International Journal of Society Reviews (INJOSER) e-ISSN: 3030-802X
Vol. 3 No. 3, March 2025, pages: 59~69

COMPARISON OF DECISION TREE AND RANDOM FOREST ALGORITHMS IN PREDICTING STUDENT GRADUATION BASED ON ACADEMIC DATA

Marlan
Universitas Nasional, Indonesia

Ahmad Rifqi
Universitas Nasional, Indonesia

Agus Iskandar
Universitas Nasional, Indonesia
Correspondence author email: Iskandaragus1005@gmail.com

Abstract

This research aims to compare the performance of the Decision Tree and Random Forest algorithms in predicting student graduation based on academic data. By utilizing data such as Grade Point Average (GPA), the number of credit hours, and course grades, this study focuses on analyzing the accuracy of both algorithms in predicting students who are at risk of not graduating on time. The results of the study indicate that the Random Forest algorithm achieves higher accuracy compared to the Decision Tree, particularly in terms of recall and precision. While Decision Tree is simpler and easier to interpret, it tends to have overfitting issues that can affect prediction results. In contrast, Random Forest overcomes these issues by producing more stable predictions through an ensemble process. This study is expected to contribute to the development of student graduation prediction systems in educational institutions. As such, institutions can use these findings as a foundation for designing intervention strategies for students at risk of not graduating on time.

***Keywords**: Decision Tree, Random Forest, Graduation Prediction, Academic Data, GPA, Credit Hour.*

INTRODUCTION

Higher education plays a very important role in preparing competent graduates who are ready to face the workforce. However, one of the challenges often faced by higher education institutions is the untimely graduation rates of students. Many educational institutions face difficulties in ensuring that their students can complete their studies within the specified timeframe. Various factors influence students' timely graduation, ranging from academic performance to involvement in campus activities and students' ability to manage their time. Therefore, it is important for higher education institutions to be able to predict student graduation in order to assist those who are at risk of facing delays in completing their studies.

Predicting student graduation can be a very useful tool for educational institutions to take preventive measures in helping students who may be at risk of not graduating on time. By utilizing predictive methods, institutions can identify students who need

additional support, both academically and non-academically. This not only helps improve graduation rates but can also enhance the overall quality of education by giving special attention to at-risk students.

Algorithm and Graduation Prediction Challenge

Several studies have attempted to utilize various methods and algorithms to predict student graduation. For example, algorithms such as C4.5 and Naïve Bayes have been used in several studies to predict student graduation. Although the results of these studies are quite promising, there is significant variation in the accuracy levels achieved by these algorithms. The accuracy levels obtained range from 72% to 99.64%, indicating that there are still several challenges to be faced in producing more consistent and accurate predictions.

In the research conducted by Wahyudi (2023), it is explained that although algorithms such as C4.5 and Naïve Bayes have been widely used in previous studies, there are still opportunities to conduct further research to improve prediction accuracy. This research highlights the importance of identifying more specific and relevant factors that influence student graduation. Additionally, the use of larger and more diverse datasets can help produce more accurate predictions and provide deeper insights into the factors affecting timely graduation.

One of the methods often used in graduation prediction is the Decision Tree, which offers advantages in terms of interpretability. However, as revealed by several studies, this algorithm has limitations in terms of accuracy, especially when faced with complex data. As an alternative, Random Forest is often used because it can overcome some of the weaknesses of Decision Trees. Random Forest combines several decision trees built randomly, thereby reducing the risk of overfitting and improving prediction accuracy.

Factors Affecting Student Graduation

Many studies have attempted to identify the factors that influence student graduation. For example, Latifah (2020) found that students' academic performance varies greatly, and this variation can be caused by various factors that were previously undetected. One way to address this issue is by applying data mining methods to identify the appropriate predictive model. By using classification techniques such as Decision Tree C4.5 and Random Forest, Latifah attempts to predict students' academic performance based on various relevant parameters, such as performance in certain subjects, involvement in campus activities, and other factors that can influence students' academic outcomes.

