Journal of Artificial Intelligence and Engineering Applications

*Website: https://ioinformatic.org/*

***15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519***

# Academic Performance Prediction from Study Habits and Lifestyle using Linear Regression

**Rafif Isdarufa Athallah[1]*, Galva Al Godzali[2], Elkin Rivalni[3]**

*[1,2,3]Informatics Enginneering, Pelita Bangsa University,Indonesia*
*rafifiathaalah@mhs.pelitabangsa.ac.id[1*], galvaalghazali@gmail.coml[2] ,elkin.rilvani@pelitabangsa.ac.id[3]*

**Abstract**

Academic performance is a critical indicator of student success in higher education, influenced by factors such as study habits, sleep patterns, and extracurricular engagement. This study presents a web-based application developed using the Streamlit framework and a linear regression model to predict students' academic Performance Index based on key predictors, including previous grades, study hours, sleep duration, practice question engagement, and extracurricular activities. Utilizing a dataset from Kaggle with 10,000 student entries, the model achieved a high R-squared ($R^2$) value of 0.9890 and a low Mean Squared Error (MSE) of 4.0826, indicating robust predictive accuracy. The application provides interactive visualizations of factor contributions and performance categories (High, Medium, Low) to support students in identifying strengths and weaknesses in their learning strategies. This study contributes to educational technology by offering a practical, data-driven tool for personalized academic improvement.

***Keywords***: *Academic Performance, Linear Regression, Machine Learning, Streamlit, Predictive Modeling, Student Success*

## 1. Introduction

Academic achievement reflects a student's intellectual capacity, discipline, and engagement during the learning process. In higher education institutions, academic success is shaped by various factors such as sleep patterns, stress levels, study habits, and participation in extracurricular activities [1]. University students frequently face significant pressure due to high academic demands, often leading them to reduce sleep, experience elevated stress, and struggle to balance academic responsibilities with part-time work or student organizations. When unmanaged, these pressures can adversely affect academic achievement and harm students' mental well-being [2].

Amid current technological advancements, particularly in machine learning, new opportunities have emerged to analyze educational data and forecast academic outcomes [3]. By applying the appropriate algorithms, patterns in historical data can be identified to predict student performance and support decision-making [4]. However, many students still struggle to detect early signs of academic challenges and lack the tools to develop effective, personalized learning strategies [5]. Predictive systems in this context not only offer early warnings but also serve as instruments for reflecting on and improving academic behaviors.

To address this gap, this study introduces a web-based application designed to predict students' academic Performance Index and present actionable insights through visualizations and tailored feedback. The application leverages the Streamlit framework and employs a linear regression model selected for its computational efficiency, transparency, and ability to model linear relationships between academic predictors and outcomes [6]. These predictors include previous academic scores, study duration, sleep hours, practice frequency, and extracurricular participation.

The application aims to empower students by providing real-time predictions and illustrating the impact of each factor on performance outcomes. By displaying the influence of individual inputs graphically, the tool enables users to recognize which aspects are most decisive and supports self-directed academic improvement. As such, a machine learning approach becomes not only a predictive mechanism but also a medium to foster awareness, reflection, and strategic learning [7].

## 2. Research Methodology

This research applies a predictive approach in the development of an academic performance monitoring system based on machine learning. The system is designed as a web-based application to predict student performance using the Multiple Linear Regression method[8]. The application was developed using the Streamlit framework and implemented in the Python programming language with relevant libraries, including scikit-learn, pandas, matplotlib, and seaborn for visualization[9].

## 2.1. Method Justification

Linear regression is selected due to its simplicity, computational efficiency, and high interpretability, making it suitable for applications in educational environments. It has been shown to remain an effective baseline model in educational prediction tasks, providing insights that are easily understood by both technical and non-technical users[14]. Furthermore, linear models can outperform more complex algorithms when data quality is sufficient and the relationships between variables are approximately linear.

By utilizing a supervised learning framework, this study aims to build a transparent model where the influence of each predictor variable (e.g., study hours, previous grades, sleep duration) is explicitly shown, thereby supporting self-directed learning strategies.

