ISSN: 2581-1274

Analysis and Prediction of Student Academic Performance Using the Random Forest Algorithm

Donna Oktar Endras Wantos[a,1,*], Dede Sahrul Bahri[a,2]

[a] University of Pamulang, Jl. Raya Puspitek, South Tangerang 15310, Indonesia

[1] *dosen03269@unpam.ac.id*;* [2] *dosen00271@unpam.ac.id*;*
** corresponding author*

ARTICLE INFO

Article history:
Published
April 12, 2026

Keywords:

Academic performance
Machine learning
Random forest
Prediction
Education

ABSTRACT

Student academic performance prediction has become an important topic in educational data mining, as it supports early intervention and data-driven decision-making. However, previous studies often lack comprehensive analysis and practical implementation of predictive models. This study aims to analyse and predict students' academic performance using the Random Forest algorithm. The dataset consists of 649 student records with 33 attributes, including demographic, behavioural, and academic variables. The final grade (G3) is used as the target variable in a classification task. The research process includes data exploration, preprocessing, feature transformation, train-test splitting, and model development using Random Forest. The model achieves an accuracy of 92.31%, with an average cross-validation accuracy of 91.21%, indicating stable performance. Additional evaluation using a confusion matrix shows that most instances are correctly classified, although some misclassifications occur in borderline cases. Feature importance analysis reveals that previous academic grades (G2 and G1) are the most influential predictors. Despite the strong performance, the model is limited by the dataset characteristics, particularly its relatively homogeneous distribution, which may affect generalization. Future work can explore more diverse datasets and alternative models to improve robustness. Overall, this study demonstrates the potential of the Random Forest algorithm in predicting students' academic performance and supporting educational decision-making.

I. Introduction

The rapid development of information technology has significantly impacted various sectors, including education [1]. The use of technology in education is no longer limited to learning media but has expanded into data-driven tools that support decision-making processes. One important application is the prediction of students' academic performance, which can be used to identify students at risk of academic decline and enable early intervention to improve learning outcomes [2]. With the increasing volume of data stored in academic information systems, there is a growing need for analytical methods capable of transforming large datasets into meaningful information. Data mining techniques have been widely applied to identify patterns and relationships among factors influencing academic performance, such as family background, learning behaviour, and previous academic achievement. In this context, machine learning has emerged as an effective approach due to its ability to build predictive models based on historical data [3].

Previous studies have applied various machine learning algorithms to predict students' academic performance, including Decision Tree, Naive Bayes, and Random Forest [4]. However, several limitations remain. Many studies lack in-depth analysis of relationships among variables, do not integrate predictive models into practical systems, and provide limited comparative evaluation between different methods. These limitations highlight the need for research that not only develops accurate predictive models but also offers more comprehensive and applicable analysis.

Random Forest is one of the widely used machine learning algorithms due to its ability to handle complex data, improve prediction accuracy, and reduce overfitting through the combination of

multiple decision trees [5]. Nevertheless, its application in the educational domain still requires further investigation, particularly in identifying the most influential factors affecting students' academic performance.

The dataset used in this study includes various attributes representing student characteristics, such as age, gender, family background, social activities, health condition, attendance, and academic grades (G1, G2, and G3). Table 1 presents a sample of the dataset used in this study.

Table 1. Student Academic Performance Data

No	School	Sex	Age	Address	FamSize	Pstatus	Medu	Fedu	Mjob	Fjob	FamRel	FreeTime	GoOut	Dalc	Walc	Health	Absences	G1	G2	G3
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	4	3	4	1	1	3	4	0	11	11
2	GP	F	17	U	GT3	T	1	1	at_home	other	5	3	3	1	1	3	2	9	11	11
3	GP	F	15	U	LE3	T	1	1	at_home	other	4	3	2	2	3	3	6	12	13	12
4	GP	F	15	U	GT3	T	4	2	health	services	3	2	2	1	1	5	0	14	14	14
5	GP	F	16	U	GT3	T	3	3	other	other	4	3	2	1	2	5	0	11	13	13

Table 1. Shows that each row represents an individual student record, while each column corresponds to a specific feature used in the analysis. The dataset includes important factors such as demographic characteristics, family background, social behavior, and academic indicators. From the sample data, it can be observed that the final grade (G3) tends to follow the pattern of previous grades (G1 and G2), indicating that prior academic performance is a strong predictor. Additionally, factors such as attendance (absences) and family conditions (famrel, Medu, and Fedu) also show potential influence on students' academic outcomes.

