Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 6 Iss. 2 April 2026, pp: 494-503
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Predicting Students' Mathematics Scores from Reading Scores Using Supervised Learning

Nofita Fitriyani

Department of Informatics Engineering, Telkom University Purwokerto, Indonesia

E-Mail: nofitafitriani0@gmail.com

Received Jan 01st 2026; Revised Feb 21th 2026; Accepted Mar 01st 2026; Available Online Apr 19th 2026
Corresponding Author: Nofita Fitriyani

Abstract

This study aims to predict students' mathematics scores based on their reading scores using a supervised learning approach. The dataset used is from Students' Performance in Exams (Kaggle), consisting of 1,000 student records, and was analyzed using Microsoft Excel and Google Colaboratory. The data was divided into training and test data with a ratio of 80:20. The research stages included descriptive statistical analysis, data visualization, Pearson correlation testing, linear regression model development, and model performance evaluation using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). Prior to modeling, regression assumptions including linearity, normality of residuals, and homoscedasticity were examined to ensure model validity. The results showed a strong positive relationship between reading and math scores with a correlation coefficient of 0.818. The linear regression model produced an MAE of 7.281, an RMSE of 8.818, and an R^2 of 0.680. Decision Tree Regressor was selected as a comparison model because it represents a non-linear and non-parametric supervised learning approach commonly used in educational data mining. This study contributes to educational data mining literature by demonstrating that interpretable regression models explain significant mathematics achievement variance, rivaling the performance of non-linear alternatives.

Keywords: Mathematics, Reading Scores, Score Prediction, Students, Supervised Learning

1. INTRODUCTION

Literacy skills, especially reading skills, play a fundamental role in students' academic success in various subjects. Reading is not merely the ability to decode text, but also involves comprehension, interpretation, reasoning, and the application of information in the context of problem solving. In mathematics education, reading literacy is particularly important when students are required to understand word problems, interpret instructions, and translate verbal information into mathematical representations. A number of studies have confirmed a significant relationship between reading comprehension and students' mathematical problem-solving abilities [1], [2], [3], [4]. However, many of these studies focus more on correlation analysis or class-based experimental designs rather than predictive modeling approaches. As a result, although the relationship between reading literacy and math performance has been recognized, its measurable predictive contribution has not been widely explored in the framework of supervised learning.

Mathematics is often perceived as a discipline dominated by logical and numerical reasoning skills. However, understanding mathematical problems, especially word problems, requires strong reading comprehension skills. Research shows that students with better reading literacy tend to perform better in solving mathematical word problems and understanding abstract mathematical concepts [1], [2]. Meta-analysis findings also confirm a consistent correlation between reading comprehension and mathematical problem-solving skills among students in Indonesia [4].

However, previous studies generally combine multiple cognitive, demographic, and socioeconomic variables, thereby obscuring the independent influence of reading ability on mathematical achievement. For example, several academic prediction studies combine learning habits, attendance, lifestyle, or e-learning activity data to improve model performance [5], [6], making it difficult to quantify the direct statistical contribution of reading ability alone.

Advances in information technology have enabled the use of educational data to analyze and predict student academic achievement using machine learning. Recent studies have applied regression-based and supervised learning models to predict academic performance using academic, behavioral, and demographic

variables [7], [8], [9], [10]. For example, Zhang and Cutumisu [8] applied machine learning techniques to predict mathematical literacy, while Patil et al [7] evaluated regression-based machine learning models for predicting student performance. Bhutto et al. [9] demonstrated the effectiveness of supervised learning algorithms in modeling academic achievement.

In addition, Kassy [11], Rismaya et al. [6] and Athallah et al. [12] show that linear regression is still widely used in academic performance prediction due to its computational simplicity and interpretability, especially when transparency is prioritized over algorithm complexity.

Although these studies report promising prediction accuracy, many of them use numerous predictors and relatively complex algorithms, including tree-based models such as Decision Tree [5]. While this approach can improve prediction performance, it often reduces model interpretability and limits transparency in explaining how each predictor affects the outcome variable. Guevara-Reyes et al [10] emphasize that interpretability is crucial in the context of educational decision-making, as stakeholders require clear and easily understandable explanations, not just high accuracy metrics.