## 2.2. Dataset

The dataset used was obtained from Kaggle (Student Performance – Multiple Linear Regression) *(https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression)* and contains 10,000 records. The attributes include:

**Table 1:** Attributes of Student Performance Dataset

| No | Feature Name | Description | Data Type / Scale |
|---|---|---|---|
| 1 | Previous Scores | Student's previous academic scores | Numerical (0–100 scale) |
| 2 | Hours Studied | Average number of hours studied per day | Numerical (continuous) |
| 3 | Sleep Hours | Average daily sleep duration | Numerical (continuous) |
| 4 | Sample Question Papers Practiced | Total number of practice questions completed | Numerical (integer count) |
| 5 | Extracurricular Activities | Participation status in extracurricular activities | Categorical (0 = No, 1 = Yes) |
| 6 | Performance Index | Target variable representing academic performance | Numerical (0–100 scale) |

All attributes are numeric, and preprocessing steps were performed to normalize input values and encode categorical variables. For user-friendly input, binary features are displayed as "Yes" or "No" in the interface, then converted back to numerical format during model computation.

## 2.3. System Development Flow

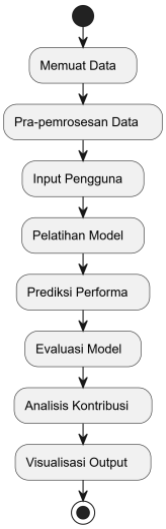


**Fig. 1:** Flow diagram of the academic performance prediction system

1. Loading and Preview Datasets
   Datasets are imported and partially rowed to allow users to view the contents of the data before they are used in modeling.
2. Data Pre-processing and Encoding
   Data is processed to be ready for use, including transformation if needed and to ensure its quality [10]. All features are numerical, including Extracurricular Activities which were originally 0 and 1, but are displayed to users in the form of "Yes" and "No" options.
3. Input Data
   Users fill out interactive input forms based on their learning habits, such as study hours, sleep, etc.
4. Linear Regression Model Training
   The data is divided into 80% of the training data and 20% of the test data [11]. The model is trained using trained data and evaluated on test data.
5. Model Evaluation
   Model performance was measured using Mean Squared Error (MSE) and R-squared ($R^2$)[12], and visualized in an actual vs. predicted value graph.
6. User Performance Prediction
   The model predicts the Performance Index based on user input and categorizes the results into three levels: High, Medium, or Low[13].
7. Visualization of Feature Contributions
   The app displays a graph of each feature's contribution to the user's predictive results, based on the input value multiplied by the regression coefficient.

8. Result Display
   The prediction results are presented in the form of numerical scores, performance level indicators (color/status), and explanatory narratives and recommendations.

### 2.4. Evaluation Method

The model is evaluated using two primary regression metrics:

a. Mean Squared Error (MSE), which measures the average squared difference between predicted and actual values,
b. R-squared ($R^2$), which represents the proportion of variance in the dependent variable explained by the independent variables.

In this study, the model achieved an MSE of 4.0826 and an $R^2$ of 0.9890, indicating a high level of predictive performance.

## 3. Result and Discussion

This section presents the outcomes of the implementation of the multiple linear regression model for predicting academic performance, followed by a discussion of model performance, visualization results, and interpretation of feature contributions.

### 3.1. Application Interface and Functionality

The application provides an interactive interface for predicting the academic achievement index (Performance Index) based on user input. Developed using the Streamlit framework, the application features a clean, browser-based interface that is intuitive and easy to use.