Despite the dataset containing relevant attributes, it has limitations in terms of representativeness, as the data tends to be relatively homogeneous in certain aspects, such as school type and residential area. This condition may introduce potential bias in the predictive model and should be considered when interpreting the results.

Based on these issues, this study aims to develop a predictive model for students' academic performance using the Random Forest algorithm and to analyse the factors influencing academic outcomes. The main contribution of this study is providing a more structured analysis of student academic data and demonstrating the potential application of machine learning to support decision-making in the educational field. The results are expected to provide insights for improving learning strategies and serve as a reference for future research in educational data mining.

II. The Proposed Method/Algorithm

This study adopts a quantitative approach using a machine learning method to analyze and predict students' academic performance based on the Random Forest algorithm. The dataset consists of 649 student records with 33 attributes, including demographic characteristics, family background, behavioral factors, and academic performance indicators. The final grade (G3) is used as the target variable in this study [9]. Prior to model development, a data validation process was conducted to ensure data quality. Missing value detection was performed through data inspection procedures, confirming that all attributes contain complete values without missing data. Therefore, no imputation process was required. Furthermore, categorical variables such as school, sex, address, and job-related attributes were transformed into numerical values using label encoding techniques to ensure compatibility with the machine learning algorithm.

The research process begins with exploratory data analysis to understand the characteristics of the dataset. Descriptive statistical analysis and data visualization were used to examine the

distribution of variables. The results indicate that the average student age is approximately 16.74 years, and the academic performance variables tend to follow a normal distribution within a moderate range

[10]. In the preprocessing stage, feature selection was conducted to identify variables relevant to academic performance prediction. The selected features include demographic attributes, family background variables, behavioral indicators, and previous academic grades. To improve model performance, all numerical features were normalized to ensure consistent scale across variables.

The dataset was then divided into training data (80%) and testing data (20%) using a random state value of 42 to ensure reproducibility. The Random Forest model was constructed using several important hyperparameters, including the number of trees (n_estimators = 100), maximum tree depth (max_depth), and random_state = 42. These hyperparameters were selected to balance model complexity and prediction accuracy while minimizing the risk of overfitting.

Model training was performed using the training dataset, and predictions were generated on the testing dataset. Model performance was evaluated using multiple evaluation metrics, including accuracy score, confusion matrix, and classification report. The evaluation results indicate that the Random Forest model achieved an accuracy of 92%, demonstrating strong predictive capability with relatively low error rates [11].

To further assess model robustness, k-fold cross-validation with five folds was applied. The results show an average accuracy of 91.21%, indicating that the model has good stability and generalization ability across different subsets of data. In addition, feature importance analysis was conducted to identify the most influential variables in predicting academic performance. The analysis reveals that previous academic grades (G2 and G1) are the most significant predictors of the final grade (G3), followed by absenteeism and other behavioral factors.

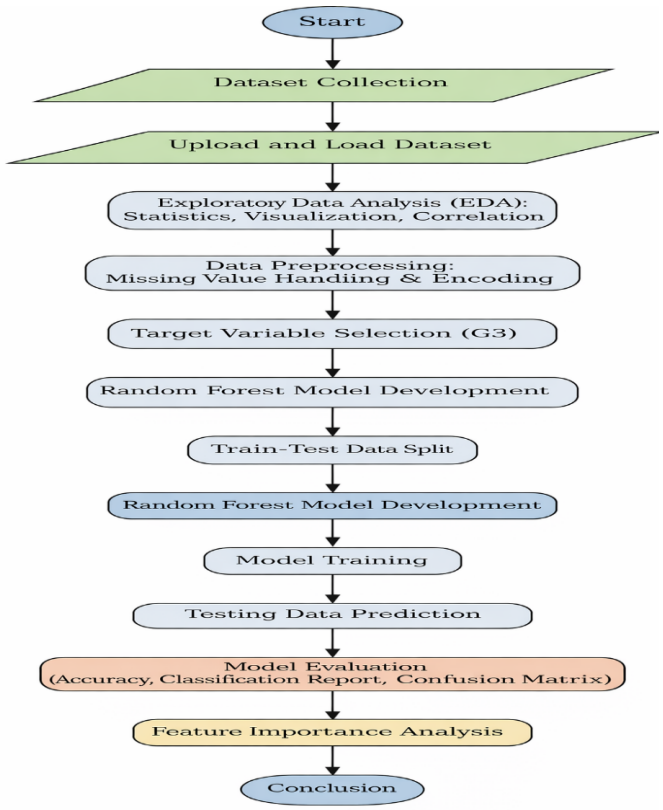

Fig 1. Research Flowchart

Figure 1. Illustrates the overall research workflow using the Random Forest algorithm. The process begins with data input, where the student academic performance dataset is collected and prepared. This is followed by the data preprocessing stage, which includes data validation, handling of categorical variables through encoding, and feature selection to ensure that relevant attributes are used in the model. After preprocessing, the dataset is divided into training and testing subsets. The training data is used to build the Random Forest model by generating multiple decision trees and