However, in many predictive studies, the explanatory role of individual academic competencies such as reading ability has not been explicitly quantified. Furthermore, comparative studies between linear regression and tree-based methods often focus on performance metrics without discussing the practical implications of model transparency in the educational context [5], [7].

Therefore, a focused, interpretive, and statistically measurable modeling approach is needed to clearly identify the proportion of variance in mathematics achievement explained by reading ability. Unlike previous studies that emphasized model complexity and multi-variable optimization, this study adopted a linear regression framework with a single predictor to isolate and quantify the statistical relationship between Reading Scores and Math Scores.

By emphasizing transparency and interpretability, this study aims to show that simple supervised learning models can still provide meaningful predictive performance while providing clear insights for educators and policymakers. The objectives of this study are: (1) to analyze the characteristics and statistical relationship between Reading Scores and Math Scores, and (2) to evaluate the effectiveness of linear regression in predicting Math Scores based on Reading Scores. The results of this study are expected to contribute to the application of interpretive machine learning in educational data analysis and support literacy-based strategies to improve student achievement in mathematics.

2. RESEARCH METHODOLOGY

This research methodology was systematically developed to predict students' Math Scores based on their Reading Scores using a supervised learning approach with linear regression. The research flow refers to the stages of data analysis and predictive modeling, as shown in Figure 1 for Research Methodology.

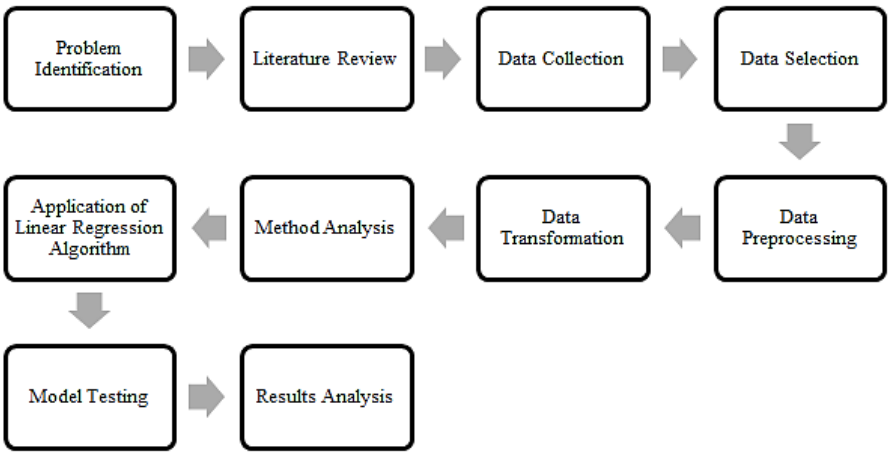

Figure 1. Research Methodology

2.1. Problem Identification

The use of educational data to analyze and predict student academic achievement has been widely practiced with the development of technology and machine learning approaches. However, in practice, student grade data is often only used as academic reports without further analysis to explore patterns of relationships between learning variables. The Students' Performance in Exams dataset contains various student assessment variables, including Reading and Math scores, that could provide insight into the relationship between reading literacy skills and math achievement.

Although several studies have discussed the prediction of academic achievement using various supervised learning algorithms, studies that specifically model the relationship between Reading and Math scores using a simple, interpretive, and easy-to-understand approach are still relatively limited. The available data has not been optimally utilized to explain the extent to which reading ability contributes to students' Math scores quantitatively. Therefore, an analysis method is needed that can model the linear relationship between variables clearly and precisely. In this study, a linear regression model was used to predict Math scores from students' Reading scores.

2.2. Literature Review

A literature review was conducted as a theoretical basis to support this study. The literature reviewed focused on the concepts of reading literacy, supervised learning, linear regression, and the application of machine learning in education. Several previous studies have shown that reading ability has a significant relationship with students' ability to understand story problems and mathematical concepts, thereby impacting mathematics learning outcomes [1], [4], [13], [14].

In addition, linear regression is widely used in academic achievement prediction studies because it has low computational complexity, is transparent, and is easy to interpret. This approach allows researchers and education practitioners to understand the direct contribution of input variables to output variables. Referring to this literature review, this study uses linear regression as the main method for modeling the relationship between students' reading and math scores [11], [12].