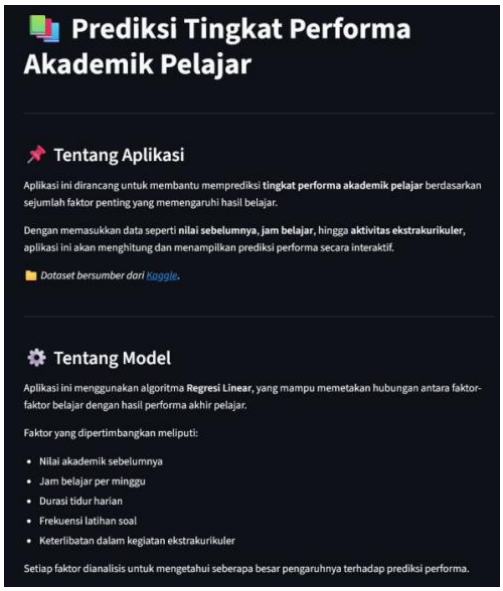


**Fig. 2:** Application initial view

Fig. 2 presents the homepage of the academic performance prediction application. This initial interface is divided into two main sections: *About Application* and *About Model*. The *About Application* section explains that the tool is designed to help predict students' academic performance based on various influential factors, such as previous grades, daily study hours, and extracurricular involvement. User input is collected interactively, and predictions are generated using a linear regression model. Additionally, this section specifies that the dataset used originates from the Kaggle platform.

The *About Model* section describes the algorithmic approach used in the system, which is linear regression. This model is designed to map the relationship between several learning factors and academic outcomes, including previous scores, weekly study hours, daily sleep duration, frequency of practice questions attempted, and participation in extracurricular activities. Each variable is analyzed to determine its relative impact on the predicted performance. This section aims to provide users with a general understanding of the application's mechanism and its modeling foundation.


**Fig. 3:** Correlation and Regression Coefficients of Predictive Features

Fig. 3 displays a table containing the correlation values and regression coefficients of each variable affecting the Performance Index. The correlation values represent the strength of the linear relationship between each factor and academic performance, based on historical data. Meanwhile, the regression coefficients quantify the contribution of each factor within the trained linear regression model.

From the table, it is evident that Previous Grade has the highest correlation (0.915) and a coefficient of 1.018, indicating that prior academic performance is the most significant predictor of future performance. Study Hours also shows a substantial impact with a coefficient of 2.853, although it has a lower correlation (0.374), suggesting that increased study time can meaningfully influence predictions. Other factors, such as Sleep Hours, Practice Questions, and Extracurricular Activities, demonstrate lower but still notable contributions.

This visualization provides users with a clearer understanding of which factors most influence the predicted outcome, thereby supporting transparency and interpretability in an educational context.

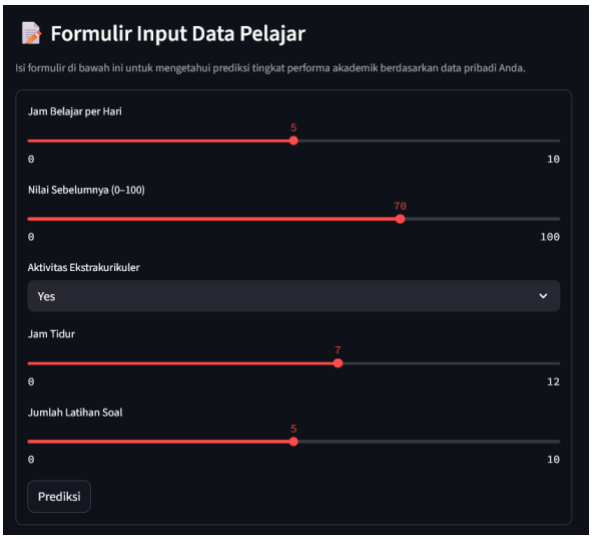


**Fig. 4:** Student Input Form Interface

Fig. 4 illustrates the input form interface within the application, allowing users to provide personal data relevant to academic performance prediction. The form captures various fields including average study hours per day, prior academic scores (on a 0–100 scale), extracurricular participation (via "Yes"/"No" selection), average sleep duration, and number of practice questions completed.

Input components are implemented through sliders and dropdown menus to streamline user interaction. Once the form is completed, users can press the Predict button, which triggers the linear regression model to calculate the Performance Index. The design of the interface is intuitive and responsive, facilitating exploration and reflection on personal academic habits.