aggregating their results. The trained model is then applied to the testing data to produce predictions. The final stage involves model evaluation using performance metrics such as accuracy, confusion matrix, and classification report. This workflow ensures that the prediction model is developed systematically and can provide reliable results for academic performance prediction.

III. Results and Discussion

3.1 Dataset Collection

In this study, a student academic performance dataset consisting of 649 records with 33 attributes is used. The dataset includes variables related to student characteristics, family background, social behavior, and academic performance. The final grade (G3) is used as the prediction target. The dataset is considered appropriate for this study due to its comprehensive representation of factors influencing academic performance. However, potential limitations related to data representativeness should be considered, as the dataset may not fully capture variations across different educational environments.

3.2 Uploading and Loading the Dataset

The dataset is processed using Google Colab with the Python programming language. Data structure inspection is performed to verify the number of attributes, data types, and data completeness. The dataset consists of 16 numerical attributes and 17 categorical attributes. Missing value validation is conducted using data inspection techniques, confirming that no missing values are present. This ensures that the dataset can be directly used for modelling without additional imputation. The presence of both numerical and categorical variables requires appropriate preprocessing, particularly encoding techniques, to ensure compatibility with machine learning algorithms.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is conducted to understand the dataset characteristics through statistical analysis and visualization. The descriptive statistics show that the average student age is 16.74 years, with a range between 15 and 22 years. The mean academic scores are 11.39 (G1), 11.57 (G2), and 11.91 (G3), indicating relatively consistent academic performance across different evaluation stages.

Table 2. Descriptive Statistics of the Dataset

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Age	649	16.74	1.22	15	16	17	18	22
Medu	649	2.51	1.13	0	2	2	4	4
Fedu	649	2.31	1.10	0	1	2	3	4
Traveltime	649	1.57	0.75	1	1	1	2	4
Studytime	649	1.93	0.83	1	1	2	2	4
Failures	649	0.22	0.59	0	0	0	0	3
Famrel	649	3.93	0.96	1	4	4	5	5
Freetime	649	3.18	1.05	1	3	3	4	5
Goout	649	3.18	1.18	1	2	3	4	5
Dalc	649	1.50	0.92	1	1	1	2	5
Walc	649	2.28	1.28	1	1	2	3	5
Health	649	3.54	1.45	1	2	4	5	5
Absences	649	3.66	4.64	0	0	2	6	32
G1	649	11.40	2.75	0	10	11	13	19
G2	649	11.57	2.91	0	10	11	13	19
G3	649	11.91	3.23	0	10	12	14	19

Table 2. Presents the descriptive statistics of the dataset, providing an overview of student characteristics and academic performance. The results indicate that the average student age is 16.74 years, suggesting that the dataset mainly represents students in a typical high school age range.The mean values of G1, G2, and G3 (11.40, 11.57, and 11.91) show a consistent pattern of academic performance, indicating that students' final grades are strongly influenced by their previous academic achievements. This supports the assumption that historical academic data is a key predictor in the model. The low average value of the failure variable (0.22) suggests that most students do not have a history of academic failure, while the relatively low number of absences (3.66) indicates good attendance behavior. These factors are likely to contribute positively to academic performance. In addition, parental education levels (Medu and Fedu) are within a moderate range, indicating a balanced socio-educational background. Behavioral variables such as free time and social activities also show moderate values, suggesting that students maintain a balance between academic and social life. Overall, the dataset demonstrates sufficient variability and balanced distribution, making it suitable for predictive modelling. However, the relatively concentrated distribution of academic scores may limit the model's ability to distinguish extreme cases, which should be considered in further analysis.