2.3. Data Collection

The data collection stage is the first step in research that focuses on data analysis and modeling. The dataset used in this study was obtained from the Kaggle platform with the title Students Performance in Exams. A total of 1,000 student data were collected and are quantitative in nature. Each data point contains information on students' academic scores, including Reading and Math scores, which serve as the main variables in this study. These scores range from 0 to 100 and represent student academic achievement in each subject. This dataset is then used as the basis for descriptive analysis, regression modeling, and model performance evaluation [15].

2.4. Data Selection

The data selection stage is carried out to ensure that only data relevant to the research objectives is used in the analysis process. From all the attributes available in the Students Performance in Exams dataset, this study only utilizes two main variables, namely Reading Score as the input variable (X) and Math Score as the output variable (Y). This selection process produces a more focused and appropriate data subset for regression-based supervised learning analysis.

2.5. Data Preprocessing

Data Preprocessing was conducted to ensure data quality, consistency, and suitability before the modeling stage, as data quality significantly influences machine learning performance [16], [17], [18]. The preprocessing procedure was carried out systematically using Microsoft Excel and the Python environment in Google Colaboratory. The preprocessing steps included:

1. Duplicate Removal
   The dataset was inspected to identify potential duplicate records. Duplicate checking was performed to maintain data integrity before analysis.

2. Missing Value Handling
   All selected variables were examined to detect incomplete or null values. If missing values were identified, appropriate handling techniques such as imputation or record removal would be applied [19].

3. Data Type and Range Validation
   The numerical data types of the selected variables were verified to ensure compatibility with regression modeling. Range validation was also performed to confirm that all values were within the expected score interval.

4. Outlier Detection
   Outlier screening was conducted using descriptive statistical analysis and visualization techniques such as boxplots. In addition, standardized score (z-score) analysis was used to identify potential extreme observations that could distort regression results [19].

2.6. Data Transformation

Data transformation was performed to adjust the data format so that it could be processed properly at the modeling stage. In this study, the Reading and Math scores were already numerical, so there was no need to transform the categorical data.

However, to ensure consistency in the analysis process and to facilitate statistical calculations and visualization, the data were reorganized into a structured format in both Microsoft Excel and Google Colaboratory. In addition, the dataset was divided into training data and test data with a ratio of 80% for training data and 20% for test data. This transformation was intended to support the objective evaluation of the model and avoid bias in the prediction results [20], [21].

2.7. Method Analysis

Linear regression is a supervised learning method used to model linear relationships between input variables and output variables based on labeled data. This method aims to build a prediction model capable of quantitatively estimating the value of dependent variables. In this study, linear regression was used to predict Mathematics scores (Math Score) based on Reading scores (Reading Score).

The application of the linear regression method was carried out systematically and structurally, as shown in Figure 2, which presents the analysis process flow from variable determination to prediction result evaluation.

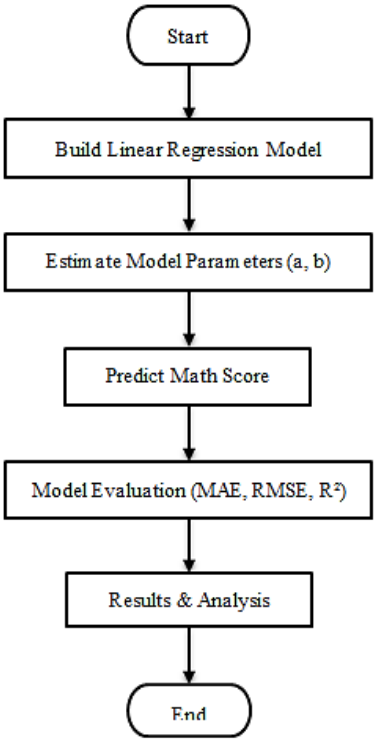

Figure 2. Flowchart of Linear Regression Method

Here is an explanation of the flowchart above:

1. Start
   This stage marks the beginning of the linear regression analysis process. The dataset used in this stage has undergone data collection, variable selection, preprocessing, and training and testing data division as described in the previous section.

