Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9947

Analysis of the Decision Tree (C4.5) and Random Forest Algorithms to Determine Student Eligibility for Final Project Assignments Based on Academic Requirements

Eisyaniah Desvazulinda*, Muhammad Iqba, Muhammad Syahputra Novelan

[1]Master of Information Technology, University Pembangunan Panca Budi, Medan, Indonesia
Email: [1]eisyaniahdesva96@gmail.com, [2]muhammadiqbal@dosen.pancabudi.ac.id, [3]putranovelan@dosen.pancabudi.ac.id
Correspondence Author Email: eisyaniahdesva96@gmail.com*
Submitted: **08/06/2026**; Accepted: **29/06/2026**; Published: **30/06/2026**

Abstract— Determining student eligibility for undertaking a final project is an important process in higher education, which is often still conducted manually and subjectively. This study aims to develop a classification model based on machine learning to determine the eligibility of students at Batam University using Decision Tree (C4.5) and Random Forest algorithms. The data used includes Grade Point Average (GPA), total completed credits (SKS), prerequisite course grades, and academic records. This research employs a quantitative approach with stages including data collection, data preprocessing, model development, and performance evaluation using accuracy, precision, and recall metrics. The results show that both algorithms are capable of classifying student eligibility effectively. The Decision Tree (C4.5) algorithm produces an interpretable model in the form of decision rules, while Random Forest demonstrates superior performance in terms of accuracy and prediction stability. The comparison indicates that Random Forest is more effective in handling complex data, whereas C4.5 provides better model transparency. In conclusion, the implementation of Decision Tree (C4.5) and Random Forest algorithms can serve as an effective solution to support objective and data-driven academic decision-making. The resulting model has the potential to be developed into a decision support system to improve the efficiency and quality of determining student eligibility for final project enrollment.

Keywords: Data Mining; Decision Tree (C4.5; Random Forest; Student Eligibility; Final Project; Machine Learning.

1. INTRODUCTION

The final project is a crucial component of higher education, serving as an indicator of a student's ability to integrate the knowledge and skills acquired during their studies. A student's success in completing a final project is not solely determined by their individual abilities but also by their academic preparedness, which includes their Grade Point Average (GPA), the number of Semester Credit Units (SKS) completed, and their completion of prerequisite courses. However, in practice, not all students who meet the administrative requirements are truly ready to undertake a final project, creating uncertainty in the process of determining academic eligibilit [1].

This issue poses a significant challenge for universities, particularly in objectively and accurately determining which students are eligible to undertake final assignments. In practice, the eligibility assessment process is still largely manual, based on subjective academic considerations, such as lecturer evaluations or administrative policies that are not always based on comprehensive data analysis. This situation results in inconsistent decision-making and the potential for bias, especially when the number of students to be evaluated is quite large. Furthermore, decisions not supported by systematic data analysis can result in inaccurate determinations of student readiness, resulting in students who are not actually ready to undertake final assignments, while eligible students experience delays [2] [3].

The impact of these issues is felt not only by students in the form of delays in their studies, but also by educational institutions, which must bear the burden of increased administrative burdens and reduced efficiency in academic management. In this context, decision-making based solely on intuition or experience is considered less effective than a data-driven approach, which emphasizes the use of factual data to improve decision quality. Various international studies have shown that the use of educational data mining and machine learning techniques can help educational institutions identify student academic patterns, predict performance, and support more accurate and consistent decision-making [4].

Furthermore, a data-driven approach also enables the development of decision support systems (DSS) that can process multiple academic variables simultaneously, resulting in more objective recommendations than manual methods. With such a system, universities can minimize subjectivity in assessments and increase efficiency in the student eligibility evaluation process. Therefore, a transformation from a conventional approach to a data- and technology-driven approach is necessary to ensure that the process of determining student eligibility for final assignments can be carried out more precisely, transparently, and measurably [5].

With the advancement of information technology, particularly in the fields of data mining and machine learning, opportunities have emerged to utilize student academic data as a basis for more objective, data-driven decision-making. Classification techniques in data mining allow for grouping student data based on specific attributes to generate more accurate predictions or decisions. In this context, algorithms such as Decision Tree (C4.5) and Random Forest have been widely used in research to predict student academic performance and graduation status [6][7].

