Terbit online pada laman web jurnal: **https://rjti.upp.ac.id/index.php/rjmldm**

RJTI
Riau Jurnal Teknik Informatika

Vol. 4 No. 3 (2025) 441– 445 | ISSN: 2829-2529(Media Online)

Predicting Student Academic Performance Using Learning Activity Data: A Comparative Study of Random Forest and Decision Tree Models

Rahmat Hidayat[1], Herwis Gultom[2], Yuda Samudra[3]
[1,2,3]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang
[1]dosen02675@unpam.ac.id , [2]dosen02535@unpam.ac.id, [3]dosen02623@unpam.ac.id

Abstract

This study compares the effectiveness of the Random Forest and Decision Tree algorithms in predicting students' academic performance based on learning activities. The data used included reading scores, writing scores, math scores, and demographic variables such as gender, race/ethnicity, parental level of education, lunch, and test preparation course. The research was carried out through the stages of data cleaning, training and test data sharing, model training, and evaluation using confusion matrix and accuracy, precision, recall, and F1-score metrics. The results show that Random Forest performs best with 97% accuracy, surpassing Decision Tree which has 94% accuracy. The feature importance analysis revealed that cognitive ability—especially in the reading score, writing score, and math score features—had the greatest influence on prediction results. These findings confirm that the Random Forest model is more reliable and effective as a prediction tool in the academic decision support system to detect the potential for decline in student achievement early.

Keywords: Academic Performance, Learning Activities, Random Forest, Decision Tree, Machine Learning

Abstrak

Penelitian ini membandingkan efektivitas algoritma *Random Forest* dan *Decision Tree* dalam memprediksi kinerja akademik mahasiswa berdasarkan aktivitas belajar. Data yang digunakan mencakup nilai *reading score, writing score, math score,* serta variabel demografis seperti *gender, race/ethnicity, parental level of education, lunch,* dan *test preparation course*. Penelitian dilakukan melalui tahapan pembersihan data, pembagian data latih dan uji, pelatihan model, serta evaluasi menggunakan *confusion matrix* dan metrik akurasi, *precision, recall*, dan *F1-score*. Hasil menunjukkan bahwa *Random Forest* menghasilkan performa terbaik dengan akurasi 97%, melampaui *Decision Tree* yang memiliki akurasi 94%. Analisis *feature importance* mengungkap bahwa kemampuan kognitif—terutama pada fitur *reading score, writing score*, dan *math score*—memiliki pengaruh terbesar terhadap hasil prediksi. Temuan ini menegaskan bahwa model *Random Forest* lebih andal dan efektif sebagai alat bantu prediksi dalam sistem pendukung keputusan akademik untuk mendeteksi potensi penurunan prestasi mahasiswa secara dini.

Kata kunci: Kinerja Akademik, Aktivitas Belajar, *Random Forest, Decision Tree, Machine Learning*

1. Introduction

Student academic performance is one of the main indicators in assessing the success of the educational process in higher education. Academic results reflect the extent to which students are able to understand the lecture material, adapt to the academic environment, and implement effective learning strategies. In the context of modern education, especially with the increasing use of digital learning systems (Learning Management System), student learning activities can be recorded and analyzed comprehensively[1][2].

However, the problem that often arises is the many factors that affect academic achievement, making it difficult for lecturers or the campus to identify students who have the potential to experience an early decline in performance. Factors such as cognitive ability, learning style, socioeconomic conditions, and family support contribute to complex learning outcomes that are difficult to analyze manually. Therefore, a data-driven approach is needed that is able to utilize student learning activity data to build accurate and reliable predictive models.

Activity data such as frequency of attendance, participation in online discussions, assignment results, and test scores can provide valuable information to predict students' future academic performance[3].