2. Build Linear Regression Model
   At this stage, a simple linear regression model was constructed to model the relationship between the independent variable Reading Score and the dependent variable Math Score. The linear regression model is expressed in the following Equation 1.

$$y = a + bx \tag{1}$$

   where y is the predicted Math Score value, x is the Reading Score, a is the constant (intercept), and b is the regression coefficient.

3. Estimate Model Parameters (a,b)
   The regression parameters a and b are calculated using the least squares method with the aim of minimizing the difference between the actual and predicted values. The regression coefficient b is calculated using Equation 2.

$$b = \frac{\left(\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \right)}{\left(\Sigma_{i=1}^{n}(x_i - \bar{x})^2 \right)} \quad (2)$$

   Next, the constant value a is calculated using Equation 3.

$$a = \bar{y} - b\,\bar{x} \quad (3)$$

   Where x_i and y_i are the data pairs for i, $\bar{x}$ and $\bar{y}$ are the average values for Reading Score and Math Score, and n is the number of training data.

4. Predict Math Score
   After the regression parameters were obtained, the model was used to predict the Math Score values in the test data. The prediction process was carried out by entering the Reading Score values into the regression Equation 4.

$$\hat{y}_i = a + b\,x_i \quad (4)$$

   Where $\hat{y}_i$ is the predicted Math Score value for the data at i.
5. Model Evaluation (MAE, RMSE, R^2)
   Model performance is evaluated by comparing predicted and actual values using three metrics: MAE, RMSE, and R^2.
   Mean Absolute Error (MAE) is calculated using Equation 5:

$$MAE = \left(\frac{1}{n}\right) \Sigma_{i=1}^{n} | y_i - \hat{y}_i | \quad (5)$$

   Root Mean Squared Error (RMSE) is calculated using the Equation 6.

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (6)$$

   The coefficient of determination (R^2) is calculated using the equation:

$$R^2 = 1 - \left(\frac{\Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n} (y_i - \bar{y})^2} \right) \quad (7)$$

6. Results & Analysis
   The evaluation results were used to analyze the performance of the linear regression model. The MAE and RMSE values indicate the level of prediction error, while the R^2 value describes the model's ability to explain the variation in Math Score based on Reading Score.

7. End
   This stage marks the end of the linear regression analysis process. The results obtained are then used as the basis for discussion and drawing conclusions from the research.

The application of linear regression in this study was carried out through the stages of dividing the training data and test data, as well as evaluating the model performance using the MAE, RMSE, and coefficient of determination (R^2) metrics [22]. This approach shows that linear regression is a quantitative statistical method used to objectively model the relationship between independent and dependent variables.

The similarity between this study and previous studies lies in the use of linear regression as the main predictive model in analyzing students' mathematics scores. This study applies linear regression with the numeric variable of reading scores as predictors of math scores, thus being methodologically consistent with previous studies in terms of modeling, data training and testing processes, and model performance

evaluation. This study reinforces the use of linear regression as a relevant and consistent method in analyzing the relationship between academic variables in the field of education.

2.8. Regression Assumption Testing

Before interpreting the regression results, basic regression assumptions were examined through exploratory visualization and residual analysis. Linearity between Reading Score and Math Score was evaluated using a scatter plot to observe whether the relationship followed a linear pattern. To examine model validity, residual analysis was conducted by plotting prediction errors against predicted values. This residual plot was used to assess whether errors were randomly distributed around zero, indicating the absence of systematic bias and suggesting homoscedasticity. Additionally, statistical significance testing was performed using the Ordinary [23], [24].

2.9. Model Configuration

Two supervised learning models were implemented: Linear Regression and Decision Tree Regressor. The Linear Regression model was built using the default scikit-learn configuration. The dataset was split into 80% training data and 20% testing data using train_test_split with random_state = 42 to ensure reproducibility. The Decision Tree Regressor was implemented as a comparative model using default parameters with random_state = 42. No additional hyperparameter tuning, such as max_depth or pruning optimization, was applied, allowing a fair baseline comparison with the linear regression model. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). Five-fold cross-validation was also conducted to assess model stability.