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9947

The Decision Tree algorithm (C4.5) has the advantage of producing easy-to-understand models because it is rule-based, making it easier for non-technical users to interpret. This algorithm works by using the concept of information gain or gain ratio to determine the best attributes for building a decision tree. On the other hand, Random Forest, as an ensemble learning method, can improve prediction accuracy and stability by combining multiple decision trees and reducing the risk of overfitting. Various studies have shown that Random Forest tends to outperform a single Decision Tree in handling complex data [8].

At Batam University, the use of academic data to support decision-making regarding student eligibility for final assignments has not been optimal. Available data is still underutilized as a strategic source of information for predictive analysis. Therefore, a machine learning-based approach capable of systematically processing academic data to produce accurate and reliable predictive models is needed [9].

Based on these issues, this study was conducted to analyze and compare the performance of the Decision Tree (C4.5) and Random Forest algorithms in determining student eligibility for final assignments. The results are expected to contribute to the development of more objective, efficient, and data-driven academic decision support systems in higher education settings [10][11].

2. RESEARCH METHODOLOGY

This study uses a quantitative approach with data mining and machine learning methods to analyze students' eligibility for final assignments based on academic data. This approach was chosen because the study focuses on processing numerical and categorical data such as Grade Point Average (GPA), number of credits, prerequisite course grades, and students' academic history. This study is explanatory and predictive, meaning it not only explains the relationships between academic variables but also builds predictive models using the Decision Tree (C4.5) and Random Forest algorithms.[12]. In addition, this study also uses a comparative approach to compare the performance of the two algorithms in producing optimal classification models.

Figure 1. Research Flow Diagram

The data source in this study comes from the academic information system of Batam University which includes dataStudent study history. The data used consists of quantitative data such as GPA, number of credits, and course grades, as well as categorical data such as course completion status and student eligibility status for final assignments. The data is compiled into a structured dataset that is used as input in the classification process. In addition, this study also utilizes supporting data in the form of academic documentation related to the rules and criteria for final assignments [13].

Data collection techniques involved secondary data collection from academic information systems and a study of documentation regarding applicable academic regulations. Furthermore, the obtained data underwent a verification and data cleaning process to remove incomplete data, duplication, and inconsistencies, ensuring the dataset was of good quality and ready for analysis.

The research was conducted systematically, starting with problem identification and then developing a classification model using the Decision Tree (C4.5) algorithm and the Random Forest algorithm to produce a more stable and accurate model. Random Forest is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. Mathematically, the Random Forest prediction is formulated as follows:

$\hat{y}$=mode{ (),(),...,()}$\hbar_1 x \hbar_2 x \hbar_n x$ (1)

where () is the prediction of each i-th decision tree and the final result is determined based on majority voting. The next stage is the collection of relevant student academic data, followed by data pre-processing to clean and transform the data to suit the model's needs. After that, a classification model is formed using the Decision Tree algorithm (C4.5) to generate decision rules, as well as the Random Forest algorithm to produce a more stable and accurate model. Next, both models are evaluated and compared using performance metrics such as accuracy, precision, recall, and confusion matrix to determine the best model. The final stage is the preparation of conclusions and recommendations as the basis for developing an academic decision support system.$\hbar_i x$[14][15].

Data analysis in this study was conducted by dividing the dataset into training data and testing data. Models were built using the two predetermined algorithms, then tested to determine the performance of each model. The evaluation results were used to determine the most effective algorithm in accurately and consistently predicting student eligibility for final assignments [17],[18].

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9947

To support the research, the hardware used was a computer with an Intel Core i3 processor, 6 GB of RAM, and a 500 GB hard disk. The software used included the Windows 10 operating system, Microsoft Office 2019, and a dedicated software package.

3. RESULTS AND DISCUSSION

Based on the results of research conducted using academic data of Batam University students, a classification model for eligibility for taking the final assignment was successfully built by applying the Decision Tree (C4.5) and Random Forest algorithms. The data processing process begins with the pre-processing stage which includes data cleaning, attribute transformation, and the formation of a dataset consisting of main variables such as GPA, number of credits, prerequisite course grades, and student academic status. The dataset is then divided into training data and testing data to ensure the model can be evaluated objectively. The modeling results show that both algorithms are able to classify students into categories eligible and unfit to take the final assignment with a good level of performance.

In the Decision Tree model (C4.5), the results obtained are in the form of a decision tree structure that describes the relationship between academic attributes and student eligibility status. GPA and number of credits are the dominant factors in determining the branching of the decision tree, followed by prerequisite course grades. This model produces easy-to-understand decision rules (rule-based), for example, students with a GPA above a certain threshold and a sufficient number of credits tend to be classified as eligible. The main advantage of this model lies in its interpretability, so that academics can easily understand the factors that influence student eligibility in a transparent manner.

