Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Predicting Student Academic Performance Based on Learning Habits Using XGBoost and SHAP

Siti Latifah[1]*, Martanto[2], Raditya Danar Dana[3], Fatihanursari Dikananda[4], Umi Hayati [5]

[1,2,4,5]*Computer Science, STMIK IKMI Cirebon, Indonesia*
[3]*Information Management STMIK IKMI Cirebon, Indonesia*
faalatiif19@gmail.com[1*], martantomusijo@gmail.com[2] , radith_danar@yahoo.com[3] , fatiha.dikananda@gmail.com[4], umi@stmik-amikbandung.ac.id[5]

Abstract

This study developed a model for predicting student academic achievement based on learning habits using the XGBoost algorithm and SHAP interpretability techniques. The secondary dataset contains 1,000 entries and 16 variables (for example, hours of study per day, mental health, frequency of exercise, social media use, hours of sleep) pre-processed including cleaning, imputation, encoding, and normalization before being divided into train–test (80:20) and validated using 5-fold CV. Three models were tested: Linear Regression, Random Forest, and XGBoost. Evaluation using RMSE, MAE, and R^2 showed that XGBoost achieved RMSE = 0.335, MAE = 0.266, and R^2 = 0.882, while Linear Regression showed the best performance according to R^2 in certain configurations (R^2 = 0.888; RMSE = 0.326). SHAP analysis revealed that the most influential features were hours of study per day, mental health scores, exercise frequency, duration of social media use, and hours spent watching Netflix. The findings confirm that students' study habits and psychological conditions are the main determinants of academic achievement variation; the use of interpretable features strengthens the readability of the model for education stakeholders. Research recommendations include testing the model on longitudinal datasets, integrating socioeconomic factors, and implementing data privacy procedures before institutional-scale implementation.

Keywords: *XGBoost; study habits; academic achievement prediction; SHAP; model evaluation*

1. Introduction

Advances in computer technology and data analysis have changed the way we understand and improve student learning outcomes in higher education. Recent research shows that learning habits, such as time management, active learning methods, reading skills, independent learning, and healthy lifestyle habits such as getting enough sleep and avoiding distractions, greatly influence academic performance. These learning habits not only show how a person learns, but also serve as key indicators that explain differences in academic achievement in various educational environments.

Meanwhile, advances in machine learning (ML) have opened up new opportunities in educational data mining (EDM), which is the processing of educational data to predict and understand how effectively students learn. Ensemble-based models, such as Extreme Gradient Boosting (XGBoost), have proven effective in capturing the complex relationship between learning behavior and academic outcomes [1]. XGBoost is capable of combining various types of information, such as demographic data, learning behavior, and previous grades, to provide accurate predictions and early detection of students at risk of academic difficulties [2].

Furthermore, combining XGBoost with interpretability methods such as SHAP allows researchers and educators to understand how much each variable contributes to the prediction results [3]. With this approach, factors such as learning duration, independent learning ability, distraction levels, and participation in quizzes and online activities can be identified as key factors that influence academic success. Thus, the ability of the model to be clearly explained is important in linking data analysis results with targeted educational decisions.

In the context of higher education, combining the XGBoost method with SHAP not only maximizes accuracy in projecting academic performance, but also helps educators to create more personalized and data-driven learning intervention plans [4]. This is in line with the global trend towards adaptive learning systems, with a focus on improving the student learning experience through evidence-based data analysis. However, there is still a lack of research on how learning habits can be systematically modeled using explainable machine learning approaches.

Previous studies have mostly focused on accuracy levels without exploring the meaning and explanation of each variable that affects academic performance [5]. Therefore, an approach is needed that can provide explanations that can be used directly by educators and policymakers. Based on this description, this study focuses on the application of the XGBoost model combined with SHAP analysis to predict student academic performance based on their learning habits. This approach is expected to provide more comprehensive insights

into the factors that determine academic success, as well as support the development of more effective and adaptive data-driven learning strategies in higher education.