3. RESULTS AND DISCUSSION

This chapter presents the results of applying research methods and analysis in predicting students' Mathematics scores based on their Reading scores using a supervised learning approach. The research stages are described in a structured manner, including data requirements analysis, data selection and preprocessing, data transformation, application of Linear Regression and Decision Tree Regression algorithms, as well as testing and analysis of results. This discussion aims to evaluate the model's performance and identify the relationship between reading scores and math scores as a basis for drawing research conclusions.

3.1 Data Requirements Analysis

A data requirements analysis was conducted to determine the type of data needed to build a model to predict students' mathematics scores. This study used students' academic data, which was numerical in nature and sourced from a secondary dataset obtained through the Kaggle platform. The data used consists of two main attributes, namely Reading Score as an independent variable and Math Score as a dependent variable. The selection of the Reading Score is based on its correlation with the ability to understand questions, while the Math Score is used as the prediction target. Details of the data requirements used in this study are presented in Table 1 as a basis for the next stage of data processing.

Table 1. Data Requirements Analysis

No	Data Attribute	Data Type	Role	Description
1	Reading Score	Numerical	Independent Variable	Student reading ability score used as a predictor
2	Math Score	Numerical	Dependent Variable	Student mathematics score as prediction target

3.2 Data Selection

Based on the data selection process described in Chapter 2, this study uses two numerical variables, namely reading score and math score. Of the total 1000 student data, all data met the selection criteria and were used in the next stage of analysis. The final dataset used is shown in Table 2.

3.3 Data Preprocessing

The preprocessing stage was conducted to ensure data quality before model development. The dataset consisted of 1,000 student records, from which Reading Score and Math Score were selected for analysis. The duplicate inspection confirmed that no duplicate records were found. Missing value examination indicated that both variables contained no null entries; therefore, no imputation was required. Data type validation verified that both variables were numerical and within the valid score range of 0–100. Outlier screening using boxplots and z-score analysis did not identify any extreme values requiring removal. Overall, the dataset was considered complete, consistent, and suitable for subsequent regression and supervised learning analysis.

Table 2. Dataset After Data Selection Process

No	Reading Score	Math Score
1	72	70
2	90	88
3	85	84
…	…	…
1000	68	65

3.4 Data Transformation

The dataset was split into training (80%) and test (20%) sets using a train-test split with random_state = 42, yielding 800 training and 200 test samples for modeling and testing. The entire process was carried out using the Python programming language in Google Colab.

3.5 Application of Linear Regression Methods

Linear regression models were used to examine the relationship between reading scores (independent variables) and math scores (dependent variables). The training process was carried out using training data (80%) with Python programming language implementation and scikit-learn library on Google Colab. The training results produced the following linear regression equation:

$$Y = 0{,}8465X + b \tag{8}$$

The regression coefficient value shows that every 1-point increase in reading score is followed by an increase in math score of ±0.85 points. These results indicate a positive linear relationship between students' reading and math abilities, as shown in Figure 3.

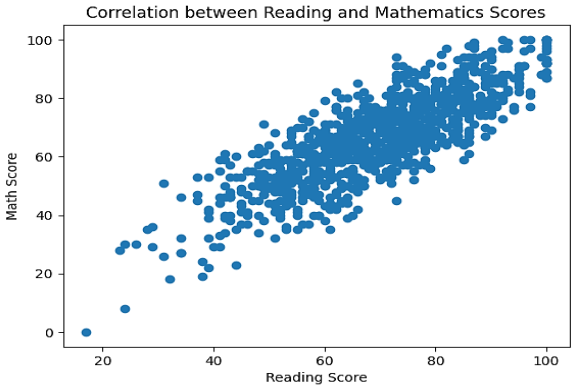

Figure 3. Scatter Plot Reading Score and Math Score

3.6 Testing

The testing was conducted using test data (20%) to evaluate the model's ability to predict students' math scores based on their reading scores. The model's performance was evaluated using the MAE, RMSE, and R^2 metrics. For comparison, testing was conducted using the Decision Tree Regressor model. A comparison of the performance of the two models is shown in Table 3, which shows that the linear regression model produces more accurate and stable predictions than the Decision Tree.