Meanwhile, the Random Forest model demonstrated superior performance compared to C4.5 in terms of prediction accuracy and stability. By building multiple decision trees and combining their results, Random Forest was able to reduce the risk of overfitting that typically occurs in single-tree models. Evaluation results using metrics such as accuracy, precision, and recall showed that Random Forest had a higher level of consistency in classifying student data, especially on datasets with complex attribute variations. This indicates that Random Forest is more effective in handling multidimensional academic data.[16].

A comparison of the two algorithms shows that while Random Forest performs better in terms of accuracy, Decision Tree (C4.5) still has an advantage in terms of model interpretability. In the context of academic decision-making, these two algorithms can complement each other, with Random Forest being used to produce more accurate predictions, while C4.5 is used to provide more understandable explanations of the classification results. Thus, the combination of the two can be a strong foundation for developing academic decision support systems.

Overall, the results of this study indicate that the application of machine learning algorithms, specifically Decision Tree (C4.5) and Random Forest, can provide a more objective and data-driven solution in determining student eligibility for final assignments. The resulting model not only improves the efficiency of the academic evaluation process but also helps the university reduce subjectivity in decision-making. Therefore, the implementation of this model has great potential for further development into an integrated decision support system within the academic information system of Batam University.

Table 1. Data Preprocessing

GPA	0
Failed Course	0
Number of Academic Leaves	0
Work While Studying	0
Number of Semesters	0
TA_Eligibility_Status	0
Average IPS	0
Final Semester Social Studies	0
IPS Trends	0
Attendance Category	0

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9947

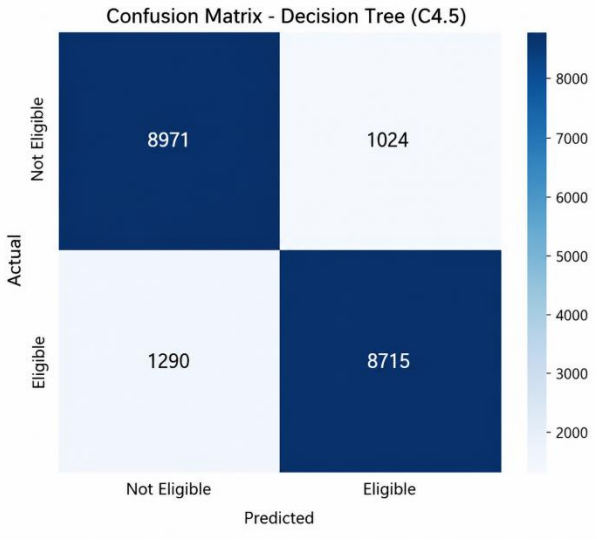

Figure 1. Confusion matrix - Decision Tree (C4.5)

The confusion matrix in the Decision Tree model (C4.5) shows that the model is able to classify the data quite well, with a True Negative value of 8971 and a True Positive value of 8715. However, there are still classification errors in the form of 1024 False Positives and 1290 False Negatives. This shows that although the model has high performance in recognizing both classes, there are still a number of errors, especially in the prediction of the positive class which is predicted as negative.

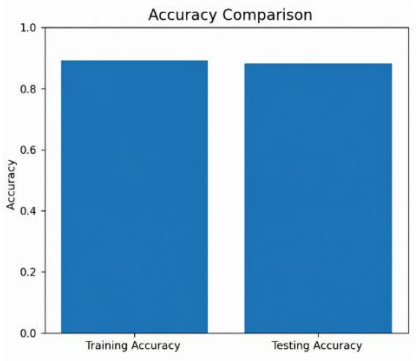

Figure 2. Accuracy Comparison

The accuracy comparison graph shows that the training and testing accuracy values are approximately the same, around 0.88–0.89. This indicates that the model has stable performance and is not experiencing overfitting, as the model's performance on the training and test data is not significantly different.