2. Research Methodology

This study uses a quantitative approach with an experimental design. The secondary dataset was obtained from a public platform (Kaggle) containing 1,000 student data and 16 variables. The research steps included data pre-processing, dividing the data into train-test (80:20), training models using Linear Regression, Random Forest, and XGBoost, and evaluation using the RMSE, MAE, and R^2 metrics. In addition, the results were interpreted using the SHAP method to understand the contribution of each feature to the model prediction.
The goal is to build an early warning system and pedagogical action recommendations [6]. This approach enables educational institutions to identify low-risk students and understand the factors that influence their learning outcomes. Empirically, ensemble tree-based models such as XGBoost show better prediction performance than traditional algorithms when properly optimized and validated [7]. Therefore, this study uses best practices in EDM by combining learning habit feature engineering, supervised boosting models, and model interpretability to support data-driven decision making.

2.1. Reasearch Subject

The object of this study is a dataset on student learning habits and academic performance that includes a number of variables such as daily study duration, hours of sleep, mental health, frequency of exercise, and intensity of social media and entertainment platform (Netflix) use. This dataset reflects a combination of behavioral, cognitive, and psychological factors that influence students' academic achievement.
This dataset does not originate from an LMS (Learning Management System), but is rather aggregated data from measurements of students' learning habits and lifestyles. This approach was used to complement previous studies that focused only on online activity data without considering students' affective and mental factors. According to [8] and [9], prediction models that combine behavioral and psychological data are capable of producing more accurate and representative predictions of students' academic performance.

2.2. Data Collection Method

This study uses secondary data in CSV format obtained from a public dataset and stored in Google Drive. The data is then processed using Google Colab with the Python programming language. The data collection process is carried out in the following stages:

1. The dataset is downloaded from Google Drive to the Colab environment.
2. The data is preliminarily checked to ensure the form, data type, and completeness of the values.
3. The data was pre-processed using steps such as removing missing data, normalizing numerical data, converting categorical variables using LabelEncoder, and standardizing features with StandardScaler.
4. The data was validated through correlation analysis, outlier detection, and checking the distribution of the results.

The dataset is divided into two parts with an 80:20 ratio for training and testing data. To reduce the risk of overfitting, a cross-validation method is used (k-fold cross-validation, k=5). After the model is trained, an interpretation analysis is performed using SHAP to determine the contribution of each feature to the prediction results. These steps ensure that the research results are clear, repeatable, and meet the validation standards for quantitative research.

2.3. Research Instrument

This research instrument uses a combination of software and supporting libraries in the Python ecosystem, namely:

a. Pandas and NumPy: Used to process and manipulate data.
b. Matplotlib and Seaborn: Used to create data visualizations and illustrate the results of correlation analysis between variables.
c. Scikit-learn: Used to split the dataset, transform feature data, create basic models, and evaluate model performance using MAE, RMSE, and R^2 metrics.
d. XGBoost: Used to create prediction models based on efficient and accurate gradient boosting algorithms.
e. SHAP: Used to analyze the model's ability to provide quantitative explanations of the influence of each feature on the prediction results.

This software is used to create a standardized, repeatable modeling workflow (pipeline) that is in line with modern research methods as recommended by [3] and [7].

2.4. Data Analysis Techniques

Data analysis was conducted in two main stages, namely exploratory data analysis (EDA) and predictive analysis (modeling). In the first stage, EDA used Pearson and Spearman correlation tests to examine the relationship between variables and heatmap visualization to determine the level of feature correlation. To ensure that there were no redundant features, multicollinearity analysis was performed using the Variance Inflation Factor (VIF) method. The second stage is predictive analysis, in which three regression models are compared, namely Linear Regression, Random Forest, and XGBoost.
The results are evaluated using several metrics, namely Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2). The comparison results show that the best model is the one with the highest accuracy and the smallest prediction error. After the modeling process is complete, the results are interpreted using SHAP.
SHAP is used to measure the contribution of each feature to the prediction results. SHAP provides both a global understanding (the most important features overall) and a local understanding (the influence of features on individual predictions). This approach increases model transparency and supports evidence-based decision-making in education [10][11][3].