Table 3. Comparison of Model Evaluation Results

Model	Dataset	MAE	RMSE	R^2
Linear Regression	Testing	7.280882	8.818137	0.680447
Linear Regression	Training	6.999839	8.704344	0.663932
Decision Tree	Testing	7.567415	9.129746	0.657464
Decision Tree	Training	6.610437	8.312882	0.69348

3.7 Analysis of Results

The performance evaluation indicates that both models demonstrate good predictive capability in modeling the association between Reading Score and Math Score. On the testing dataset, Linear Regression achieved an R^2 value of 0.680447, while the Decision Tree model achieved 0.657464. The MAE and RMSE values between the two models are relatively close, indicating comparable predictive accuracy. However, Linear Regression shows slightly better generalization performance on unseen data.

To reduce the risk of overfitting bias, training and testing performances were compared. For Linear Regression, the R^2 value increased slightly from 0.663932 (training) to 0.680447 (testing), indicating stable generalization and no evidence of overfitting. In contrast, the Decision Tree model showed a decrease in R^2 from 0.693480 (training) to 0.657464 (testing), suggesting a mild generalization gap. Although the decline is not substantial, it indicates that the Decision Tree model fits the training data more closely than Linear Regression. Overall, Linear Regression performs more consistently across datasets.

An R^2 value of 0.680447 indicates that approximately 68% of the variance in Math Score is statistically associated with variation in Reading Score. This result suggests a strong relationship between reading literacy and mathematics achievement. However, the remaining 32% of the variance is influenced by other factors not included in this model. These may include socioeconomic background, parental education level, learning environment, teaching quality, and student motivation. Therefore, mathematics performance is multidimensional and cannot be fully explained by reading ability alone.

Beyond numerical metrics, visual diagnostic analysis was conducted to ensure that the statistical indicators are supported by the observed data distribution. Visual inspection is an important component of regression evaluation because it confirms whether the model structure appropriately represents the data pattern. The relationship between Reading Score and Math Score is illustrated in Figure 4. The scatter plot combined with the regression line serves as a graphical validation of the model's predictive structure.

Figure 4. Visualization of Linear Regression

The regression line demonstrates a clear positive trend, and the concentration of data points around the line visually supports the R^2 value of 0.68. This alignment between statistical metrics and graphical representation strengthens the credibility of the linear regression model.

In addition, to evaluating predictive performance, residual diagnostics were conducted to assess potential bias and model assumptions. The residual distribution is presented in Figure 5.

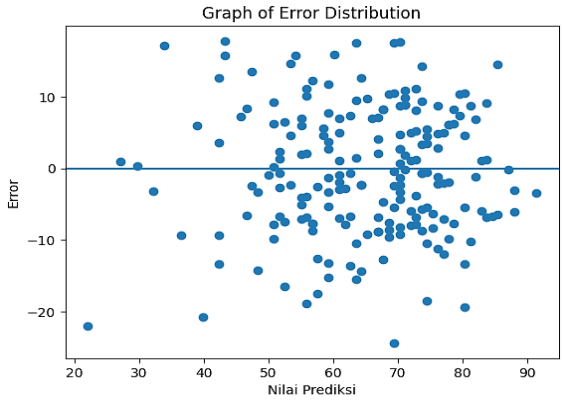

Figure 5. Graph of Error Distribution

The residuals are randomly distributed around zero without observable systematic patterns. This visual diagnostic suggests that the model does not exhibit major heteroscedasticity or structural bias, thereby reinforcing the reliability of the regression results.

Although the Decision Tree model achieved slightly higher performance during training, Linear Regression offers greater interpretability and stability. Linear Regression provides a transparent mathematical

equation that quantifies the statistical relationship between Reading Score and Math Score. This interpretability is particularly important in educational research contexts, where explainable findings are necessary for informed academic decision-making. In contrast, Decision Tree models generate rule-based splits that are less straightforward in expressing direct quantitative relationships.

Nevertheless, several limitations should be acknowledged. The dataset was obtained from Kaggle and does not specifically represent Indonesian students. Differences in curriculum structure, cultural context, assessment systems, and educational environments may limit the generalizability of these findings to the Indonesian education context. Furthermore, the model includes only one predictor variable, which restricts its explanatory scope. Future research should incorporate additional relevant predictors and utilize locally collected educational data to enhance contextual validity and improve model robustness.