Another study conducted by Darmawan et al. (2023) highlights the importance of non-academic factors in predicting student graduation. In this study, it was found that admission test scores are not always an accurate predictor of on-time graduation. On the contrary, academic grades from previous education, such as high school grades, can be a better predictor. This shows that students' academic success is not only

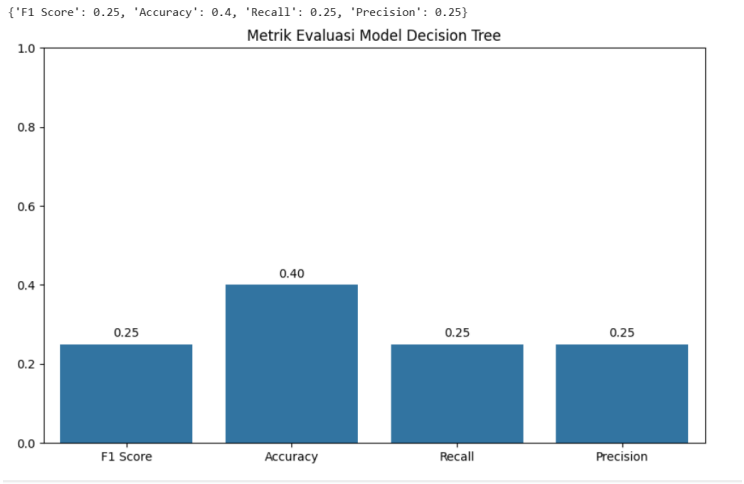
determined by cognitive abilities but also by self-regulation skills, self-efficacy, as well as positive social support and academic environment.
In addition, the research by Darmawan et al. also conducted a performance comparison between two different algorithms, namely Support Vector Machine (SVM) and Random Forest, in predicting the graduation of students from Madrasah Aliyah Swasta (MAS). The results of this study are expected to provide deeper insights into the effectiveness of these two algorithms in the context

RESEARCH METHOD

This research focuses on predicting student graduation using Decision Tree and Random Forest algorithms. The object of this research is the academic data of students taken from the relevant university. The data used includes information such as course grades, Cumulative Grade Point Average (GPA), the number of credits taken, as well as data from students who have completed at least 4 semesters. The data is focused on students from a specific faculty to maintain relevance with the research objectives. This research also collects data from students who have graduated within the last five years to ensure that the analysis results are more representative.
The academic data used includes several important variables, namely GPA, the number of credits taken each semester, and other related academic information. This data was collected from secondary sources, namely the university's academic database, which was then processed for predictive analysis. The algorithms used in this analysis are Decision Tree and Random Forest, both of which aim to predict student graduation categorized as "on time" and "not on time." The main focus of this research is to evaluate the extent to which these two algorithms are able to predict graduation based on the available academic data.
This research also compares the two classification algorithms in terms of accuracy, precision, and recall. In other words, this research aims to assess the effectiveness of Decision Tree and Random Forest in predicting student graduation, as well as to determine which algorithm is superior in predictive analysis of academic data. In addition, this research aims to identify the most dominant factors influencing graduation predictions. Several variables analyzed include the Student Identification Number (NPM), GPA, the number of credits taken, as well as other variables relevant to students' academic performance.
In the context of data collection, the technique used is the documentation method. Data is taken from the existing university academic information system, so the data collection process is carried out electronically and efficiently. The collected data includes academic information such as student names, GPA, number of credits, and student attendance percentage. The data collection process begins with a formal request for access to the university's data management, accompanied by the necessary permissions to access the academic database.

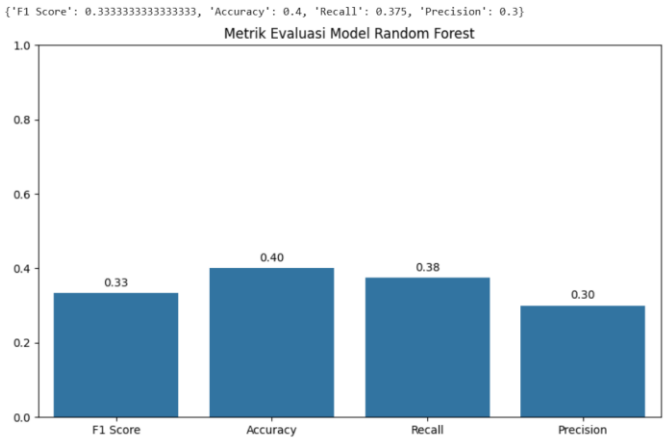
In addition, this research also highlights the differences between primary and secondary data. Primary data is collected directly by researchers through surveys, interviews, or observations, while secondary data is taken from existing sources, such as official reports or data from academic institutions. In this study, the secondary data taken from the university database is more relevant because it includes historical information about the academic achievements of graduated students. By using secondary data, this research is able to provide a broader and more representative picture of student graduation patterns.