Research by Rifa Andriani Saputri, Asrianda, and Lidya Rosnita (2025) shows that the Random Forest algorithm is effective in predicting the academic achievement of

students in SMAN 1 and SMAN 3 Aceh Barat Daya with an accuracy of 87.40%[4], while Rohman et al. (2025) developed a Hybrid Logistic Regression–Random Forest approach with SMOTE techniques and hyperparameter tuning, which increases the accuracy by up to 95.74%. These two studies prove that Random Forest-based models—both single and hybrid—have great potential in supporting accurate data-driven academic prediction systems that can be used to assist educational institutions in making strategic decisions and early detection of the risk of declining student performance[5][6].

Based on research by Micheline A. Gotardo and Md. Mahadhi Hasan et al, the application of the Decision Tree algorithm has been proven to be able to provide accurate results in predicting student academic performance. Gotardo's research (2019) shows that the J48 Decision Tree model can identify important factors such as Midterm and Final scores as the main determinants of student success in the Data Structures and Algorithms course, with an accuracy rate of 91.67%. Meanwhile, the research of Hasan et al. (2025) combines Decision Tree and K-Means Clustering methods to analyze academic performance while identifying student learning behavior in college, with a prediction accuracy of above 98%. Based on these two studies, the use of data mining algorithms such as Decision Tree is not only effective in predicting learning outcomes, but also provides valuable insights for educational institutions to recognize student behavior patterns, classify learner types (active, passive, and risky), and design adaptive learning strategies to improve the quality and effectiveness of the educational process[7], [8].

Based on research by Kesgin et al, Gufroni et al, and Nurbaeti et al, the Decision Tree and Random Forest algorithms have proven to be effective in predicting student academic performance. Kesgin et al. highlight the importance of accurate and fair models through the integration of SMOTE, SHAP, and adversarial debiasing. Gufroni et al. found that Random Forest achieved the highest accuracy of 79%, while Decision Tree excelled in interpretability. Meanwhile, Nurbaeti et al. showed that Random Forest with a Variance Threshold yielded an accuracy of 0.77 and an F1-score of 0.84, better than Decision Tree. Overall, the study confirms that Random Forest is the most superior model for accurate, stable, and applicable academic predictions in data-driven educational decision support systems.[9], [10], [11]

The main advantage of the Random Forest method over Decision Tree lies in its ability to combine many simple models to produce more stable predictions. Random Forest is also more resistant to data noise and class imbalances, as seen in the distribution of the data in this study where the performance category is dominating. In addition, this algorithm provides feature importance analysis that helps understand which variables have the most influence on academic results, so that it can be used as a basis for decision-making in the development of more effective learning strategies[12][13].

Based on this background, this study focuses on the analysis of student academic performance predictions based on learning activities by comparing the performance of the Random Forest and Decision Tree algorithms. This research is expected to contribute to the development of an academic prediction system that is able to identify students with a risk of declining achievement early[14]. Thus, educational institutions can design more targeted interventions, such as tutoring programs or academic mentoring, to improve the quality of learning and overall academic outcomes of students.

2. Research Methods

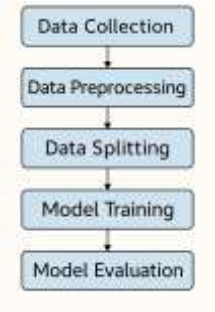

Figure 1. Stages of Research Method

The explanation of the stages of the research method above is:

a. Data Collection

The data used in this study was obtained from student academic results which included cognitive and non-cognitive ability variables. The features used include reading score, writing score, math score, gender, race/ethnicity, parental level of education, lunch, and test preparation course. The total amount of data analyzed was 1,000 students, with three categories of academic performance, namely low (0), medium (1), and high (2). The distribution image shows that the middle category class dominates with 510 students, the low category with 297 students, and the high category with 193 students. This unbalanced distribution is one of the challenges in the model training process because it can affect classification results, especially in categories with a smaller amount of data[15].

