Abro et al., 2025) demonstrated that simpler regression-based approaches can provide competitive performance when applied to primary education data with limited features.

In contrast to studies that reported superior performance of complex models such as Random Forest in large-scale or high-dimensional educational datasets (Begum & Padmannavar, 2023; Xu & Hoang, 2021). From a practical perspective, these findings provide important implications for educational institutions and policymakers. The limited predictive power of demographic variables suggests that relying solely on basic student profile data may not be sufficient for accurately identifying students at risk of academic difficulties. Educational institutions should consider integrating additional data sources, such as learning behavior indicators, classroom engagement metrics, and formative assessment results, to support more effective predictive analytics. By incorporating richer educational data, predictive models may provide more accurate insights that can assist teachers and administrators in designing early interventions to improve student learning outcomes.

Case Study

As an additional validation of the main findings of this study, a case study was conducted using a local dataset obtained from a private primary school. The dataset consists of 132 sixth-grade students with historical academic records collected over a three-year period, from grade 4 to the first semester of grade 6. The dataset includes four main predictor variables, namely mathematics score, Indonesian language score, science score, and student attendance records. These variables represent academic achievement and learning participation indicators commonly available in primary school administrative data. To construct the prediction target, a readiness score was calculated by aggregating students' academic performance across the observed period. This dataset provides a simple but realistic representation of small-scale educational data typically available at the school level.

Prior to model development, several preprocessing steps were conducted to ensure data consistency. These steps included data cleaning to address incomplete records, normalization of academic score variables to maintain comparable scales, and verification of attendance data to remove inconsistencies. Since the dataset consists primarily of numerical variables, no categorical encoding was required. These preprocessing procedures ensure that the dataset is suitable for regression-based predictive modeling.

The case study is evaluated using the same modeling framework as the main experiment by comparing Linear Regression and Random Forest Regression in predicting students' readiness scores. Model performance is assessed using regression metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). To improve the robustness of the evaluation, 5-fold cross-validation is applied during model training, allowing the models to be tested across multiple data partitions and reducing the risk of evaluation bias due to a single train–test split.

The results show that Linear Regression achieves superior predictive performance with $R^2 = 0.509$, RMSE = 6.78, and MAE = 4.89, while Random Forest Regression demonstrates lower performance with $R^2 = 0.171$, RMSE = 8.81, and MAE = 7.27. These results indicate that Linear Regression is more effective in capturing the relationships among variables within the local dataset. The stronger performance of Linear Regression suggests that the relationships among the available predictors and the readiness score are predominantly linear, making simpler regression models more suitable for this dataset.

The cross-dataset comparison between the public dataset and the local school dataset provides additional insights into the generalizability of the findings. Despite differences in dataset size and feature composition, both experiments consistently demonstrate that Linear Regression outperforms Random Forest for datasets with relatively simple structures and limited features. This consistency suggests that simpler regression models may provide more stable predictive performance when applied to small-scale educational datasets commonly available in primary school environments.

CONCLUSION

This study compares the performance of Linear Regression and Random Forest Regression in predicting students' academic performance using an educational data mining approach. The experimental results show that

*name of corresponding author

Linear Regression achieves the best performance, with a coefficient of determination (R^2) value of 0.162, while Random Forest Regression with default parameters exhibits very poor performance. After hyperparameter tuning using GridSearchCV with 5-fold cross-validation, the performance of Random Forest improves to an R^2 value of 0.112, but it still fails to outperform Linear Regression.

These findings indicate that the relationship between the factors used in this study and students' academic performance tends to be linear, making linear models more suitable than high-complexity non-linear models. This study also confirms that increasing algorithmic complexity does not necessarily correspond to improved predictive performance, particularly when the features employed have limited predictive power. Therefore, the selection of academic performance prediction models should consider data characteristics and analytical objectives rather than relying solely on the complexity of machine learning methods.

Beyond the empirical findings, this study provides several theoretical contributions to the field of Educational Data Mining. First, the results demonstrate that model effectiveness in academic performance prediction is strongly influenced by dataset characteristics rather than algorithmic complexity alone. The findings highlight that simpler regression-based models may provide more reliable predictions when applied to educational datasets with limited features and relatively linear relationships. Second, this study contributes empirical evidence from both a public educational dataset and a real-world primary school dataset, offering cross-dataset validation of model behavior in different data environments. This dual-dataset evaluation represents a methodological contribution by demonstrating how predictive models behave across datasets with different sizes and feature structures. The novelty of this research lies in systematically comparing linear and non-linear predictive models in the context of small-scale primary education datasets, a context that has received limited attention in previous educational data mining studies.

Future research should extend this work by incorporating richer educational features such as learning behavior indicators, cognitive ability measures, classroom engagement metrics, and socio-economic variables to improve predictive performance. Methodologically, future studies may explore hybrid modeling approaches that combine interpretable regression models with advanced machine learning techniques. In addition, expanding cross-dataset validation across multiple schools and educational contexts would allow researchers to better evaluate model generalizability and robustness. These directions may contribute to the development of more reliable predictive analytics systems for supporting data-driven decision making in educational institutions.