RESULT AND DISCUSSION



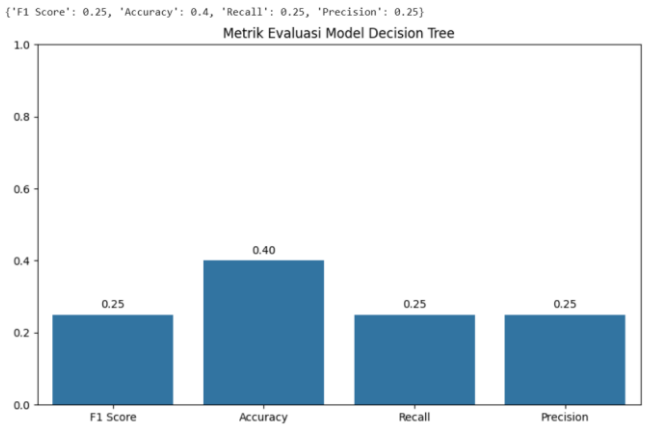
picture 1. Class of 2016 Testing

An F1 Score of 0.25 indicates that the model is not good at balancing Precision and Recall. An accuracy of 0.40 means only 40% of predictions are correct. A recall of 0.25 indicates the model only detected 25% of positive samples, indicating the model is less sensitive. Precision 0.25 means that only 25% of positive predictions are actually correct, indicating a lot of error in positive predictions.



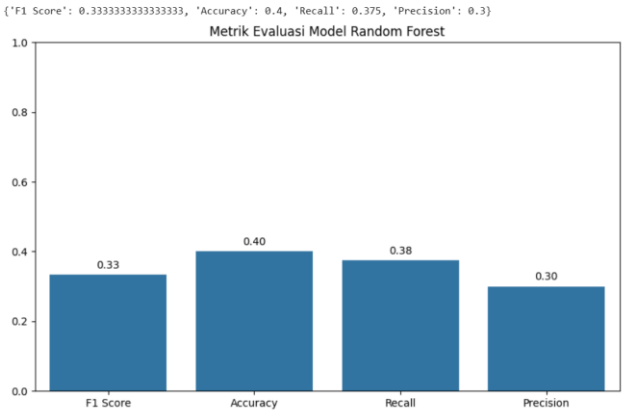
picture 2. Class of 2016 Testing

F1 Score 0.33 indicates that the balance between accuracy and sensitivity is still low. Accuracy 0.40 means that 40% of the predictions are correct, there are still many errors. Recall 0.38 indicates that the model detects 38% of positive samples, slightly better but not ideal. Precision 0.30 means that only 30% of positive predictions are correct.



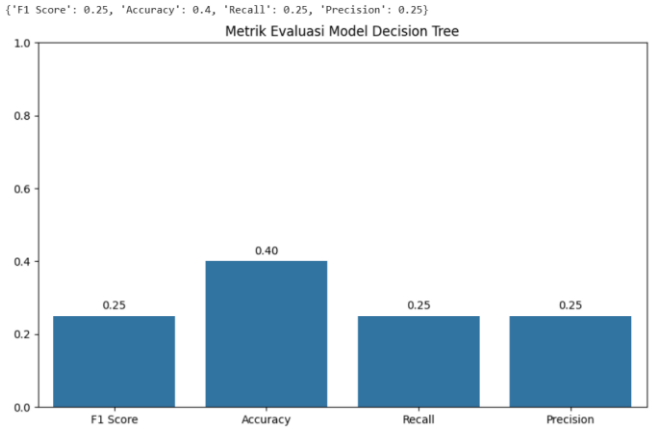
picture 3. Class of 2017 Testing

The F1 Score is 0.25, indicating that the balance between the precision and sensitivity of the model is still low. An accuracy of 0.40 means that only 40% of predictions are correct, indicating that the model often makes mistakes. A recall of 0.25 indicates that the model is only able to detect 25% of the true positive samples. Precision 0.25 means that only 25% of positive predictions are correct.



