Online version at https://journal.lenterailmu.com/index.php/jafotik

JAFOTIK

E-ISSN: 3031-2698, DOI prefix: 10.70356

Predicting Timely Graduation of Students with Disabilities at Universitas Pamulang Using Decision Tree and Random Forest Based on Historical Academic Data

Muhammad Azis Sularso[1]*, Riky Susanto[1], Juri Pebrianto[1], Ilham Firmansyah Siddhi[1]

[1]Informatics Engineering, Faculty of Computer Science, Universitas Pamulang, South Tangerang, Indonesia

ARTICLE INFO

Article history:
Received 01 August 2026
Revised 02 September 2026
Accepted 12 September 2026
Published 16 September 2026

Keywords:
Students with Disabilities
Timely Graduation
Machine Learning
Decision Tree
Random Forest

ABSTRACT

Timely graduation is an important indicator of academic success, particularly for students with disabilities who may require additional academic support. This study aims to predict timely graduation among students with disabilities at Universitas Pamulang using historical academic data and to compare the performance of Decision Tree and Random Forest algorithms. The study employed a quantitative approach and data mining methods using data from 125 students, with 100 records used for training and 25 records for testing. The prediction models were developed using academic variables, including semester GPA, cumulative GPA, total credits earned, study duration, attendance, and disability-related variables. Model performance was evaluated using accuracy, precision, recall, F1-score, confusion matrix, and 5-fold cross-validation. The results showed that Random Forest outperformed Decision Tree in terms of testing accuracy, achieving 88.00% compared with 80.00%. Random Forest also achieved a 5-fold cross-validation accuracy of 97.00% ± 0.0245. Feature importance analysis indicated that IPS_S4, IPS_S3, and AVG_IPS were among the most influential variables. These findings suggest that Random Forest can serve as a supporting tool for the early identification of students who may be at risk of delayed graduation and can help inform data-driven academic interventions.

This is an open access article under the CC BY-SA license.

1. Introduction

Inclusive higher education aims to provide equal learning opportunities for all students, including students with disabilities. In its implementation, students with disabilities may face various challenges during the academic process, such as accessibility limitations, suboptimal academic support, and the need to adapt to the learning environment. These conditions may affect the learning process and potentially influence students' ability to complete their studies on time. Universitas Pamulang is one of the higher education institutions that applies the principles of inclusive

* *Corresponding author:* Muhammad Azis Sularso
E-mail address: dosen03153@unpam.ac.id

education and provides learning opportunities for students with disabilities. Therefore, the use of historical academic data is important for monitoring students and supporting more targeted academic services. Data such as Semester Grade Point Average (IPS), Cumulative Grade Point Average (IPK), the number of credits (SKS), attendance, study duration, and disability type can provide an overview of students' academic conditions and factors associated with successful study completion.

The main problem is that monitoring students who may experience delayed graduation has not been fully supported by systematic data analysis. Academic evaluation is generally conducted by observing grade development each semester or through reports from academic advisors. This approach can make it difficult for the university to identify students who are at risk of not graduating on time at an early stage. Therefore, a data-driven approach is needed to utilize patterns in historical data to support prediction and academic decision-making. Machine learning can provide an appropriate approach because it is capable of performing classification and prediction based on patterns contained in historical data. In this study, academic data are used to develop a model that predicts graduation status and can support earlier academic intervention for students who are at risk of delayed graduation [1].

This study employs two classification algorithms, namely Decision Tree (DT) and Random Forest (RF) [2]. Decision Tree is selected because its decision structure is relatively easy to interpret, while Random Forest is an ensemble learning method that combines multiple decision trees to produce more stable predictions and reduce the risk of overfitting. Both algorithms are used to learn the relationship between students' academic characteristics and graduation status. The objectives of this study are to identify factors associated with the graduation of students with disabilities, develop prediction models using Decision Tree and Random Forest based on historical academic data, and evaluate the performance of both models using accuracy, precision, recall, and F1-score metrics [3].

The study is expected to contribute to the development of a more data-driven academic monitoring system at Universitas Pamulang. The resulting prediction model can serve as a basis for identifying students at risk of delayed graduation, enabling the university to provide academic assistance or intervention more quickly and appropriately. In addition, the study is expected to support the development of higher education that is more inclusive, adaptive, and data-driven [4].