b. Data Preprocessing

The data preprocessing stage is carried out to improve the quality of the data so that it is ready to be used in the model training process. This step includes data cleaning from empty values, duplication, or input errors, as well as transforming the data into numerical format so that it can be processed by machine learning algorithms[16]. In addition, data normalization or standardization is carried out to avoid the dominance of certain features. This process is very important because good data quality will

DOI : https://doi.org/10.52158/rjti.v4i3.4032

have a direct impact on the model's performance and the accuracy of the prediction results.

c. Data Splitting

At this stage, the dataset is divided into two parts, namely training data and test data (test set) with a general ratio such as 80:20 or 70:30. Drill data is used to build and customize models, while test data is used to measure the model's ability to predict new data that has never been seen before. This division aims to avoid overfitting, which is a condition in which the model is over-adjusted to the training data and fails to generalize to new data.

d. Model Training

This stage is at the heart of the research process, where two algorithms, namely Decision Tree and Random Forest, are used to build prediction models. Decision Tree establishes a Decision Tree structure based on data sharing rules that minimize error rates, while Random Forest combines multiple decision trees to produce more stable and accurate prediction results. The training process is carried out using training data, where the learning model recognizes the pattern of relationships between input variables (features) and outputs (student academic performance categories).

e. Model Evaluation

The final stage is model evaluation to assess how well the algorithm is able to predict the test data. Evaluation was carried out using metrics such as accuracy, precision, recall, and F1-score, as well as confusion matrix analysis to see the distribution of predicted results for each class. In addition, feature importance analysis was carried out to determine the features that had the most influence on the prediction results. Based on the results of the study, the Random Forest model showed the best performance with an accuracy of 97%, indicating that this model is more effective in predicting academic performance than the Decision Tree.

3. Results and Discussion

a. Pemprosesan Data

Data setelah encoding:

	gender	race/ethnicity	parental level of education	lunch	\
0	0	1	1	1	
1	0	2	4	1	
2	0	1	3	1	
3	1	0	0	0	
4	1	2	4	1	

	test preparation course	math score	reading score	writing score
0	1	72	72	74
1	0	69	90	88
2	1	90	95	93
3	1	47	57	44
4	1	76	78	75

Figure 2. Data Process

The image above shows the results of the encoding process or transformation of categorical data into numerical data that is carried out at the preprocessing stage before the machine learning model is implemented. This process aims to allow algorithms such as Random Forest and Decision Tree to process all variables mathematically. Each column represents the features used in the research, such as gender, race/ethnicity, parental level of education, lunch, test preparation course, and academic scores in the form of math scores, reading scores, and writing scores. Categorical values such as gender, ethnicity, parental education, and lunch status are converted into numbers to standardize the data format. For example, in the first row, it is seen that students with gender 0, taking exam preparation courses, and having a math score of 72, reading 72, and writing 74. The results of this encoding show that the data is ready to be used for the model training stage because all variables are in a uniform numerical form and can be processed by the prediction algorithm.

b. Evaluasi Model

=== Classification Report: Decision Tree ===

	precision	recall	f1-score	support
Low	0.94	0.95	0.94	62
Medium	0.95	0.95	0.95	109
Hingh	0.96	0.90	0.93	29
accuracy				200
macro avg	0.95	0.93	0.94	200
weighted avg	0.95	0.94	0.94	200

=== Classification Report: Random Forest ===

	precision	recall	f1-score	support
Low	0.94	0.98	0.96	62
Medium	0.98	0.96	0.97	109
Higgh	1.00	0.97	0.98	29
weighted avg	0.95	0.97	0.97	200

Figure 3. Model Evaluation

The image above shows a comparison of the Classification Report between the Decision Tree and Random Forest models in predicting student academic performance. Based on the evaluation results, the Decision Tree model obtained an accuracy of 94%, with an average precision and recall values of 0.95 and 0.93, while the Random Forest model showed superior performance with an accuracy of 97% and an average precision, recall, and F1 score values of 0.97. In all categories—low, medium, and high—Random Forest produced higher F1-scores, demonstrating the model's better ability to balance precision and sensitivity. These results prove that ensemble learning methods such as Random Forest are more effective than single models such as Decision Tree in handling data with high variation and unbalanced class distribution. Overall, Random Forest is able to provide more accurate, stable, and generalist predictions, making it more recommended for use in student academic performance prediction systems.