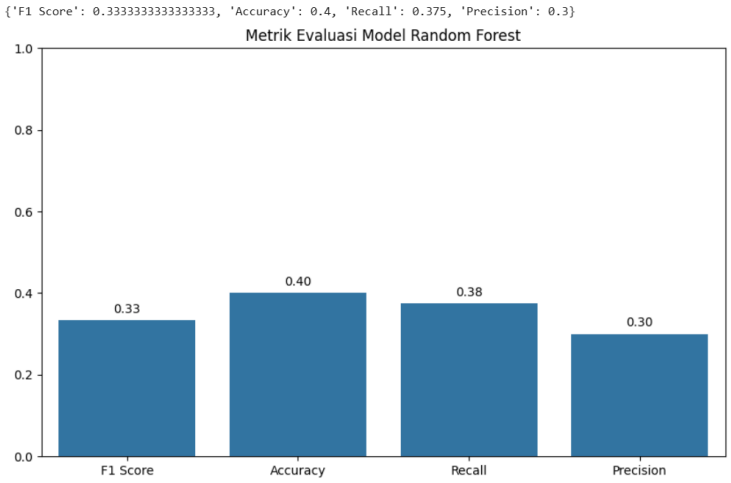
picture 4. Class of 2017 Testing

The F1 Score is 0.33, indicating that the balance between precision and recall is better than the previous model, but still low. An accuracy of 0.40 means that only 40% of predictions are correct, still showing many errors. A recall of 0.38 indicates that the model successfully detects 38% of positive samples, slightly better than before. Precision 0.30 means that only 30% of positive predictions are completely accurate.



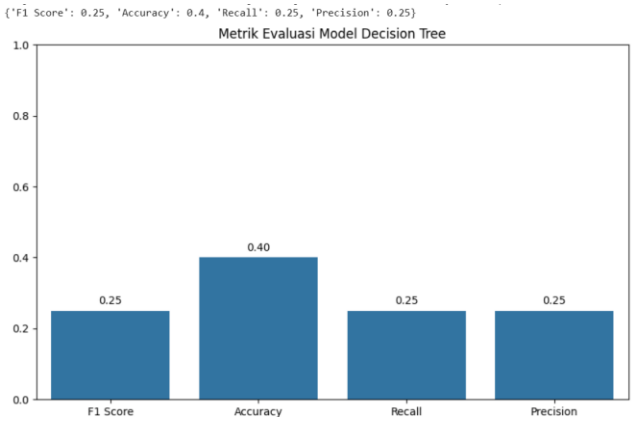
picture 5. Class of 2018 Testing

The F1 Score is 0.25, indicating that the balance between precision and sensitivity is still low. An accuracy of 0.40 means that only 40% of predictions are correct, indicating that the model still makes mistakes frequently. A recall of 0.25 indicates that the model only manages to detect 25% of positive samples, indicating that the model is not very sensitive. Precision 0.25 means that only 25% of positive predictions are correct, indicating many errors in predicting positive.



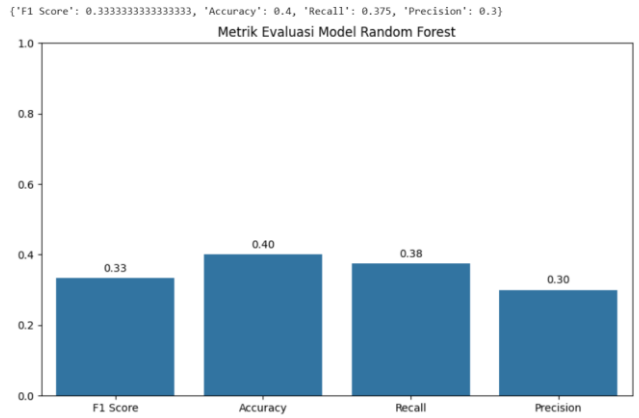
picture 6. Class of 2018 Testing

An F1 Score of 0.33 indicates that the balance between precision and recall is slightly better, but still low. Accuracy is 0.40, which means that 40% of the predictions made by the model are correct. Recall 0.38 shows that the model is able to detect 38% of positive samples, better than the previous model but not ideal. Precision 0.30 means that only 30% of positive predictions are completely accurate.