2. Method

This study employed a quantitative approach combined with data mining and machine learning methods to develop a prediction model for the timely graduation of students with disabilities at Universitas Pamulang. The quantitative approach was used to analyze historical academic data numerically and statistically, while data mining was applied to identify patterns and relationships within the academic data. The resulting patterns were then used to develop a classification model for predicting students' graduation status [5].

2.1. Data Population and Sample

The population of this study consisted of active students with disabilities at Universitas Pamulang from various study programs. The population included students with different types of disabilities, such as physical, hearing, visual, and other mild disabilities, who were officially registered as active students in the university's academic system [6]. The sample was selected using purposive sampling based on criteria relevant to the research objectives. The criteria included students with disabilities who were still active at Universitas Pamulang, had completed at least one semester of study, and had complete academic records covering academic performance, credits, attendance, and study duration [7].

The final dataset used in the analysis consisted of 125 student records, which were divided into training and testing data. The dataset was used to build and evaluate the classification models for predicting whether a student would graduate on time.

2.2. Variables and Data Preparation

The prediction target in this study was Lulus_Tepat_Waktu, representing whether a student graduated on time. The predictor variables consisted primarily of academic and student characteristics, including Semester Grade Point Average (IPS) from the first to fourth semesters, average IPS, disability type, accumulated credits, study duration, attendance, and academic performance trends [8].

Before model development, the data underwent several preprocessing stages. These stages included data cleaning, normalization where required, categorical data encoding, and feature preparation. The prepared dataset was subsequently divided into training and testing subsets using an 80:20 ratio. Feature selection was also performed to determine the variables that contributed to the prediction process. The preprocessing procedure was intended to ensure that the dataset was suitable for classification using the selected machine learning algorithms.

2.3. Classification Algorithms

Two classification algorithms were applied in this study: Decision Tree (DT) and Random Forest (RF). The two algorithms were selected because they are suitable for classification problems and provide useful interpretability in the context of educational data. The Decision Tree algorithm constructs a tree-based decision structure using attributes in the dataset. Each branch represents a decision based on an attribute, while each leaf represents a classification result. In this study, Decision Tree was used to learn the relationship between academic characteristics, such as academic performance, attendance, credits, study duration, and disability type, and the student's graduation status [9]. Random Forest was used as an ensemble learning method that combines multiple Decision Trees. Each tree produces a classification result, and the final prediction is determined through majority voting. This approach is intended to improve model stability and reduce the risk of overfitting associated with a single decision tree [10].

2.4. Model Evaluation

The performance of the two classification models was evaluated using a confusion matrix and several classification metrics, namely accuracy, precision, recall, and F1-score. These metrics were used

to measure the ability of each model to correctly classify students according to their graduation status. In addition to the train-test evaluation, 5-fold cross-validation was performed to assess the consistency of model performance across different data partitions. The model comparison was conducted by examining the performance of Decision Tree and Random Forest on the testing data as well as their cross-validation results [11]. The algorithm with the better overall performance was considered the more suitable model for predicting the timely graduation of students with disabilities at Universitas Pamulang.

2.5. Research Workflow

The overall research process consisted of data collection, data preprocessing, feature preparation, dataset splitting, model development using Decision Tree and Random Forest, model evaluation, and comparison of the resulting performance [12]–[15]. The workflow was designed to transform historical academic data into a predictive model that could support early identification of students who may be at risk of delayed graduation (Figure 1).

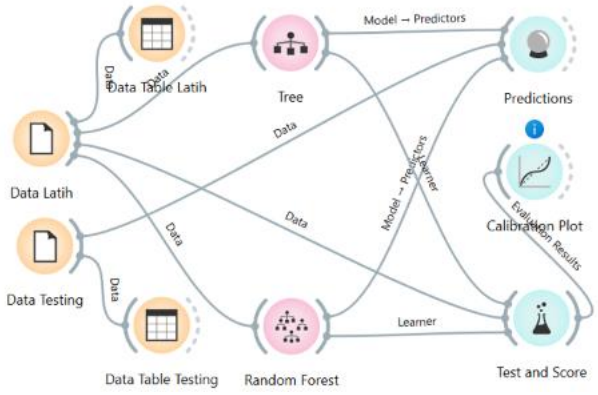

Fig. 1 - Research workflow for predicting the timely graduation of students with disabilities using Decision Tree and Random Forest.