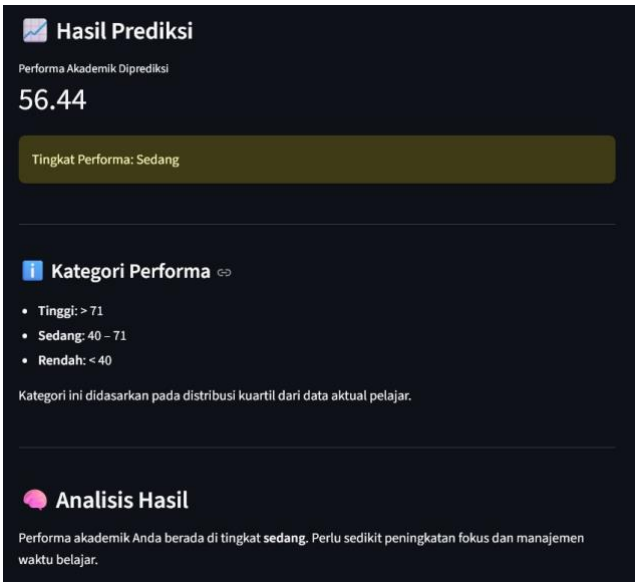


**Fig. 5:** Prediction Results and Performance Analysis

Fig. 5 displays the predicted academic performance based on the user's input. For example, a prediction result of 56.44 is automatically classified under the medium category based on quartile thresholds established from the historical dataset. The categorization is defined as:

1. High: > 71
2. Medium: 40–71
3. Low: < 40

This classification provides an accessible interpretation of the numerical prediction. In addition to the score, the application generates a brief recommendation. For instance, a medium classification may include suggestions for better time management or focused study strategies. This feature aims to offer actionable insights for students to improve their academic performance.

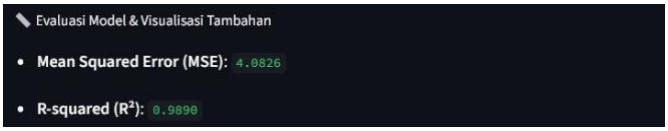


**Fig. 6:** Model Evaluation

Fig. 6 presents the evaluation metrics for the linear regression model used in the system. The model achieved a Mean Squared Error (MSE) of 4.0826 and an R-squared ($R^2$) of 0.9890. The high $R^2$ value indicates that the model can explain approximately 98.9% of the variance in the target data, which reflects excellent predictive accuracy and model fit.

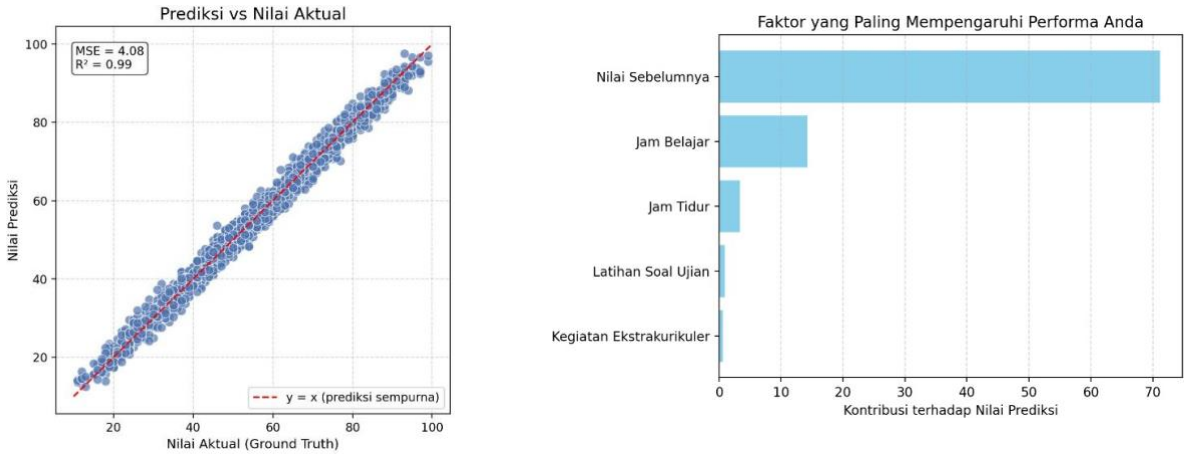


**Fig. 7:** Visualization of Feature Contributions

Fig. 7 contains two key visualizations. The first graph depicts the relationship between actual and predicted values. Each blue dot represents a prediction, while the red dotted line (y = x) indicates perfect prediction alignment. The close distribution of data points along this line suggests strong model accuracy and minimal bias.