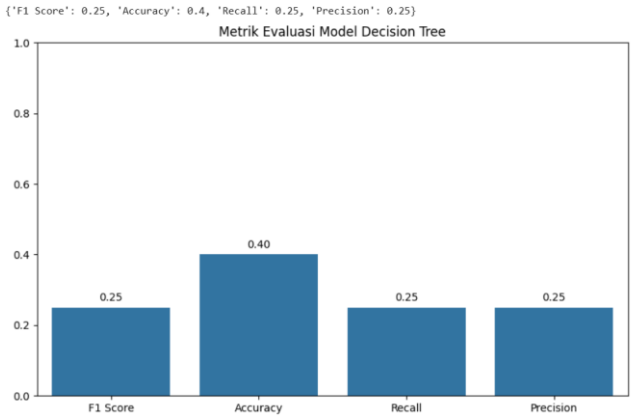
picture 7. Class of 2019 Testing

An F1 Score of 0.25 indicates that the balance between precision and sensitivity (recall) is quite low. Accuracy is 0.40, meaning that only 40% of the predictions made by the model are correct. Recall 0.25 indicates that the model can only detect 25% of the positive samples. Precision 0.25 means that only 25% of the positive predictions are actually correct.



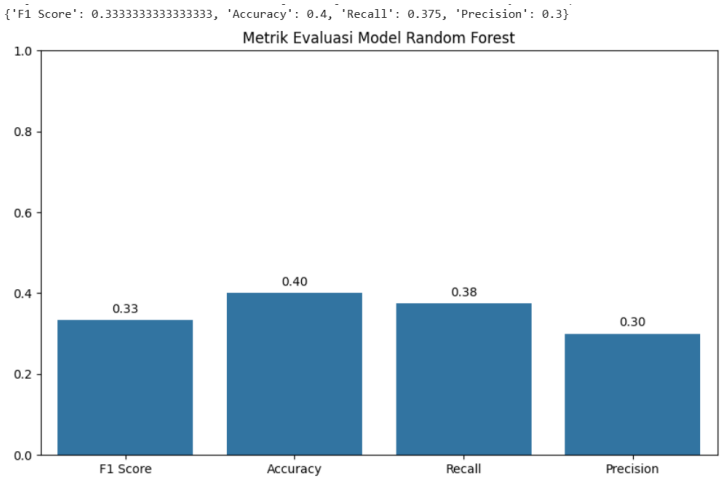
picture 8. Class of 2019 Testing

An F1 Score of 0.33 indicates a balance between accuracy and sensitivity, slightly better than the Decision Tree model. Accuracy is still 0.40, meaning that 40% of this model's predictions are correct. Recall of 0.38 indicates that the model successfully detects 38% of positive samples, slightly better than the previous model. Precision is 0.30, meaning that 30% of positive predictions are correct.



picture 9. Class of 2020 Testing

An F1 Score of 0.25 indicates a low balance between accuracy and sensitivity. An accuracy of 0.40 means that only 40% of the predictions made by this model are correct. A recall of 0.25 indicates that the model only manages to detect 25% of the positive samples. Precision is also 0.25, meaning that 25% of the model's positive predictions are correct.



picture 10. Class of 2020 Testing

An F1 Score of 0.33 indicates that the balance between accuracy and sensitivity is slightly better than the previous model. Accuracy is still 0.40, which means that 40% of this model's predictions are correct. Recall increases to 0.38, which indicates that the model successfully identifies 38% of positive samples. A precision of 0.30 indicates that 30% of the model's positive predictions are correct.

Analysis/Discussion

Table 1. Comparison Results of Decision Tree and Random Forest Algorithms

no	Angkatan	AlgoritmaDecision Tree	Algoritma Random Forest
1.	2016	F1 Score = 0.25	F1 Score = 0.33

		Accuracy = 0.40 Recall = 0.25 Precision = 0.25	Accuracy =0.40 Recall = 0.38 Precision = 0.30
2.	2017	F1 Score = 0.25 Accuracy = 0.40 Recall = 0.25 Precision = 0.25	F1 Score = 0.33 Accuracy = 0.40 Reccal = 0.38 Precision = 0.38
3	2018	F1 Score = 0.25 Accuracy = 0.40 Recall = 0.25 Precision = 0.25	F1 Score = 0.33 Accuracy = 0.40 Reccal = 0.38 Precision = 0.38
4	2019	F1 Score = 0.25 Accuracy = 0.40 Recall = 0.25 Precision = 0.25	F1 Score = 0.33 Accuracy = 0.40 Reccal = 0.38 Precision = 0.38
5	2020	F1 Score = 0.25 Accuracy = 0.40 Recall = 0.25 Precision = 0.25	F1 Score = 0.33 Accuracy = 0.40 Reccal = 0.38 Precision = 0.38

comparison between the performance of two machine learning algorithms, namely Decision Tree and Random Forest, for datasets from several years (2016 to 2020). The metrics used to measure model performance are F1 Score, Accuracy, Recall, and Precision. From this table, it can be seen that the F1 Score and Accuracy values for the Decision Tree algorithm remain consistent every year, respectively with an F1 Score of 0.25 and an Accuracy of 0.40. Apart from that, the Recall and Precision for the Decision Tree also remain at 0.25.