3. Result and Discussion

The analysis was conducted using historical academic data of students with disabilities at Universitas Pamulang. The dataset consisted of 125 student records, which were divided into 100 training records (80%) and 25 testing records (20%). The target variable was *Lulus_Tepat_Waktu*, while the predictor variables included IPS_S1, IPS_S2, IPS_S3, IPS_S4, AVG_IPS, disability type, total credits, and study duration. Before the prediction process, the dataset was cleaned to address missing values, duplicate records, and data inconsistencies.

The analysis compared two classification algorithms, namely Decision Tree (DT) and Random Forest (RF). The Orange data mining application was used for the initial analysis, while a Python-based prediction system was used to validate and compare the results. The same historical dataset was processed using the two classification models, and their performance was evaluated using confusion matrix, 5-fold cross-validation, accuracy, precision, recall, F1-score, and feature importance.

3.1. Prediction Analysis Using Orange

The first stage of the analysis was performed using the Orange data mining application. The training data were connected to the Decision Tree and Random Forest models, and the resulting predictions were evaluated using the testing data. The workflow was designed to obtain prediction results and evaluate the performance of the two classification models.

The calibration plots were also generated to visualize the relationship between the predicted probabilities and the corresponding calibration curves for the two target classes. The first calibration plot represents the *Yes* target, while the second represents the *No* target.

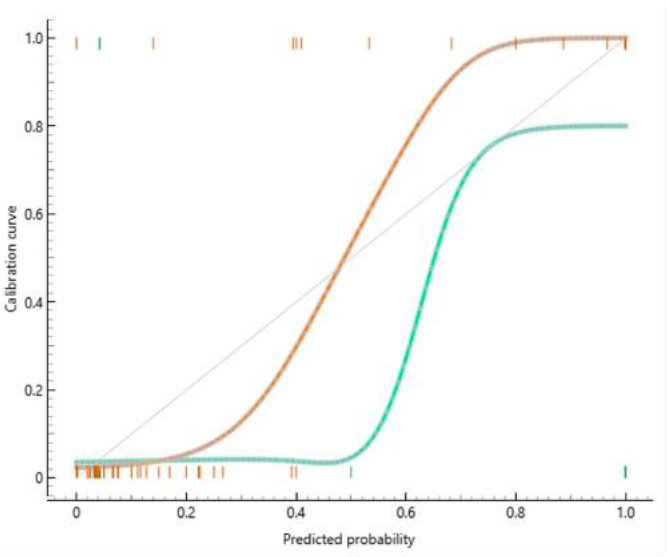

Fig. 2 - Calibration plot for the "Yes" target.

Figure 2 presents the calibration plot for the *Yes* target, representing students classified as graduating on time. The horizontal axis represents the predicted probability, while the vertical axis represents the calibration curve. The figure provides a visual representation of the predicted probabilities associated with the *Yes* classification.

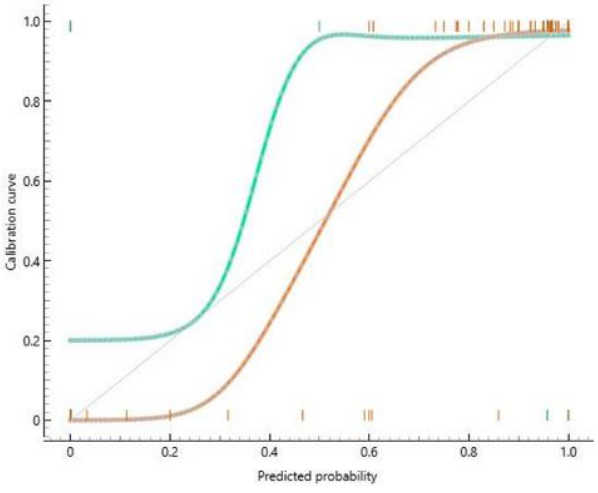

Fig. 3 - Calibration plot for the "No" target.

Figure 3 presents the calibration plot for the *No* target, representing students classified as not graduating on time. Similar

to Figure 2, the horizontal axis represents the predicted probability and the vertical axis represents the calibration curve. The figure provides a visual representation of the prediction probabilities associated with the *No* classification.

The calibration plots provide additional visualization of the prediction behavior of the models. However, the primary evaluation of model performance in this study was based on the confusion matrix, cross-validation, and classification metrics.