3. Results And Discussion

This research was conducted through several structured stages, namely the application of Linear Regression and XGBoost algorithms to analyze the effect of student learning habits on academic achievement. The stages carried out included: (1) Data pre-processing, (2) Separation of data into training data and test data, (3) Development and training of prediction models, and (4) Evaluation and interpretation of results using performance metrics and the SHAP (SHapley Additive Explanations) method to understand the contribution of each variable to the prediction results. The following is an explanation of the results of each stage carried out in this study:
In this study, the pre-processing was carried out sequentially, including examining the data structure, identifying missing values, and calculating descriptive statistics to understand the initial characteristics of the dataset.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
student_id	909	909	S1999	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	909.0	NaN	NaN	NaN	20.475248	2.302721	17.0	18.0	20.0	22.0	24.0
gender	909	3	Male	440	NaN	NaN	NaN	NaN	NaN	NaN	NaN
study_hours_per_day	909.0	NaN	NaN	NaN	3.538724	1.46973	0.0	2.5	3.5	4.5	8.3
social_media_hours	909.0	NaN	NaN	NaN	2.50462	1.164802	0.0	1.7	2.5	3.3	7.2
netflix_hours	909.0	NaN	NaN	NaN	1.830363	1.071251	0.0	1.0	1.8	2.6	5.4
part_time_job	909	2	No	713	NaN	NaN	NaN	NaN	NaN	NaN	NaN
attendance_percentage	909.0	NaN	NaN	NaN	83.880308	9.453622	56.0	77.5	84.2	90.7	100.0
sleep_hours	909.0	NaN	NaN	NaN	6.474037	1.218943	3.2	5.6	6.5	7.3	10.0
diet_quality	909	3	Fair	396	NaN	NaN	NaN	NaN	NaN	NaN	NaN
exercise_frequency	909.0	NaN	NaN	NaN	3.051705	2.035632	0.0	1.0	3.0	5.0	6.0
parental_education_level	909	3	High School	392	NaN	NaN	NaN	NaN	NaN	NaN	NaN
internet_quality	909	3	Good	410	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mental_health_rating	909.0	NaN	NaN	NaN	5.466447	2.857525	1.0	3.0	5.0	8.0	10.0
extracurricular_participation	909	2	No	620	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig. 1. Pre-processing Data Results

Overall, this pre-processing stage ensures that the dataset is ready for further analysis, such as data splitting (train-test split) and predictive model building.
The second stage in data processing is the dataset splitting process, which is done using an 80:20 ratio between training data and test data with the train-test split technique. The result is that 727 data points are used for training (training set) and 182 data points are used for testing (testing set). This division aims to ensure that the model can learn optimally from most of the data, while retaining some data to test the model's generalization ability against new data.
Next, in the third stage, the XGBoost model was trained with various parameters that had been set through a hyperparameter tuning process to obtain the best combination of values in reducing prediction errors.

```
XGBoost Performance:
RMSE = 0.335
MAE = 0.266
R² = 0.882
```

Fig. 2. XGBoost Model Training

These results show that the XGBoost approach is effective in the context of educational data mining, mainly due to its ability to handle data consisting of many variables and having non-linear relationships. Next is the model comparison stage, which is carried out to determine the performance of various machine learning models used to predict student academic performance based on learning habits, as shown in the figure below.

	Model	RMSE	R²
0	Linear Regression	0.325670	0.888134
2	XGBoost	0.334702	0.881842
1	Random Forest	0.348541	0.871870

Fig. 3. Model Comparison

The results show that Linear Regression has an RMSE value of 0.325670 and an R^2 value of 0.888134. This means that this model has the smallest error rate and the highest ability to explain data variation compared to the three models tested. XGBoost ranks second with an RMSE of 0.334702 and an R^2 of 0.881842. Although the difference is not too significant, these results show that XGBoost still performs well and is superior in capturing relationships between variables that cannot be fully explained by Linear Regression. Meanwhile, Random Forest produced the highest RMSE of 0.348541 with an R^2 of 0.871870, indicating slightly poorer performance compared to the other two models.

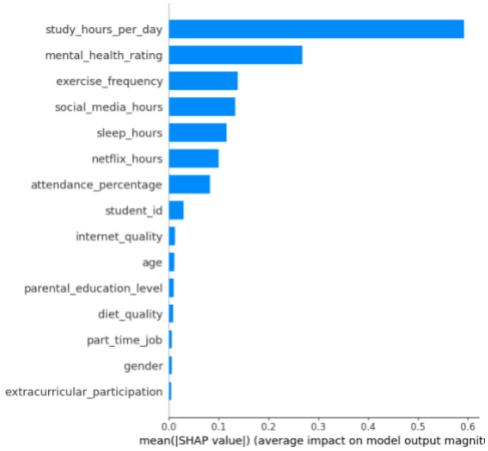

Fig. 4. SHAP Summary Plot Results

The graph above shows the fourth stage, which is the results of the SHAP summary plot. These results indicate that the higher the intensity of students' daily learning, the greater their contribution to improving academic performance. On the other hand, negative SHAP values for features such as hours spent on social media and hours spent watching Netflix indicate that the more time students spend on these activities, the more their academic performance declines.