c. Confusion Matriks

DOI : https://doi.org/10.52158/rjti.v4i3.4032

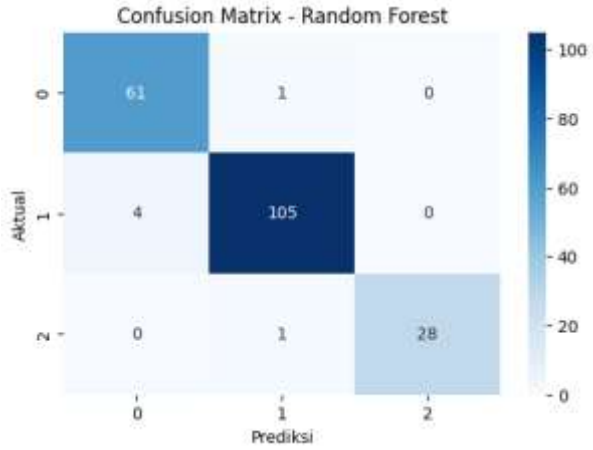

Figure 4. Confusion Matrik Random Forests

The image above shows the Confusion matrix of the Random Forest model used to predict the academic performance categories of students. This matrix shows a comparison between the actual value and the model's prediction results in three classes, namely 0 (low), 1 (medium), and 2 (high). Out of a total of 200 test data, the model correctly predicted 61 students in the low grade, 105 in the medium class, and 28 in the high grade, while the prediction error only occurred in a few cases, namely 1 low grade student who was predicted to be moderate, 4 students in the middle class who was predicted to be low, and 1 high class student who was predicted to be moderate. Overall, the model shows a very high level of accuracy with a very small number of misclassifications. This shows that the Random Forest algorithm is able to recognize data patterns well and provide accurate and stable prediction results in each category of student academic performance.

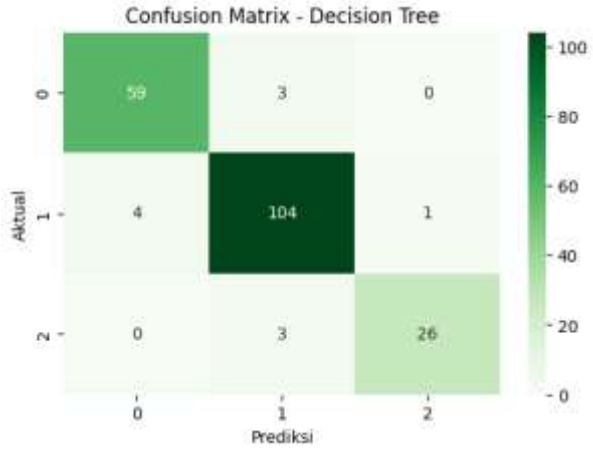

Figure 5. Confusion Matriks Decision Tree

The image above shows the Confusion matrix from the Decision Tree model which is used to predict student academic performance in three categories, namely low (0), medium (1), and high (2). From the total test data, the model managed to correctly predict 59 students in the low category, 104 students in the medium category, and 26 students in the high category. However, there are several prediction errors, including 3 students in the low category predicted as moderate, 4 students in the medium category predicted as low, 1 student in the medium category predicted as high, and 3 students in the high category predicted as moderate. Overall, the model performed quite well with most of the data being classified correctly, although there were still minor errors compared to the Random Forest model. This suggests that Decision Tree is able to recognize data patterns well, but tends to be slightly more susceptible to misclassification of data that has similar characteristics between classes.