3.2. Prediction Analysis Using Python

In addition to the Orange analysis, a prediction system was developed using Python. The Python implementation was used to validate and compare the results obtained from the Orange application. The system included several analysis options, including 5-fold cross-validation, class weighting, and feature importance. According to the research report, the Python-based analysis showed that Random Forest performed better than Decision Tree.

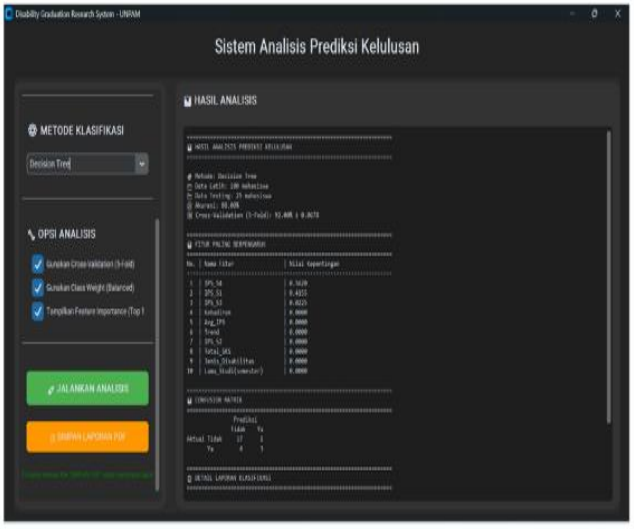

Fig. 4 - Prediction analysis using Python with the Decision Tree model.

Figure 4 presents the prediction analysis using the Decision Tree model in the Python-based system. The interface displays the selected classification method, analysis options, and the resulting model analysis. The system also provides information related to cross-validation, feature importance, confusion matrix, and classification results. These outputs were used to evaluate the performance of the Decision Tree model.

The Decision Tree model achieved an accuracy of 80.00% using 100 training records and 25 testing records. Its 5-fold cross-validation result was 92.00% ± 0.0678. The confusion matrix showed 17 correctly classified cases in the *No* class and 3 correctly classified cases in the *Yes* class. The model also produced one false positive and four false negatives.

Figure 5 presents the prediction analysis using the Random Forest model in the Python-based system. Similar to the Decision Tree analysis, the system provides information about the selected classification method, cross-validation, feature importance, confusion matrix, and classification results. The output provides a basis for comparing the Random Forest model with the Decision Tree model.

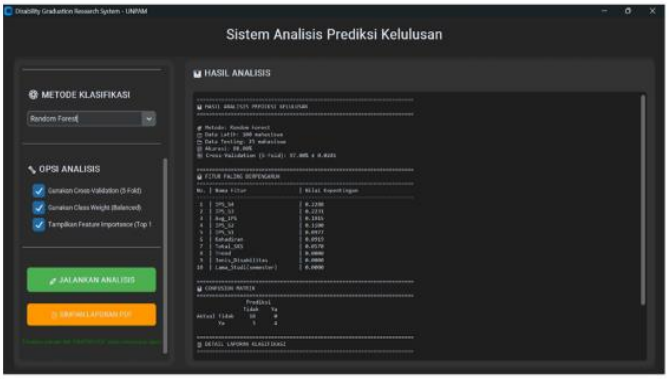

Fig. 5 - Prediction analysis using Python with the Random Forest model.

The Random Forest model achieved an accuracy of 88.00% using 100 training records and 25 testing records. The 5-fold cross-validation result was 97.00% ± 0.0245. Its confusion matrix contained 18 correctly classified cases in the *No* class and 4 correctly classified cases in the *Yes* class. There were no false positives and three false negatives.

3.3. Confusion Matrix and Cross-Validation Results

The confusion matrix results demonstrate differences between the two classification models. For Decision Tree, the model correctly classified 17 students who were in the *No* category and 3 students who were in the *Yes* category. One student was incorrectly classified as *Yes*, while four students in the *Yes* category were incorrectly classified as *No*. The resulting cross-validation performance was 92.00% ± 0.0678. For Random Forest, 18 students were correctly classified in the *No* category and 4 students were correctly classified in the *Yes* category. The model produced no false positives and three false negatives. The 5-fold cross-validation result of 97.00% ± 0.0245 was higher than the result obtained by Decision Tree.