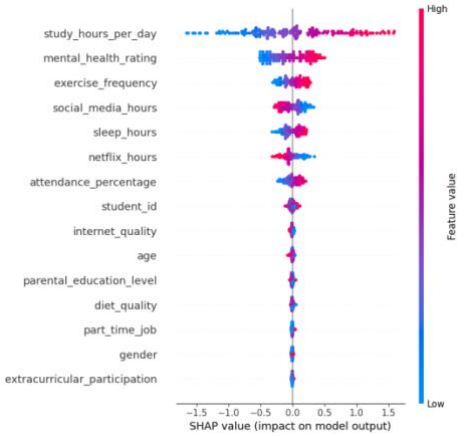

Fig. 5. Grafik Fitur Penting SHAP

The final stage can be seen in the image above, which shows the SHAP feature importance graph. The graph above explains the average contribution of each feature to the model. From these results, the study_hours_per_day feature emerges as the variable with the greatest influence, followed by mental_health_rating and exercise_frequency. These results reinforce the findings of previous analyses that study habits and mental health factors are key indicators in predicting student academic achievement.

From these two visualizations, it can be concluded that SHAP interpretation not only supports the accuracy of the XGBoost model but also provides practical guidelines for policymakers and educators in designing more targeted interventions. For example, educational institutions can emphasize the importance of managing study time, physical fitness, and mental balance for students as part of a strategy to improve academic outcomes.

Next is the final part of the quantitative analysis process, which aims to provide a general summary of the modeling results and explore the main features that influence student academic performance. This stage also serves to combine the model evaluation results with explanations of the important variables that influence the prediction results, as well as to confirm the best model found in the study.

```
==== Summary of Results ====
Best Model : Linear Regression
R² Score   : 0.888
RMSE       : 0.326

Top 5 Important Features:
                 Feature  Importance
3    study_hours_per_day    0.513670
13  mental_health_rating    0.166116
10    exercise_frequency    0.076423
4     social_media_hours    0.043201
5          netflix_hours    0.034990
```

Fig. 6. Summary of Model Results

The image above shows a summary of the model results obtained after the previous comparison and evaluation process. Based on these results, the Linear Regression model showed the best performance with an R^2 value of 0.888 and an RMSE of 0.326. An R^2 value close to 1 indicates that this model is able to explain approximately 88.8% of the variation in student test scores. The low RMSE value indicates a

small prediction error rate. These results show that the Linear Regression model has good generalization capabilities for test data compared to other models such as Random Forest or XGBoost.

In addition to model performance results, the analysis also displays the five most important features that contribute most to academic outcome predictions. The feature with the greatest influence is study_hours_per_day with an importance value of 0.513670, followed by mental_health_rating with a value of 0.166116, exercise_frequency of 0.076423, social_media_hours of 0.043021, and netflix_hours of 0.034990.

From this order, it can be concluded that study habits (number of hours of study per day) are the most dominant factor in determining student academic performance, followed by mental health and physical activity. This analysis shows a balance between cognitive and non-cognitive factors that influence academic outcomes.

Students who have regular study habits and good mental health tend to perform better academically. Conversely, excessive use of social media and digital entertainment has a negative impact on learning outcomes. Therefore, the results of this study not only demonstrate the effectiveness of the model in predicting academic performance, but also provide guidance for the development of data-driven learning strategies in higher education.

Overall, this stage forms the basis for drawing theoretical and practical conclusions from the research. By combining the quantitative results of the model and feature interpretation, educators, academic counselors, and policy makers can design intervention programs that focus on improving students' study habits and psychological well-being.