The second graph illustrates the contribution of each input factor to the final predicted score. The results show that *Previous Grade* and *Study Hours* are the most influential features, followed by *Sleep Hours*, *Practice Questions*, and *Extracurricular Activities*. This visual representation helps users understand which areas of their academic behavior most significantly influence their predicted performance, thereby guiding more effective learning strategies.

### 3.2. Model Performance and Evaluation

The predictive model employed in this application is Linear Regression, designed to estimate a student's Performance Index based on five key input variables: previous academic scores, daily study hours, sleep duration, number of practice questions completed, and participation in extracurricular activities.

To assess the model's performance, two widely accepted regression evaluation metrics were used:

1. Mean Squared Error (MSE): 4.0826
2. R-squared ($R^2$): 0.9890

Interpretation:

1. The low MSE value indicates that the average squared difference between the predicted and actual values is minimal, suggesting high predictive precision.
2. The $R^2$ value of 0.9890 demonstrates that approximately 98.9% of the variance in the Performance Index can be explained by the input variables. This reflects a strong model fit and confirms the model's effectiveness in capturing the relationship between the predictors and the target outcome.

These results support the reliability of the linear regression model in forecasting academic performance based on the selected behavioral and academic features.

### 3.3. Prediction Output Visualization

The prediction results generated by the application are presented in two formats: a numerical output representing the Performance Index and a categorical classification indicating performance level—High, Medium, or Low. These outputs provide users with both quantitative and qualitative insight into their predicted academic standing.

To enhance interpretability, the application includes a scatter plot visualization comparing the actual values with the predicted values. This allows users to observe how closely the predictions align with the real data.

The scatter plot displays the distribution of prediction points relative to the diagonal reference line ($y = x$). A concentration of points near this line indicates that the model achieves high predictive accuracy. This visual confirmation reinforces the model's reliability, especially when making predictions on previously unseen data.

### 3.4. Model Coefficients

The linear regression model generates a set of coefficients that represent the influence of each input variable on the predicted Performance Index. A higher coefficient indicates a stronger contribution of that variable to the final prediction outcome. Table 4.1 presents the coefficient values derived from the trained model.

**Table 2:** Linear Regression Coefficients and the Contribution of Predictive Features

| Feature | Coefficient |
|---|---|
| Study Hours | 2.853 |
| Previous Grade | 1.018 |
| Sleep Hours (Bedtime) | 0.481 |
| Extracurricular Activities | 0.613 |
| Practice Exam Questions | 0.194 |

Interpretation:

1. Study Hours has the highest coefficient (2.853), indicating it is the most influential factor. Increasing study time significantly improves predicted academic performance.
2. Previous Grade also contributes strongly to prediction results, emphasizing the importance of prior academic achievement.
3. Other features, such as Sleep Hours, Extracurricular Activities, and Practice Exam Questions, show positive but comparatively smaller impacts on the outcome.
4. Notably, all coefficients are positive, meaning each factor contributes constructively to the predicted Performance Index.

A horizontal bar chart is used to visualize the relative contributions of each variable. This graphical representation enables users to better understand which areas of their academic behavior have the greatest impact on predicted performance, and helps guide strategies for academic improvement.

## 4. Conclusion

This study presents a web-based academic performance prediction tool that effectively bridges machine learning technology and student self-assessment. The novelty of this application lies in its ability to transform a transparent, interpretable linear regression model into an intuitive interface that empowers students with actionable insights.

The model achieved exceptional predictive accuracy, with an $R^2$ value of 0.9890 and a low MSE of 4.0826, proving its robustness in capturing performance patterns from behavioral and academic variables. This application integrates real-time visualizations, personalized feedback, and feature contribution analysis to present complex predictive insights in a clear and accessible manner, enabling students to better understand and reflect on their academic progress.

By emphasizing interpretability, usability, and precision, this work contributes a scalable and evidence-driven solution to the field of educational technology, with potential applications in both institutional analytics and individual student support systems.

## References


[1] D. M. Jannah, M. T. Hidayat, M. Ibrahim, and S. Kasiyun, "Pengaruh kebiasaan belajar dan motivasi belajar terhadap prestasi belajar siswa di sekolah dasar," J. Basicedu, vol. 5, no. 5, pp. 3378–3384, Aug. 2021, doi: 10.31004/basicedu.v5i5.1350.