d. Accuracy Comparison

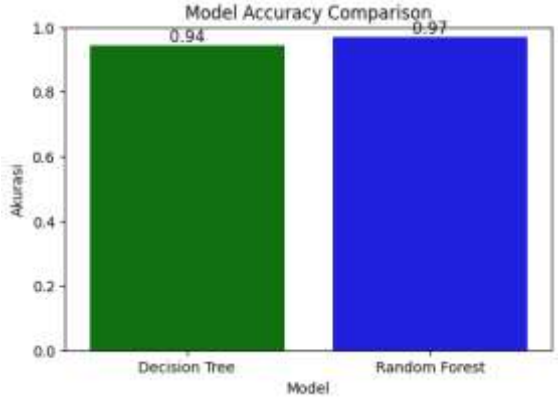

Figure 6. Accuracy Comparison

The figure above shows an accuracy comparison between two classification models, namely Decision Tree and Random Forest, which were used to predict students' academic performance based on learning activities. The results revealed that the Random Forest model achieved higher accuracy at 0.97 (97%), while the Decision Tree model obtained an accuracy of 0.94 (94%). This difference indicates that the ensemble learning approach in Random Forest effectively enhances prediction performance by combining multiple decision trees to minimize misclassification errors and prevent overfitting. Although Decision Tree demonstrates fairly good accuracy, it remains more sensitive to data variability and tends to have weaker generalization when faced with diverse datasets.

In a broader educational context, these findings highlight the potential of machine learning models—particularly Random Forest—to serve as data-driven tools for academic monitoring and decision-making. Educational institutions can utilize such predictive systems to identify students who may be at risk of academic decline, allowing for early interventions through mentoring, tutoring, or adaptive learning programs. Moreover, the insights gained from feature importance analysis—which emphasize cognitive abilities such as reading, writing, and mathematical skills—can guide

DOI : https://doi.org/10.52158/rjti.v4i3.4032

educators in designing more focused learning strategies that strengthen these core competencies. As a recommendation, future research should explore the integration of predictive analytics into real-time academic support systems, enabling continuous monitoring and dynamic feedback loops between students and educators. This integration can foster more personalized learning experiences and improve institutional effectiveness in promoting academic success and reducing student failure rates.

4. Conclusion

The results of this study demonstrate that the application of Random Forest and Decision Tree algorithms can effectively predict students' academic performance based on learning activities. Among the two models, Random Forest achieved the highest accuracy of 97%, surpassing Decision Tree, which achieved 94%. The evaluation using the confusion matrix and classification report revealed that Random Forest performed more consistently across all performance categories—low, medium, and high—showing higher values of precision, recall, and F1-score. The feature importance analysis further identified reading score, writing score, and math score as the most influential predictors of academic achievement, while demographic variables such as gender and parental education contributed less significantly. Beyond statistical performance, these findings have practical implications for the field of higher education. The use of machine learning models, particularly Random Forest, can assist educational institutions in developing data-driven decision support systems that help identify students at risk of academic decline early on. By integrating predictive analytics into academic monitoring systems, universities can design targeted interventions, personalized learning strategies, and academic support programs to improve student success rates. Moreover, this research opens opportunities for future studies to explore the implementation of real-time, data-integrated academic monitoring systems, enabling institutions to continuously track, analyze, and enhance learning outcomes. In this way, the study not only contributes to the theoretical understanding of predictive modeling in education but also provides a tangible framework for improving the effectiveness, equity, and quality of academic management in higher education.

Reference List

[1] N. Sharma, S. Appukutti, U. Garg, J. Mukherjee, and S. Mishra, "Analysis of Student's Academic Performance based on their Time Spent on Extra-Curricular Activities using Machine Learning Techniques," International Journal of Modern Education and Computer Science, vol. 15, no. 1, 2023, doi: 10.5815/ijmecs.2023.01.04.

[2] M. Furqon, P. Sinaga, L. Liliasari, and L. S. Riza, "The Impact of Learning Management System (LMS) Usage on Students," TEM Journal, vol. 12, no. 2, pp. 1082–1089, May 2023, doi: 10.18421/TEM122-54.