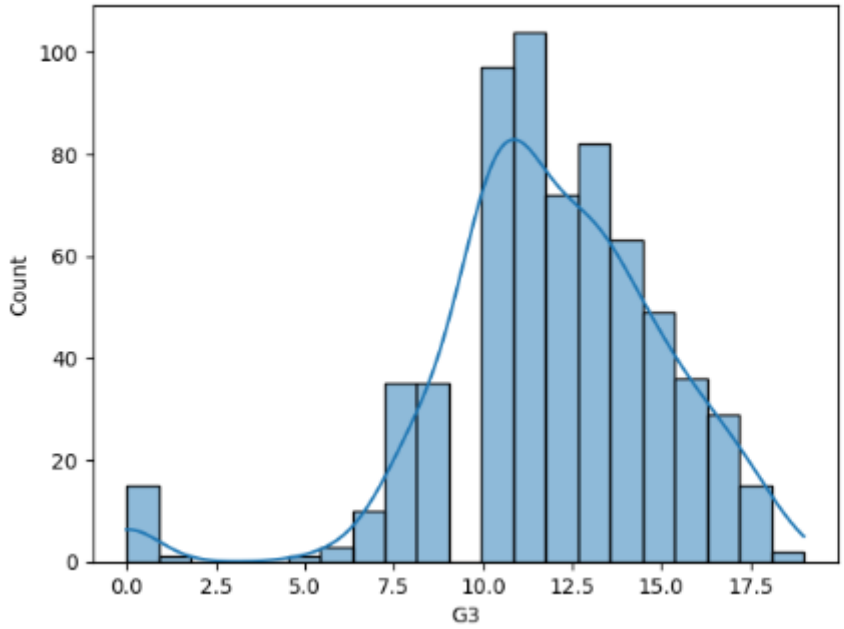

Fig 2. Distribution of final grades

Figure 2. shows that the distribution of students' final grades (G3) approximates a normal distribution, with most values concentrated between 10 and 15. This indicates that the dataset is relatively balanced and does not suffer from severe class imbalance. However, the presence of extreme values near 0 and 19 suggests that there are outliers within the dataset. These outliers may affect model performance, particularly in classification tasks where borderline cases are difficult to distinguish. The concentration of values in the mid-range also implies that the model may have limitations in differentiating high-performing and low-performing students. Therefore, robust algorithms such as Random Forest are required to handle such data variability effectively.

3.4 Data Preprocessing and Determination of Prediction Target

The preprocessing stage is conducted to prepare the dataset for modelling. Missing value validation is performed using data inspection techniques, confirming that no missing values are present. Categorical variables such as school, sex, and address are transformed into numerical representations using label encoding to ensure compatibility with machine learning algorithms. In addition, feature selection is applied to identify relevant variables that contribute to academic performance prediction. The target variable (G3) is transformed into a binary classification (pass/fail) to simplify the prediction task and improve model interpretability.

3.5 Training and Testing Data Split

The dataset is divided into training data (80%) and testing data (20%) using a random state of 42 to ensure reproducibility. This split allows the model to learn from the training data while maintaining an independent dataset for performance evaluation.

3.6 Random Forest Model Development

The prediction model is developed using the Random Forest algorithm. Key hyperparameters include n_estimators = 100 and random_state = 42, which are selected to balance model accuracy and generalization performance.Random Forest is chosen due to its ability to handle complex data patterns, reduce overfitting, and improve prediction stability through ensemble learning.

3.7 Model Training

The training process involves constructing multiple decision trees using the training dataset. The predictions from these trees are aggregated to produce a final output, improving classification accuracy and reducing variance.

3.8 Testing Data Prediction

The trained model is applied to the testing dataset to generate predictions. These predictions are then compared with actual values to evaluate the model's classification performance.

3.9 Model Evaluation

Model performance is evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and confusion matrix. While accuracy provides an overall performance measure, additional metrics are used to assess the model's effectiveness in handling class imbalance and prediction errors.The confusion matrix is used to analyze classification results in detail, identifying correct predictions and misclassification patterns. This analysis helps reveal potential weaknesses in the model, particularly in distinguishing borderline cases between classes.