The comparison indicates that Random Forest provided better performance on both the testing accuracy and cross-validation results. The accuracy increased from 80.00% for Decision Tree to 88.00% for Random Forest, while the cross-validation result increased from 92.00% ± 0.0678 to 97.00% ± 0.0245. These results indicate that Random Forest produced a more stable classification performance on the dataset used in this study.

3.4. Classification Performance

The classification report provides a more detailed comparison of precision, recall, and F1-score for the two models. For Decision Tree, the *No* class obtained a precision of 0.81, recall of 0.94, and F1-score of 0.87. For the *Yes* class, the model obtained a precision of 0.75, recall of 0.43, and F1-score of 0.55. The overall accuracy was 0.80.

For Random Forest, the *No* class obtained a precision of 0.86, recall of 1.00, and F1-score of 0.92. For the *Yes* class, the precision was 1.00, recall was 0.57, and F1-score was 0.73. The overall accuracy was 0.88, while the weighted F1-score was 0.87. The results show that Random Forest performed better than Decision Tree in identifying the *Yes* class. The recall for the *Yes* class increased from 0.43 in Decision Tree to 0.57 in Random Forest, while the F1-score increased from 0.55 to 0.73. The

precision for the *Yes* class also increased from 0.75 to 1.00. This indicates that Random Forest provided a better balance between precision and recall for the classification of students predicted to graduate on time.

3.5. Feature Importance Analysis

Feature importance analysis was conducted to determine the contribution of each predictor to the classification process. In the Decision Tree model, IPS_S4 was the most important feature with an importance value of 0.5620, followed by IPS_S1 with 0.4155 and IPS_S3 with 0.0225. The reported importance values for attendance, AVG_IPS, trend, IPS_S2, total credits, disability type, and study duration were 0.0000 in the Decision Tree analysis. For the Random Forest model, IPS_S4 had the highest importance value of 0.2288, followed by IPS_S3 with 0.2231, AVG_IPS with 0.1915, IPS_S2 with 0.1100, and IPS_S1 with 0.0977. Attendance contributed 0.0919, while total credits contributed 0.0570. The reported importance value for trend was 0.0000.

The feature importance results indicate that academic performance variables were important in the prediction process. In particular, IPS_S4 was the most influential feature in both models. In the Random Forest model, the importance was distributed more broadly across several academic variables, including IPS_S3, AVG_IPS, IPS_S2, and IPS_S1. This result suggests that the students' academic performance across semesters provided substantial information for predicting timely graduation.

3.6. Comparison of Decision Tree and Random Forest

The overall comparison demonstrates that Random Forest outperformed Decision Tree on the dataset used in this study. Random Forest achieved 88.00% accuracy, whereas Decision Tree achieved 80.00%. Random Forest also produced a higher 5-fold cross-validation result of 97.00% ± 0.0245, compared with 92.00% ± 0.0678 for Decision Tree.

The better performance of Random Forest can be considered in relation to its ensemble approach. The algorithm combines multiple decision trees and determines the final classification through majority voting. This approach is intended to provide greater model stability and reduce the risk of overfitting compared with a single Decision Tree. The research report also identifies Random Forest as the more stable and accurate model for the prediction of graduation among students with disabilities.

The findings demonstrate that historical academic data can be utilized to develop a model for predicting students' timely graduation. The Random Forest model provided the highest performance among the two algorithms evaluated in this study. Therefore, the model has potential as a supporting tool for identifying students who may require earlier academic attention or intervention.

However, the results should be interpreted within the scope of the available dataset. The study used only 125 student records, consisting of 100 training records and 25 testing records. Therefore, further research using a larger and more diverse dataset is recommended to strengthen the reliability and generalizability of the prediction model.

4. Conclusion

This study demonstrates that machine learning can be applied to predict the timely graduation of students with disabilities using historical academic data. The comparison between Decision Tree and Random Forest showed that Random Forest achieved better overall performance, with a testing accuracy of 88.00% compared with 80.00% for Decision Tree and a 5-fold cross-validation accuracy of 97.00% ± 0.0245. Random Forest also achieved a precision of 1.00, recall of 0.57, and F1-score of 0.73 for students predicted to graduate on time. Feature importance analysis indicated that IPS_S4, IPS_S3, and AVG_IPS were among the most influential variables in the Random Forest model. Therefore, Random Forest can be considered a useful approach for supporting early identification of students who may require academic attention. However, further research is recommended using larger and more diverse datasets, stronger validation methods, and other machine learning or ensemble models to improve prediction reliability and support the development of a web-based academic monitoring application.