[3] M. P. R. I. R. Silva, R. A. H. M. Rupasingha, and B. T. G. S. Kumara, "A Comparative Study of Predicting Students' Academic Performance Using Classification Algorithms," in ICARC 2022 - 2nd International Conference on Advanced Research in Computing: Towards a Digitally Empowered Society, 2022. doi: 10.1109/ICARC54489.2022.9753729.

[4] R. Andriani Saputri and L. Rosnita, "A Random Forest-Based Predictive Model for Student Academic Performance: A Case Study in Indonesian Public High Schools," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[5] M. Ghofar Rohman, Z. Abdullah, and S. Kasim, "INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage : www.joiv.org/index.php/joiv INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION Hybrid Logistic Regression Random Forest on Predicting Student Performance." [Online]. Available: www.joiv.org/index.php/joiv

[6] B. Deddy Setiawan and D. Wibisono, "PREDICTING STUDENT PERFORMANCE USING MACHINE LEARNING FOR STUDENT MANAGEMENT IN UNIVERSITY," Jurnal Ilmiah Indonesia, vol. 7, no. 10, 2022, doi: 10.36418/syntax.

[7] M. A. Gotardo, "Using Decision Tree Algorithm to Predict Student Performance," Indian J Sci Technol, vol. 12, no. 8, pp. 1–8, Feb. 2019, doi: 10.17485/ijst/2019/v12i5/140987.

[8] M. Hasan et al., "Predicting student performance and identifying learning behaviors using decision trees and K-means clustering," International Journal of Evaluation and Research in Education, vol. 14, no. 5, pp. 3872–3881, Oct. 2025, doi: 10.11591/ijere.v14i5.33815.

[9] N. Sulistiyaningsih and R. Rismayati, "Comparison of Random Forest, Decision Tree, and XGBoost Models in Predicting Student Academic Success," Journal of Artificial Intelligence and Software Engineering, vol. 5, no. 3, pp. 920–930, 2025, doi: 10.30811/jaise.v5i3.7138.

[10] K. Kesgin, S. Kiraz, S. Kosunalp, and B. Stoycheva, "Beyond Performance: Explaining and Ensuring Fairness in Student Academic Performance Prediction with Machine Learning," Applied Sciences (Switzerland), vol. 15, no. 15, Aug. 2025, doi: 10.3390/app15158409.

[11] A. I. Gufroni, P. Purwanto, and F. Farikhin, "INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION journal homepage : www.joiv.org/index.php/joiv INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION Academic Performance Prediction Using Supervised Learning Algorithms in University Admission." [Online]. Available: www.joiv.org/index.php/joiv

[12] W. Chong-Wen, L. Sha-Sha, and E. Xu, "Predictors of rapid eye movement sleep behavior disorder in patients with Parkinson's disease based on Random Forest and decision tree," PLoS One, vol. 17, no. 6 6, 2022, doi: 10.1371/journal.pone.0269392.

[13] X. Wang, L. Zhang, and T. He, "Learning Performance Prediction-Based Personalized Feedback in Online Learning via Machine Learning," Sustainability (Switzerland), vol. 14, no. 13, 2022, doi: 10.3390/su14137654.

[14] M. S. Ibrahim Alsumaidaie, K. M. Ali Alheeti, and A. K. Alaloosy, "An Assessment of Ensemble Voting Approaches, Random Forest, and Decision Tree Techniques in Detecting Distributed Denial of Service (DDoS) Attacks," Iraqi Journal for Electrical and Electronic Engineering, vol. 20, no. 1, 2024, doi: 10.37917/ijeee.20.1.2.

[15] M. Schonlau and R. Y. Zou, "The Random Forest algorithm for statistical learning," Stata Journal, vol. 20, no. 1, 2020, doi: 10.1177/1536867X20909688.

[16] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," 2021. doi: 10.3389/fenrg.2021.652801.