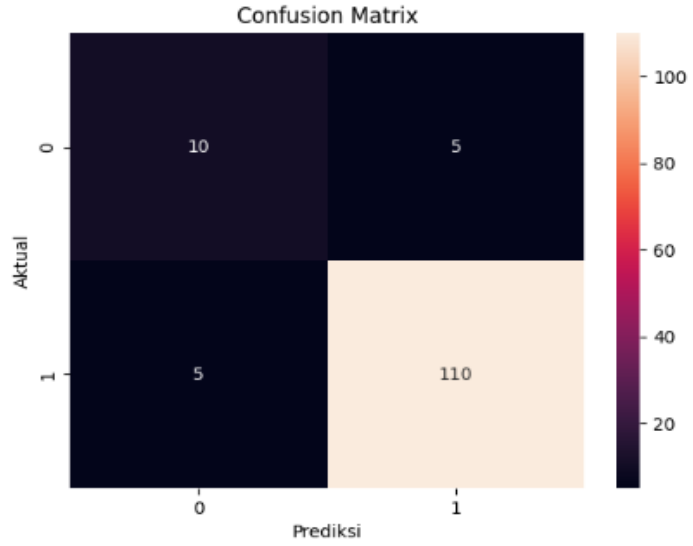

Fig 3. Confusion matrix of the Random Forest model

Figure 3. presents the confusion matrix of the Random Forest model. The model correctly classified 110 instances in the pass category and 10 instances in the fail category. However, misclassifications are still observed, with 5 false positives and 5 false negatives. These misclassification results indicate that the model encounters difficulty in distinguishing borderline cases, particularly when the feature values between pass and fail categories overlap. This suggests that some features may not provide sufficient discriminatory power for certain instances. Despite the high number of correct predictions, the presence of both false positives and false negatives highlights

potential limitations of the model. False negatives (students predicted as pass but actually fail) are particularly critical, as they may lead to missed early intervention opportunities. Overall, the model demonstrates strong performance; however, further improvement can be achieved by optimizing feature selection or incorporating additional relevant variables.

3.10 Feature Importance Analysis

Feature importance analysis is conducted to identify the most influential variables in predicting students' academic performance. This analysis provides insights into the contribution of each feature to the prediction results.

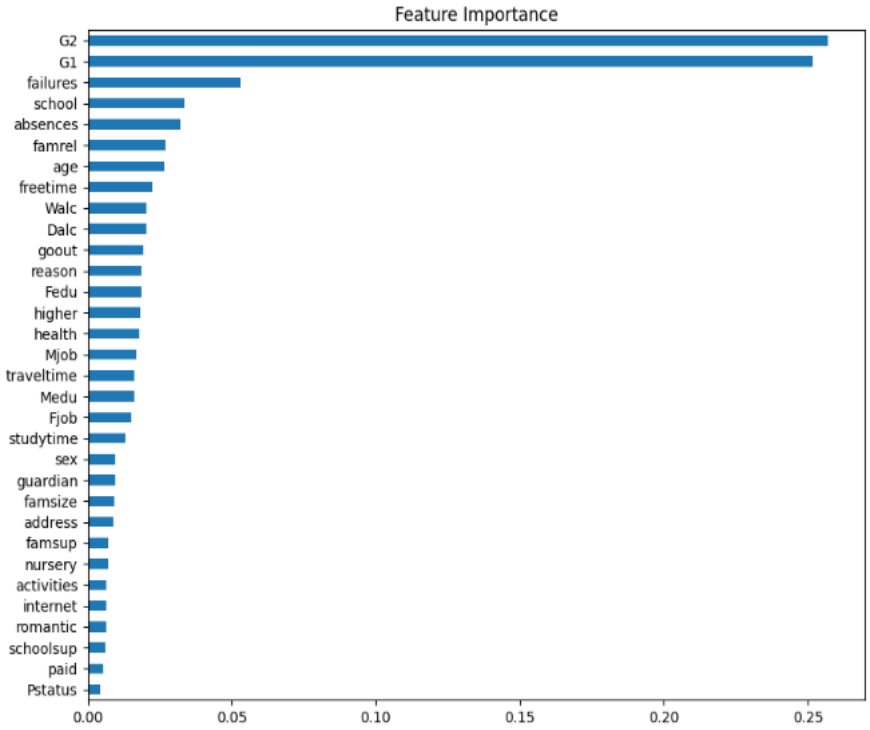

Fig 4. Feature importance of the Random Forest model

Figure 4. shows that previous academic performance variables, particularly G2 and G1, are the most influential predictors of the final grade (G3). This indicates a strong dependency between historical and future academic performance, confirming that prior achievement is a key determinant in predictive modelling. In addition, variables such as failures and absences also contribute significantly, suggesting that academic discipline and attendance play important roles in student performance. These findings highlight that behavioural factors, combined with academic history, influence learning outcomes. On the other hand, variables such as family background, age, and social activities have relatively lower importance, indicating that their impact is less significant within this dataset. These results suggest that predictive models in education should prioritize academic history and behavioural indicators to improve prediction accuracy. However, the dominance of G1 and G2 may also indicate potential redundancy, which could be further explored in future studies through feature selection optimization.