4. Conclusion

This study proves that the XGBoost algorithm with SHAP interpretability support is capable of predicting student academic achievement with high accuracy while providing transparent explanations. Learning habits and mental health factors have been proven to have a significant effect on academic achievement. For further research, it is recommended to use longitudinal datasets and integrate socioeconomic factors to improve the validity of the prediction model.

This study also has several limitations, including the use of secondary data sourced from public datasets with limited variables, without considering socioeconomic factors, intrinsic motivation, or learning quality. In addition, the developed model has not been tested longitudinally, so it does not describe changes in student learning behavior over time.

For further research, it is recommended to use a more diverse dataset sourced directly from educational institutions so that the results are more representative. The integration of online behavior data, psychological factors, and social context is also necessary to improve the accuracy and relevance of the model. The explainable machine learning approach, such as XGBoost–SHAP, remains important to ensure that the prediction results are not only accurate but also understandable and useful for educators and policy makers in improving the quality of learning in higher education.

Based on the research results and conclusions obtained, several recommendations can be drawn for future research development and implementation. The use of larger, more diverse, and longitudinal datasets needs to be considered so that prediction models have stronger generalization capabilities, including adding other variables such as learning motivation, cognitive style, and socioeconomic factors to improve accuracy. The findings of this study can also be used by educators to identify patterns of student learning habits at an earlier stage so that adaptive interventions for students at low risk of academic achievement can be provided more appropriately. At the institutional level, the application of a data-driven approach in academic policy-making is important, especially in planning learning support and mental health programs. XGBoost and SHAP-based prediction models have the potential to be integrated into academic information systems or LMS to provide automatic and transparent analysis of student performance, and can be further developed in the form of easy-to-understand visual dashboards. Furthermore, further research needs to pay more attention to the ethical aspects of student data use, including privacy, security, and anonymity, so that the use of machine learning technology remains in line with ethical research principles and responsible academic practices.

References

[1] A. Asselman, M. Khaldi, and S. Aammou, “Enhancing the prediction of student performance based on the machine learning XGBoost algorithm,” *Interact. Learn. Environ.*, vol. 31, no. 6, pp. 3360–3379, 2021.

[2] R. Ed-Daoudi, M. Azhari, B. Ettaki, and J. Zerouaoui, “Academic Performance Prediction in Virtual Environments Using Big Data and Machine Learning,” *J. Electr. Syst.*, vol. 20, no. 3, 2024.

[3] S. Wang and B. Luo, “Academic achievement prediction in higher education through interpretable modeling,” *PLoS One*, vol. 9, p. e0309838, 2024.

[4] M. Al-Okaily, S. Magatef, A. Al-Okaily, and F. S. Shiyyab, “Exploring the factors that influence academic performance in Jordanian higher education institutions,” *Heliyon*, vol. 10, no. 13, 2024.

[5] K. Mukesh Kumar, N. Singh, J. Wadhwa, P. Singh, G. Kumar, and A. Qtaishat, “Utilizing Random Forest and XGBoost data mining algorithms for anticipating students’ academic performance,” *Int. J. Mod. Educ. Comput. Sci.*, vol. 16, no. 2, pp. 29–44, 2024.

[6] Tao-Hongli, “Educational data mining for student performance prediction: feature selection and model evaluation,” *J. Electr. Syst.*, vol. 20, no. 3, 2024.

[7] Z. Ersozlu, S. Taheri, and I. Koch, “A review of machine learning methods used for educational data,” *Educ. Inf. Technol.*, vol. 29, pp. 22125–22145, 2024.

[8] J. Lu *et al.*, “Machine learning analysis of factors affecting college students’ academic performance,” *Front. Psychol.*, vol. 15, 2024.

[9] A. J. L. Brambila-Tapia, E. U. Velarde-Partida, L. A. Carrillo-Delgadillo, S. Ramírez-De los Santos, and F. Macías-Espinoza, “Correlation between studying strategies, personal and psychological factors with academic achievement and intelligence in health sciences university students: A cross-sectional study,” *BMC Med. Educ.*, vol. 24, p. 881, 2024.

[10] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, 2018, vol. 30, pp. 4765–4774.

[11] M. Lünich and B. Keller, “Explainable artificial intelligence for academic performance prediction: An experimental study on the impact of accuracy and simplicity of decision trees on causability and fairness perceptions,” *Front. Artif. Intell.*, vol. 7, p. 1486392, 2024.