Importantly, the findings of this study indicate statistical association rather than causal relationships. Although Reading Score is strongly associated with Math Score, the analysis does not establish that reading ability directly causes improvements in mathematics performance. Further longitudinal or experimental research would be required to examine potential causal mechanisms.

4. CONCLUSION

This study concludes that the Linear Regression method is capable of effectively modeling and predicting students' mathematics scores based on their reading scores. The modeling results show a positive linear relationship between reading scores and math scores, with a coefficient of determination (R^2) value of 0.68, indicating that approximately 68% of the variance in students' math scores is statistically explained by reading scores within the model, rather than representing a direct causal contribution. Model evaluation using MAE and RMSE metrics shows a relatively low and stable prediction error rate, as well as better linear regression model performance compared to the Decision Tree Regressor method. The visualization of the regression line and residual distribution reinforces the finding that the model does not experience systematic bias and is able to represent the linear relationship well. The results of this study indicate that reading ability is strongly associated with mathematics academic achievement within the observed dataset; however, this relationship should not be interpreted as causal. The limitations of this study lie in the use of a single independent variable and the application of a model that is still relatively simple. Therefore, further research is recommended to add other relevant variables, use more complex regression methods or machine learning algorithms, apply multi-variable modeling approaches, and implement more robust validation techniques such as k-fold cross-validation to obtain more comprehensive and generalizable prediction results.

REFERENCES

[1] N. Fauziah, W. Hadi, and Y. Sari, "The Relationship between Reading Comprehension Ability and the Ability to Solve Mathematics Story Problems for Class V Elementary School," *Jurnal Gentala Pendidikan Dasar*, vol. 9, no. 1, pp. 55–58, Jun. 2024, doi: 10.22437/GENTALA.V9I1.32978.

[2] I. R. Boctot, D. M. Enriquez, and C. P. Yurango, "Reading Comprehension as a Predictor of Mathematical Word Problem-solving Ability among Grade 7 Students," *Asian Journal of Education and Social Studies*, vol. 51, no. 7, pp. 1115–1121, Jul. 2025, doi: 10.9734/AJESS/2025/V51I72196.

[3] Fitri Anisa Kusumastuti, Novela Wulandari, Muh. Khaedir Lutfi, and Aeni Rohmawati, "Kemampuan Membaca Teks Matematika Sebagai Prediktor Literasi Matematis Siswa Sekolah Menengah Pertama," *JIPMat*, vol. 10, no. 2, pp. 198–211, Oct. 2025, doi: 10.26877/jipmat.v10i2.2663.

[4] R. P. Mentari, R. Tuanaya, and M. Albrecht, "Correlation Of Reading Comprehension Skill And Ability To Solve Mathematics Story Questions Of Students In Indonesia: A Meta-Analysis," *Matematika Dan Pembelajaran*, vol. 11, no. 2, pp. 154–168, Nov. 2023, doi: 10.33477/mp.v11i2.5514.

[5] I. Simbolon, P. Aditya, and E. Br Purba, "Prediksi Performa Akademik Siswa Berdasarkan Kehadiran dan Aktivitas E-Learning Menggunakan Algoritma Decision Tree," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 2, pp. 4899–4910, Jul. 2025, doi: 10.31004/riggs.v4i2.1352.

[6] Riska Rismaya, Dwi Yuniarto, and David Setiadi, "Penerapan Algoritma Machine Learning dalam Prediksi Prestasi Akademik Mahasiswa," *Router : Jurnal Teknik Informatika dan Terapan*, vol. 3, no. 1, pp. 15–23, Feb. 2025, doi: 10.62951/router.v3i1.389.

[7] K. V. Patil, K. D. Yesugade, and K. B. Naikwadi, "A Study on Regression Based Machine Learning Models to Predict the Student Performance," *Journal of Engineering Education Transformations*, vol. 38, no. 2, pp. 177–186, Oct. 2024, doi: 10.16920/jeet/2024/v38i2/24200.

[8] Y. Zhang and M. Cutumisu, "Predicting the Mathematics Literacy of Resilient Students from High-performing Economies: A Machine Learning Approach," *Studies in Educational Evaluation*, vol. 83, p. 101412, Dec. 2024, doi: 10.1016/J.STUEDUC.2024.101412.