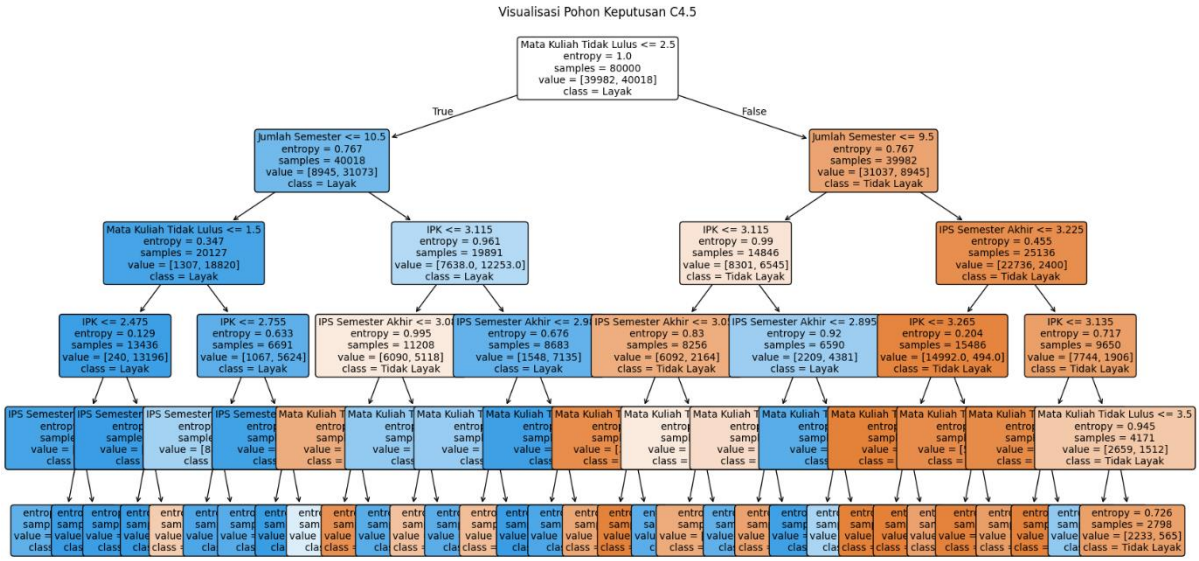

Figure 3. C45 Decision Tree Visualization

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9947

The image depicts a portion of a page containing a mathematical formula or analytical calculation, likely related to the calculation process within a specific model or method. The presence of mathematical symbols such as variables, exponents, and roots indicates that the image represents the steps of a mathematical calculation or formulation. Furthermore, there is a handwritten section at the bottom, indicating a manual calculation process or a detailed explanation of the formula. Overall, the image depicts the stages of mathematical calculations used to support analysis or modeling, whether in the context of statistics or machine learning.

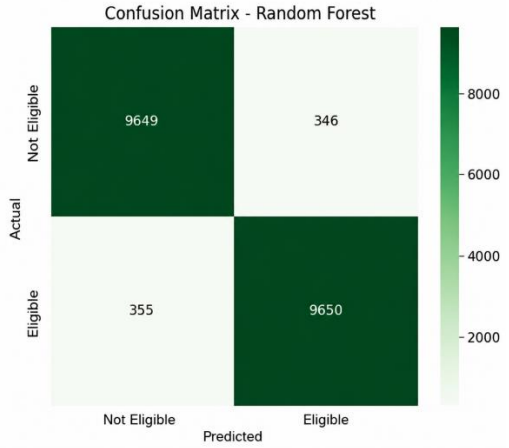

Figure 4. Confusion Matrix - Random Forest

The confusion matrix of the Random Forest model shows improved performance compared to the Decision Tree model. The True Negative count is very high at 9,649, while the False Positive count is only 346, demonstrating the model's excellent ability to recognize the negative class. This indicates that the Random Forest model is more accurate and stable in its classification than the previous model.

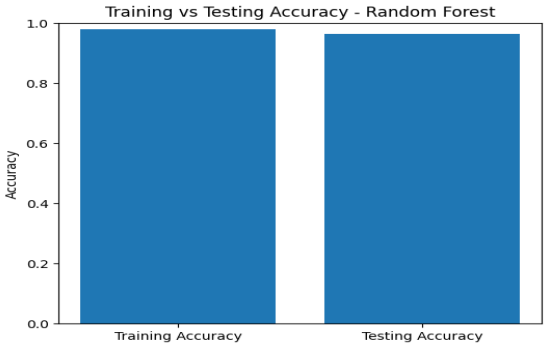

Figure 5. Accuracy Testing Training - Random Forest

The figure shows a comparison between the training and testing accuracy of the Random Forest model. The accuracy value on the training data is very high, approaching 1.0, while the accuracy on the testing data is slightly lower but still high. This indicates that the Random Forest model is able to learn data patterns very well on the training data and also has quite good generalization capabilities on the test data. The small difference between the training and testing accuracy indicates that the model does not experience significant overfitting. Thus, the model can be said to be stable and quite optimal in classifying.