Acknowledgements

The authors would like to express their sincere gratitude to Universitas Pamulang for providing permission and support for the implementation of this research. The authors also gratefully acknowledge the Faculty of Computer Science, the Informatics Engineering Program, the Research and Community Service Institute (LPPM) of Universitas Pamulang, fellow lecturers, and students who contributed to the research process. Special appreciation is also extended to the students with disabilities of Universitas Pamulang for their cooperation and contribution in providing the information required for this study.

REFERENCES

[1] T. A. Marzuqi, E. Kristiani, and Marcel, "Prediksi Mahasiswa Drop-Out Di Universitas XYZ," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 6, 2024.

[2] Jupron and Sutrisno, "Analysis of Heart Disease Using the Random Forest Method," *J. Inotera*, vol. 10, no. 1, pp. 167–174, 2025, doi: https://doi.org/10.31572/inotera.Vol10.Iss1.2025.ID452.

[3] A. Y. Pratama, I. S. Maulana, F. K. Sari, S. D. Tiara, and I. Darmawan, "Prediksi Risiko Depresi pada Mahasiswa Menggunakan Algoritma Random Forest Berdasarkan Data Akademik dan Gaya Hidup," *JSITIK J. Sist. Inf. dan Teknol. Inf. Komput.*, vol. 4, no. 1, 2025, doi: https://doi.org/10.53624/jsitik.v4i1.696.

[4] Faradillah, M. F. Alie, and I. Saluza, "Perbandingan Model Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree dan Naive Bayes," *Decod. J. Pendidik. Teknol. Inf.*, vol. 5, no. 2, 2025, https://doi.org/10.51454/decode.v5i2.1230.

[5] K. Kurniawati, R. Kusumawati, and M. A. Yaqin, "Klasifikasi Data Mahasiswa Lampau Menggunakan Metode Decision Tree dan Support Vector Machine,"

ILKOMNIKA, vol. 6, no. 3, 2024.

[6] I. Nur Sofiyanto, A. Bagus Saputra, R. Fauzi Rakhmadianto, H. Nur Eka Febrianto, A. Malika Endiwan, and A. Arum Sari, "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbor (KNN) dengan Klasifikasi Biner," *Pros. Semin. Nas. Teknol. Inf. dan Bisnis*, 2025.

[7] Muflihatul Hasanah and Zaehol Fatah, "Penerapan Decision Tree Pada Klasifikasi Kelulusan Mahasiswa," *J. Mhs. Tek. Inform.*, vol. 4, no. 2, 2025, doi: https://doi.org/10.35473/jamastika.v4i2.4487.

[8] A. A. Sitanggang, M. Y. Lase, and S. P. Sipayung, "Prediksi Kelulusan Mahasiswa Menggunakan Logistic Regression dan Random Forest Berdasarkan Data Akademik," *J. Pendidik. Tambusai*, vol. 10, no. 1, 2026.

[9] M. Putra and Erwin Harahap, "Machine Learning pada Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Random Forest," *J. Ris. Mat.*, 2024, doi: https://doi.org/10.29313/jrm.v4i2.5102.

[10] U. Indahyanti, N. L. Azizah, and H. Setiawan, "Pendekatan Ensemble Learning Untuk Meningkatkan Akurasi Prediksi Kinerja Akademik Mahasiswa," *J. Sains dan Inform.*, vol. 8, no. 2, 2022.

[11] U. Suriani, "Penerapan Data Mining untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree C4.5," *Journalcisa*, vol. 3, no. 2, 2023.

[12] H. Putra, K. Nasution, and E. Rilvani, "PREDIKSI KELULUSAN MAHASISWA TEPAT WAKTU MENGGUNAKAN DECISION TREE: STUDI PERBANDINGAN ALGORITMA ID3 DAN C4.5," *J. Media Akad.*, vol. 3, no. 7, 2025.

[13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: https://doi.org/10.1023/A:1010933404324.

[14] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. 1993," *Machine Learning*, vol. 16, pp. 235–240, 1994, doi: https://doi.org/10.1007/BF00993309.

[15] M. M. Bal, "Performance Evaluation of the Machine Learning Algorithms Used in Inference Mechanism of a Medical Decision Support System," *The Scientific World Journal*, 2014, doi: https://doi.org/10.1155/2014/137896.

.