[9] S. Bhutto, I. F. Siddiqui, Q. A. Arain, and M. Anwar, "Predicting Students' Academic Performance Through Supervised Machine Learning," *ICISCT 2020 - 2nd International Conference on Information Science and Communication Technology*, Feb. 2020, doi: 10.1109/ICISCT49550.2020.9080033.

[10] R. Guevara-Reyes, I. Ortiz-Garcés, R. Andrade, F. Cox-Riquetti, and W. Villegas-Ch, "Machine learning models for academic performance prediction: interpretability and application in educational decision-making," *Front. Educ. (Lausanne).*, vol. 10, p. 1632315, Aug. 2025, doi: 10.3389/FEDUC.2025.1632315/BIBTEX.

[11] M. K. Kassy, "Predicting Student Performance using Linear Regression," *Data Science Insights*, vol. 3, no. 2, pp. 66–74, Aug. 2025, doi: 10.63017/jdsi.v3i2.104.

[12] R. I. Athallah, G. Al Godzali, and E. Rivalni, "Academic Performance Prediction from Study Habits and Lifestyle using Linear Regression," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 5, no. 1, pp. 337–343, Oct. 2025, doi: 10.59934/jaiea.v5i1.1313.

[13] D. Kristiani and N. Tupulu, "Pengaruh Kemampuan Membaca dan Motivasi Belajar Terhadap Kemampuan Pemecahan Masalah pada Soal Cerita Matematika," *Jurnal Pendidikan Matematika*, vol. 4, pp. 789–797, Aug. 2025, doi: 10.56916/jp.v4i3.2205.

[14] H. ∗ Volia, U. Citra, B. Roswita, L. Nahak, U. Citra Bangsa, and C. A. Naitili, "Pengaruh Literasi Digital dan Minat Baca Terhadap Motivasi Belajar Siswa SD GMIT Kuanino 3 Kupang," *Jurnal Jendela Pendidikan*, vol. 5, Nov. 2025, doi: 10.57008/jjp.v5i04.1793.

[15] "Students Performance in Exams." Accessed: Dec. 17, 2025. [Online]. Available: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams?utm_source

[16] B. Mahendra, D. Pratama, A. Faqih, and R. Kurniawan, "Evaluasi Pengaruh Kualitas Data Terhadap Performa Model Machine Learning Menggunakan Pendekatan Data-Centric AI," *Jurnal Sistem Informasi dan Teknologi (SINTEK)*, doi: 10.56995/sintek.v6i1.211.

[17] Pratik Mahajan, "Machine Learning-Based Data Preprocessing as well as Visualization Techniques for Predicting Students' Tasks," in *Demystifying Emerging Trends in Machine Learning*, BENTHAM SCIENCE PUBLISHERS, 2025. doi: 10.2174/9789815305395125020l.

[18] V. Çetin and O. Yıldız, "A comprehensive review on data preprocessing techniques in data analysis," *Pamukkale University Journal of Engineering Sciences*, vol. 28, no. 2, pp. 299–312, Apr. 2022, doi: 10.5505/pajes.2021.62687.

[19] A. M. Sharifnia, D. E. Kpormegbey, D. K. Thapa, and M. Cleary, "A Primer of Data Cleaning in Quantitative Research: Handling Missing Values and Outliers," *J. Adv. Nurs.*, vol. 82, no. 1, pp. 970–975, Jan. 2026, doi: 10.1111/jan.16908.

[20] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Min.*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.

[21] J. J. Salazar, L. Garland, J. Ochoa, and M. J. Pyrcz, "Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy," *J. Pet. Sci. Eng.*, vol. 209, p. 109885, Feb. 2022, doi: 10.1016/j.petrol.2021.109885.

[22] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, Jul. 2021, doi: 10.7717/PEERJ-CS.623.

[23] I. Shatz, "Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics," *Behavior Research Methods 2023 56:2*, vol. 56, no. 2, pp. 826–845, Mar. 2023, doi: 10.3758/s13428-023-02072-x.

[24] S. Midway and J. W. White, "Testing for normality in regression models: mistakes abound (but may not matter)," *R. Soc. Open Sci.*, vol. 12, no. 4, p. 241904, Apr. 2025, doi: 10.1098/rsos.241904.