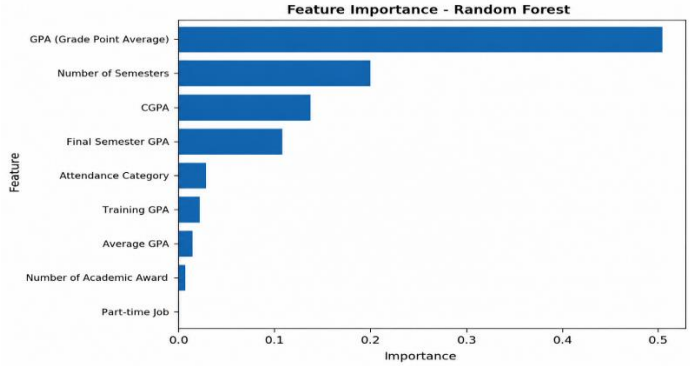

Figure 6. Feature Importance - Random Forest

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9947

The feature importance graph shows that the Final Semester Social Studies feature has the greatest influence on predicting the results. This is followed by features such as attendance category and Social Studies trend, while the number of academic leave features has the least influence. This indicates that recent academic performance is the primary factor in the classification.

Table 2. Feature Importance Random Forest

	Feature	Importance
1	Failed Course	0.502842
4	Number of Semesters	0.195845
0	GPA	0.129419
6	Final Semester Social Studies	0.109208
8	Attendance Category	0.027246
7	IPS Trends	0.024879
5	Average IPS	0.008730
2	Number of Academic Leaves	0.001164
3	Work While Studying	0.000667

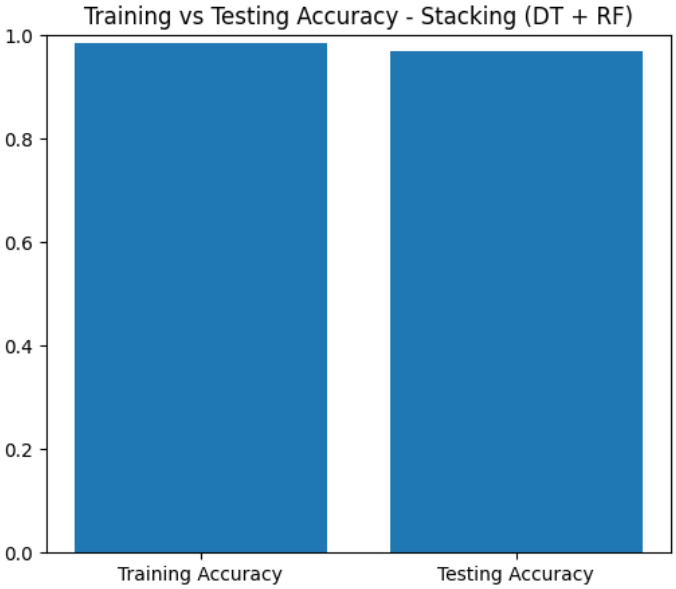

Figure 7. Training Vs Testing Accuracy – Staking (DT + RF)

The figure shows a comparison graph of the accuracy of the stacking model (Decision Tree/C4.5 + Random Forest) between the training and testing data. It can be seen that both accuracy values are at a high level and almost close to the maximum value, indicating that the model has excellent classification capabilities. Furthermore, the difference between training and testing accuracy is very small, indicating that the model does not experience overfitting and is able to generalize well to new data. This shows that the stacking method that combines two algorithms can improve the stability and performance of the model as a whole, resulting in more accurate and consistent predictions.

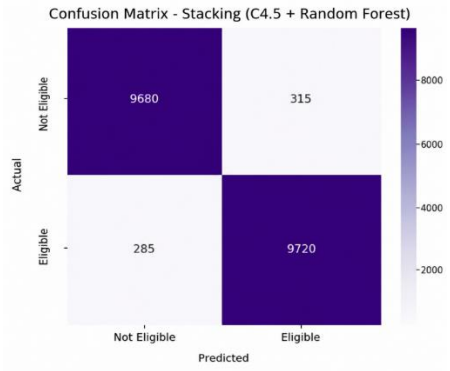

Figure 9. Confusion Matrix - stacking (C4.5 and Random Forest)

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9947

The confusion matrix in the figure shows the performance of the stacking model (a combination of C4.5 and Random Forest) in classifying two classes, namely Unfit and Feasible. It can be seen that 9680 Unfit data were successfully predicted correctly (True Negative), while only 315 data were incorrectly predicted as Feasible (False Positive). In the Feasible class, there were 9720 data that were correctly predicted (True Positive) and only 285 data that were incorrectly predicted as Unfit (False Negative). These results indicate that the stacking model has excellent performance with a very high number of correct predictions and relatively small errors. Overall, the model is able to distinguish the two classes very accurately and balanced, so it can be said to be the optimal model in this classification task.