[2] A. Djohar and B. Hermawan, "Hubungan kualitas tidur dan motivasi belajar terhadap prestasi belajar pada mahasiswa Fakultas Kedokteran Universitas Muhammadiyah Surakarta," in Proc. Univ. Muhammadiyah Surakarta, 2023.

[3] J. D. Prabowo, H. Setiawan, and M. F. Nugraha, "Analisis komparasi algoritma machine learning dalam prediksi performa akademik mahasiswa: literature review," J. Ilmu Komput. Inform. (JIKI), vol. 4, no. 2, pp. 143–149, 2023.

[4] M. Putra and E. Harahap, "Machine learning pada prediksi kelulusan mahasiswa menggunakan algoritma Random Forest," J. Riset Mat., vol. 4, no. 1, 2024.

[5] K. Afandi and M. H. Arief, "Analisis performa akademik mahasiswa menggunakan social network analysis (studi kasus: Prodi Bisnis Digital Universitas dr. Soebandi)," J. Technol. Inform. (JoTI), vol. 5, no. 2, pp. 181–195, 2024, doi: 10.37802/joti.v5i2.514.

[6] F. A. Pribadi and P. D. Ramadhan, "Sistem pendukung keputusan untuk menentukan harga bahan pokok menggunakan regresi linier," J. Inform. Polinema, vol. 10, no. 1, pp. 70–78, Nov. 2023, doi: 10.33795/jip.v10i1.1421.

[7] H. Handoyo and M. Najib, "Prediction of student performance index based on hours studied, previous scores, extracurricular activities, sleep hours, and sample question papers practiced with least-squares linear regression," Al-Aqlu: J. Pemikir. Penelit. Ekon. Islam, vol. 7, no. 1, pp. 55–67, 2025.

[8] M. R. Al Fadry and A. R. Susanti, "Pengembangan dashboard untuk analisis informasi ijazah menggunakan Streamlit," Karimah Tauhid, vol. 3, no. 11, pp. 78–85, 2024.

[9] M. Sholeh, S. Suraya, and D. Andayati, "Machine linear untuk analisis regresi linier biaya asuransi kesehatan dengan menggunakan Python Jupyter Notebook," J. Eduk. Penelit. Inform. (JEPIN), vol. 8, no. 1, pp. 20–27, Apr. 2022.

[10] E. Etriyanti, "Perbandingan tingkat akurasi metode KNN dan Decision Tree dalam memprediksi lama studi mahasiswa," J. Ilm. Bin. STMIK Bina Nusantara Jaya Lubuklinggau, vol. 3, no. 1, pp. 6–14, 2021, doi: 10.52303/jb.v3i1.40.

[11] W. Musu, A. Ibrahim, and H. Heriadi, "Pengaruh komposisi data training dan testing terhadap akurasi algoritma C4.5," SISITI (Sistem dan Informasi Teknologi Informasi), vol. 10, no. 1, Mar. 2021, doi: 10.36774/sisiti.v10i1.802.

[12] H. Tobing, M. F. Arfa, M. R. Al Fathan, and R. Rahmaddeni, "Prediksi harga cryptocurrency dengan metode linier regresi," SENTIMAS: Semin. Nas. Penelit. Pengab. Masy., vol. 1, no. 1, pp. 8–15, Aug. 2023, doi: 10.36774/sisiti.v10i1.802.

[13] S. Arti and E. Suherlan, "Evaluasi kinerja machine learning dalam memprediksi kemampuan adaptasi mahasiswa pada lingkungan pembelajaran daring," J. Pustaka AI, vol. 5, no. 1, pp. 50–57, Apr. 2025, doi: 10.55382/jurnalpustakaai.v5i1.901.

[14] B. C. Ngongoloy dan M. A. I. Pakereng, "Penerapan Metode Linear Regression dan Correlation Pearson Dalam Menganalisis Pengaruh Kualitas Pembelajaran Online Terhadap Prestasi Akademik," *Progresif: Jurnal Ilmiah Komputer*, vol. 5, no. 2, pp. 115–124, Ags. 2023