Overall, this study emphasizes that selecting predictive models in educational data mining should be guided not only by algorithmic sophistication but also by the structural characteristics of the available educational datasets.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the primary school that provided access to the local dataset used in the case study

REFERENCES

Abro, M., Husain, I., Hassan Zaidi, S. M., Sheikh, F., & Murtaza, G. (2025). A Predictive Model and Performance Evaluation in Mathematics for Primary Education. *Journal of Computing and Biomedical Informatics*, *9*(2). https://www.scopus.com/inward/record.uri?eid=2-s2.0-105027171524&partnerID=40&md5=18424cc099eb6457a06e345b46e99f6a

Ali, J. A., Abdi, M. K., Ali, T. A., Muse, A. H., & Cumar, M. A. (2025). Geographic and school-level disparities as primary predictors of numeracy skills: A supervised machine learning approach of Somaliland's national learning assessment. *Social Sciences and Humanities Open*, *12*(July), 102305. https://doi.org/10.1016/j.ssaho.2025.102305

Begum, S., & Padmannavar, S. S. (2023). Student Performance Analysis using Bayesian Optimized Random Forest Classifier and KNN. *International Journal of Engineering Trends and Technology*, *71*(5), 132–140. https://doi.org/10.14445/22315381/IJETT-V71I5P213

Bulut, O., Tan, B., Mazzullo, E., & Syed, A. (2025). Benchmarking Variants of Recursive Feature Elimination: Insights from Predictive Tasks in Education and Healthcare. *Information (Switzerland)*, *16*(6), 1–21. https://doi.org/10.3390/info16060476

Bussaman, S., Nasa-Ngium, P., Nuankaew, W. S., Sararat, T., & Nuankaew, P. (2024). Ensemble Learning Approaches to Strategically Shaping Learner Achievement in Thailand Higher Education. *Lecture Notes in Electrical Engineering*, *1258*, 329 – 339. https://doi.org/10.1007/978-981-97-7356-5_27

Deleña, R. D., Dia, N. J., Sacayan, R. R., Sieras, J. C., Khalid, S. A., Macatotong, A. H. T., & Gulam, S. B. (2025). Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic

*name of corresponding author

and academic factors. *Systems and Soft Computing*, *7*(June). https://doi.org/10.1016/j.sasc.2025.200352

Eriksson, M., Malefors, C., Secondi, L., & Marchetti, S. (2021). Guest attendance data from 34 Swedish pre-schools and primary schools. *Data in Brief*, *36*, 107138. https://doi.org/10.1016/j.dib.2021.107138

Hegde, V., Abhinav, M. R., & Roshin, C. (2023). Predicting Student Placement using PCA and Machine Learning Technique. *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*. https://doi.org/10.1109/ICCCNT56998.2023.10307185

Jabir, B., Hamzaoui, R., Rahali, E. A., & Falih, N. (2025). A machine learning framework for early intervention in e-learning environments. *EDPACS*, *70*(12), 53 – 66. https://doi.org/10.1080/07366981.2025.2515737

Kostopoulos, G., Tsiakmaki, M., & Kotsiantis, S. (2026). Benchmarking Statistical and Deep Generative Models for Privacy-Preserving Synthetic Student Data in Educational Data Mining. *Algorithms*, *19*(1), 39. https://doi.org/10.3390/a19010039

Ling, N. Y., Tin, T. T., Keat, T. C., Khattak, U. F., & Almaiah, M. A. (2024). Educational Big Data Analytics: Machine Learning Based Academic Performance Predictive Modelling. *Pakistan Journal of Life and Social Sciences*, *22*(2), 7442–7477. https://doi.org/10.57239/PJLSS-2024-22.2.00562

Lyu, H., & Xu, K. (2025). A SYSTEMATIC REVIEW OF AI-DRIVEN ANALYTICS IN EDUCATION: MAPPING THE EVIDENCE FOR PREDICTING AND ENHANCING STUDENT SUCCESS. *Journal of Environmental Protection and Ecology*, *26*(7), 2767 – 2778. https://www.scopus.com/inward/record.uri?eid=2-s2.0-105025931874&partnerID=40&md5=a658789dad60ef69a485203a77bbe7b0

Nugraha, F. M., Dewi, K. K., Gunawan, A. A. S., & Tedjasulaksana, J. J. (2025). Leveraging Regression-Based Machine Learning for Predicting Middle School Student Passing Grades. *2025 IEEE International Conference on Artificial Intelligence and Mechatronics Systems, AIMS 2025*. https://doi.org/10.1109/AIMS66189.2025.11229636

Poh, Z. X., & Khor, E. T. (2024). Predictive Analytics for Student Online Learning Performance Using Machine Learning and Data Mining Techniques. *International Journal on E-Learning: Corporate, Government, Healthcare, and Higher Education*, *23*(3), 269 – 283. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85214129363&partnerID=40&md5=95d98e3926257f20d0c1079327255cd4

Qureshi, R., & Lokhande, P. S. (2024). A Comprehensive Review of Machine Learning techniques used for Designing An Academic Result Predictor And Identifying The Multi-Dimensional Factors Affecting Student's Academic Results. *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry, IDICAIEI 2024*. https://doi.org/10.1109/IDICAIEI61867.2024.10842901

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), e1355. https://doi.org/https://doi.org/10.1002/widm.1355

Soares, W. L., Pereira De Carvalho, H. D., Santos, W. B., & Andrade De A. Fagundes, R. (2022). Regression models based in optimized Ensemble of Extreme Learning Machine Networks. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, *2022-Octob*, 1140 – 1146. https://doi.org/10.1109/SMC53654.2022.9945088

Thaher, T., & Jayousi, R. (2020). Prediction of Student's Academic Performance using Feedforward Neural Network Augmented with Stochastic Trainers. *14th IEEE International Conference on Application of Information and Communication Technologies, AICT 2020 - Proceedings*. https://doi.org/10.1109/AICT50176.2020.9368820

Xu, W., & Hoang, V. T. (2021). MapReduce-Based Improved Random Forest Model for Massive Educational Data Processing and Classification. *Mobile Networks and Applications*, *26*(1), 191 – 199. https://doi.org/10.1007/s11036-020-01699-w

*name of corresponding author