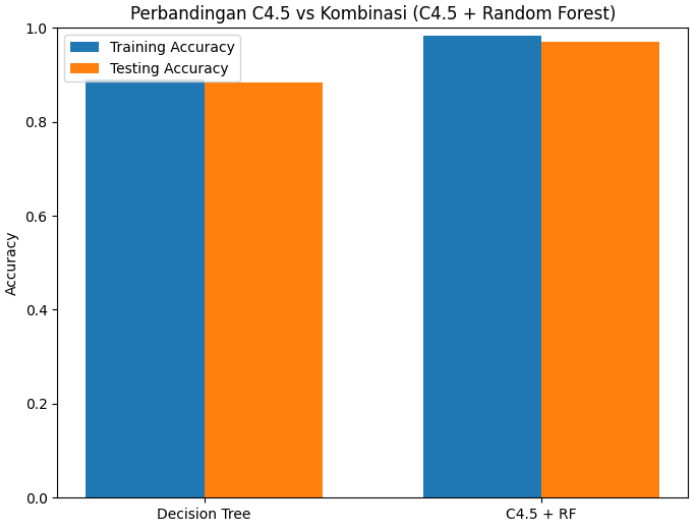

Figure 8. Comparison of C4.5 vs. Combination (C4.5 + Random Forest)

The figure shows a comparison of accuracy between the Decision Tree model (C4.5) and the combination model (C4.5 + Random Forest) on training and testing data. It can be seen that the Decision Tree model has a fairly high accuracy value, but still below the combination model. Meanwhile, the combination model (C4.5 + Random Forest) shows improved performance with a training accuracy value approaching 1.0 and a testing accuracy that is also higher than the Decision Tree. In addition, the difference between the training and testing accuracy of the two models is not too large, indicating that neither model experiences significant overfitting. However, the combination model remains superior because it is able to provide higher accuracy and more stable results, so it can be concluded that the use of the ensemble method can improve model performance in the classification process.

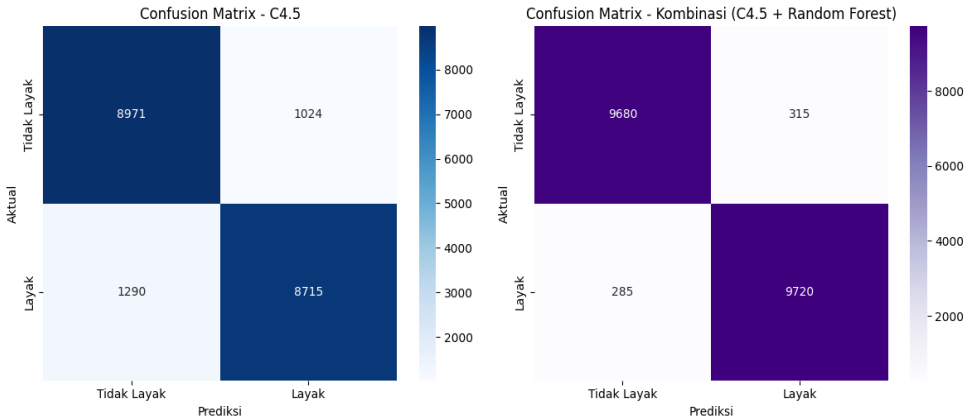

Figure 9. Comparison of the Confusion For Matrix - C4.5 and Random Forest

The figure shows a comparison of the confusion matrix between the C4.5 model and the combination model (C4.5 + Random Forest) in classifying two classes, namely Unfit and Feasible. In the C4.5 model, it can be seen that there are 8971 Unfit data that are correctly predicted (True Negative) and 8715 Eligible data that are correctly predicted (True Positive), but there are still quite large errors, namely 1024 False Positive and 1290 False Negative. Meanwhile, in the combination model (C4.5 + Random Forest), the number of correct predictions increased significantly with 9680 True Negative and 9720 True Positive, as well as much smaller errors, namely only 315 False Positive and 285 False Negative. This shows that the combination model has a much better performance than the C4.5 model, both in reducing prediction errors and in increasing overall

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9947

accuracy. Thus, the use of the ensemble method through a combination of C4.5 and Random Forest has been proven to be able to improve the model's ability to distinguish the two classes more accurately and balanced.