IV. Conclusion

Based on the results of this study, the Random Forest algorithm demonstrates strong performance in predicting students' academic performance, achieving an accuracy of 92.31%. The evaluation results, including the confusion matrix and cross-validation (average accuracy of 91.21%), indicate that the model is capable of producing consistent and reliable predictions within the given dataset. However, the findings should be interpreted with caution. The model heavily relies on previous academic grades (G1 and G2), which may limit its generalization when applied to different datasets or educational contexts. In addition, the dataset used in this study may have limitations in terms of

representativeness, as it reflects a relatively homogeneous data distribution. This condition may also introduce potential bias and affect model performance in broader applications. Furthermore, although the model shows stable performance, the possibility of overfitting cannot be entirely ruled out, especially due to the dominance of certain predictive features. Therefore, further validation using more diverse datasets is necessary to ensure the robustness of the model. The feature importance analysis confirms that prior academic performance, along with attendance and academic failure history, are key factors influencing students' final grades. These findings highlight the importance of combining academic and behavioural indicators in predictive modelling. For future work, it is recommended to explore additional features, apply alternative machine learning algorithms, and utilize larger and more diverse datasets to improve model generalization. Moreover, integrating the predictive model into a real-world application system could enhance its practical usefulness in supporting data-driven decision-making in education.

Acknowledgment

The authors would like to express their gratitude to all parties who have provided support and assistance so that this research could be completed successfully. It is hoped that the results of this study will provide benefits and contribute to the development of knowledge, particularly in the fields of information technology and healthcare.

References

[1] L. H. Alamri, R. S. Almuslim, M. S. Alotibi, D. K. Alkadi, I. U. Khan, and N. Aslam, "Predicting Student Academic Performance Using Support Vector Machine and Random Forest," ACM International Conference Proceedings Series, vol. PartF168981, pp. 100–107, 2020.

[2] M. Nachouki and M. A. Naaj, "Predicting Student Performance to Improve Academic Advising Using the Random Forest Algorithm," International Journal of Distance Education Technologies, vol. 20, no. 1, pp. 1–17, 2022.

[3] C. Beaulac and J. S. Rosenthal, "Predicting University Students' Academic Success and Major Using Random Forests," Research in Higher Education, vol. 60, no. 7, pp. 1048–1064, 2019.

[4] J. Fathurahman, D. Hartanti, and S. Sopingi, "Sentiment Analysis of Joglo Wifi UMKM Service Using the Naive Bayes Method," Jurnal Inotera, vol. 9, no. 2, pp. 370–377, 2024.

[5] Y. Abubakar, N. Bahiah, and H. Ahmad, "Prediction of Students' Performance in an E-Learning Environment Using Random Forest," International Journal of Innovative Computing, vol. 7, no. 2, pp. 1–5, 2017.

[6] R. Amin and A. S. F. Utami, "Prediction of Exam Scores Based on Student Habits Using the Random Forest Regressor Algorithm," Information Systems Education Professionals Journal, vol. 10, no. 2, p. 149, 2025.

[7] M. H. Sukri and Y. Handrianto, "Application of the C4.5 Algorithm in Predicting Student Achievement at SMPN 51 Jakarta," Informatics and Computer Engineering Journal, vol. 4, no. 1, pp. 11–24, 2024.

[8] R. A. Saputri, A. Asrianda, and L. Rosnita, "A Random Forest-Based Predictive Model for Student Academic Performance: A Case Study in Indonesian Public High Schools," Journal of Applied Informatics and Computing, vol. 9, no. 3, pp. 1042–1049, 2025.

[9] R. Z. Arifin, H. Firmansyah, and W. Asriyani, "Prediction of Student Graduation Based on Demographic and Academic Data Using the Student Performance Dataset," Journal of Community Service and Educational Research, vol. 4, no. 2, pp. 13300–13307, 2025.

[10] A. Fatunnisa and H. Marcos, "Prediction of On-Time Graduation for Computer Engineering Vocational School Students Using the Random Forest Algorithm," Journal of Informatics Management, vol. 14, no. 4, pp. 101–111, 2024.

[11] Jupron and Sutrisno, "Analysis of Heart Disease Using the Random Forest Method," Jurnal Inotera, vol. 10, no. 1, pp. 167–174, 2025.