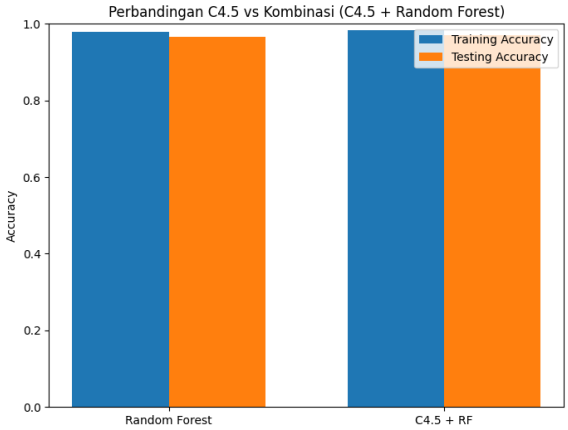

Figure 10. Comparison of C4.5 vs. Combination (C4.5 + Random Forest)

The figure shows a comparison of the accuracy between the Random Forest model and the combined model (C4.5 + Random Forest) on the training and testing data. It can be seen that both models have very high accuracy values, both on the training and testing data, indicating that both are capable of good classification. However, the combined model (C4.5 + Random Forest) shows slightly superior performance compared to pure Random Forest, both in terms of training accuracy and testing accuracy. Furthermore, the difference between the training and testing accuracy of the two models is relatively small, indicating that the models do not experience significant overfitting and have good generalization capabilities.

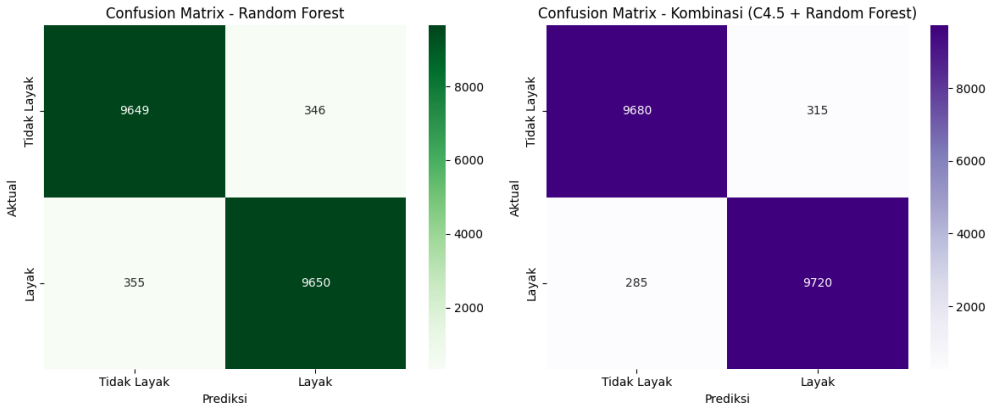

Figure 11. Comparison of the Confusion For Random Forest and Combination (C4.5 + Random Forest)

The figure shows a comparison of the confusion matrix between the Random Forest model and the combination model (C4.5 + Random Forest) in classifying two classes, namely Unfit and Feasible. In the Random Forest model, it can be seen that 9649 Unfit data were predicted correctly (True Negative) and 9650 Feasible data were predicted correctly (True Positive), with errors of 346 False Positive and 355 False Negative. Meanwhile, in the combination model (C4.5 + Random Forest), the number of correct predictions increased to 9680 for Unfit and 9720 for Feasible, and the number of errors decreased to 315 False Positive and 285 False Negative. This shows that the combination model has better performance than Random Forest, because it is able to increase the number of correct predictions while reducing classification errors. Thus, the use of the ensemble method through a combination of algorithms provides more accurate and balanced results in distinguishing the two classes.

4. CONCLUSION

Based on the research results, it can be concluded that the implementation of the Decision Tree (C4.5) and Random Forest classification algorithms can be used effectively in determining the eligibility of Batam University students to take the final assignment based on academic data such as GPA, number of credits, prerequisite course grades, and study history. Both algorithms demonstrated good capabilities in identifying patterns of relationships between academic variables and producing predictive models that can be used as a basis for more objective decision-making.

The Decision Tree (C4.5) model offers the advantage of an easily understood and interpreted decision tree structure, making it easier for academics to identify the key factors influencing student eligibility. Meanwhile, Random Forest demonstrates superior performance in terms of prediction accuracy and stability, as it is able to reduce overfitting through the combination of multiple decision trees. Thus, the results of this study indicate that the combination of these two algorithms can provide a balance between model accuracy and interpretability. The implementation of this classification model has the potential to become the basis for developing a data-driven academic decision support system, thereby increasing efficiency, consistency, and objectivity in the process of determining student eligibility for final assignments at Batam University.

REFERENCES

[1] Ş. K. Aydoğan, T. Pura, and F. Bİngül, "Predicting Students' Academic Performances Using Machine Learning Algorithms in Educational Data Mining," vol. 12, no. 4, 2024.
[2] JM Zain, H. Chiroma, and T. Herawan, "Data Mining for Education Decision Support:," pp. 4–19.
[3] MD Alviansyah, "Application of Data Mining in Supporting Scholarship Recipient Decision Support Systems at Universities: Literature Review," vol. 4, no. 2, pp. 149–156, 2024.
[4] M. Gebser, B. Kaufmann, and T. Schaub, "Conflict-driven answer set solving : From theory to practice☆," Artif. Intell., vol. 187–188, pp. 52–89, 2012, doi: 10.1016/j.artint.2012.04.001.
[5] *No Title*.
[6] VN Juli, AS Gustian, F. Mahardika, and US April, "Analysis of Student Dropout Risk Classification Using Decision Tree and Random Forest Algorithms This research focuses on the development and evaluation of classification models for Random Forest. By using historical academic data, the developed model," 2025.
[7] "Tahta Media Group".
[8] MS Novelan, S. Aryza, U. Pembangunan, and P. Budi, "OPTIMIZATION CVRP WITH MACHINE LEARNING FOR IMPROVED CLASSIFICATION OF IMBALANCED DATA FOOD DISTRIBUTION," vol. 10, no. 4, p. 917–925, 2025, doi: 10.33480/jitk.v10i4.6467.OPTIMIZATION.
[9] M. Amin, C. Rizal, and IM R, "Instal : Jurnal Komputer," vol. 17, pp. 576–581, 2025.
[10] ND Rumlaklak, DR Sina, and TN Sooai, "Comparative Analysis of C4.5 and Random Forest for Analyzing Factors Affecting Undergraduate Students' Final Project Completion in Higher Education," p. 350–359, 2025.
[11] M. Indra and D. Nasution, "Analysis of Nursing Home Residents' Identity Completeness Classification Using the Decision Tree Algorithm," vol. 1, no. 2, p. 158–161, 2024, doi: 10.61306/jitcse.v1i2.
[12] S. Barutu, M. Iqbal, and D. Nasution, "Comparative Analysis of the C4.5 and Random Forest Algorithms for the Prediction of Diarrheal Disease," vol. 4, no. 07, p. 625–635, 2025, doi: 10.58471/esaprom.v4i07.
[13] M. Iqbal, BS Anggoro, and R. Rakhmawati, "DEVELOPING CONTEXTUAL-BASED STUDENT WORKSHEETS," vol. 3, no. 1, pp. 1–9, 2020.
[14] P. Algorithm and C. For, "Application of c4.5 algorithm for student graduation prediction based on academic data," vol. 3, no. 8, 2025.
[15] C. Algoritma, B. Mobile, and E. Hariyanto, "Analysis of Tourist Satisfaction in Binjai City Using," vol. 4, no. 1, pp. 89–98, 2023.
[16] MT Information, U. Development, and P. Budi, "Analysis of ASN Professionalism Index Improvement Patterns in Asahan Regency Using Random Forest and Gradient Boosting Machines," vol. 07, no. 02, pp. 141–149, 2025.
[17] A. Aman, N. P. Rahrahima, and A. Fitri, "Implementation of Machine Learning Algorithms for Predicting Student Academic Performance," IJATIS: Indonesian Journal of Applied Technology and Innovation Science, vol. 3, no. 1, pp. 1–9, Feb. 2026, doi: 10.57152/IJATIS.v3i1.1871.
[18] G. Gunawan, H. Hanes, and C. Catherine, "C4.5, K-Nearest Neighbor, Naïve Bayes, and Random Forest Algorithms Comparison to Predict Students' On Time Graduation," Indonesian Journal of Artificial Intelligence and Data Mining, vol. 4, no. 2, 2021, doi: 10.24014/ijaidm.v4i2.10833.