Meanwhile, Random Forest consistently performs slightly better than Decision Tree. The F1 Score for Random Forest is always higher, namely 0.33, and accuracy remains at 0.40. Recall and Precision Random Forest are also better with Recall values varying from 0.38, although in some parts there are typos in writing "Recall" as "Reccal". Precision for Random Forest varies between 0.30 to 0.38 across periods. Overall, Random Forest provides superior results to Decision Tree on most metrics, although the differences are not very significant in some aspects.

CONCLUSION

Based on the research conducted, it appears that the Random Forest algorithm is superior to the Decision Tree in predicting student graduation based on academic data. The evaluation results show that Random Forest has higher accuracy, especially in terms

of recall and precision. This shows that Random Forest is more stable in making predictions and better at handling potential data variations.

On the other hand, Decision Trees, although simpler and easier to interpret, as well as faster in processing, tend to experience overfitting. This makes the predictions generated less optimal when faced with new data that differs from the training data.

Research also reveals that academic factors such as the Cumulative Grade Point Average (GPA), the number of credit hours, and course grades significantly influence student graduation. Students with a high GPA and sufficient credit hours have a greater chance of graduating on time.

Therefore, the development of a prediction system based on academic data is very important for educational institutions. With this system, institutions can identify students at risk of not graduating on time and provide early interventions, helping them to complete their studies more successfully.

REFERENCES

Amri, Z., Kusrini, K., & Kusnawi, K. (2023). Prediksi Tingkat Kelulusan Mahasiswa menggunakan Algoritma Naïve Bayes, Decision Tree, ANN, KNN, dan SVM. *Edumatic: Jurnal Pendidikan Informatika*, *7*(2), 187–196. https://doi.org/10.29408/edumatic.v7i2.18620

Budiyantara, A., & A, I. (2018). Prediksi Mahasiswa Lulus Tepat Waktu. *Infotech: Journal of Technology Information*, *5*(2), 7–13. https://doi.org/10.37365/it.v5i2.39

Darmawan, A., Yudhisari, I., Anwari, A., & Makruf, M. (2023). Pola Prediksi Kelulusan Siswa Madrasah Aliyah Swasta dengan Support Vector Machine dan Random Forest. *Jurnal Minfo Polgan*, *12*(1), 387–400. https://doi.org/10.33395/jmp.v12i1.12388

Latifah, S. L. S. N. H. (2020). *Prediksi Prestasi Akademik Mahasiswa Menggunakan Random Forest dan C.45*. *VIII*(1), 47–52. www.bsi.ac.id

Linawati, S., Nurdiani, S., Handayani, K., & Latifah, L. (2020). Prediksi Prestasi Akademik Mahasiswa Menggunakan Algoritma Random Forest Dan C4.5. *Jurnal Khatulistiwa Informatika*, *8*(1), 47–52. https://doi.org/10.31294/jki.v8i1.7827

Permatasari, R. P. (2021). *Implementasi algoritma decision tree untuk prediksi kelulusan mahasiswa tepat waktu laporan skripsi*.

Plaosan, van S. (2019). Algoritma Random Forest. *Http://Learningbox.Coffeecup.Com/05_2_Randomforest.Html*, *18*(1), 10–14.

Prediksi kelulusan pelajar menggunakan decision tree. (2023). 90–94.

Satrio Junaidi, Valicia Anggela, R., & Kariman, D. (2024). Klasifikasi Metode Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa dengan Algoritma Naïve Bayes, Random Forest, Support Vector Machine (SVM) dan Artificial Neural Nerwork (ANN). *Journal of Applied Computer Science and Technology*, *5*(1), 109–119. https://doi.org/10.52158/jacost.v5i1.489

Wahyudi, A. (2023). Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Decision Tree Dan Naïve Bayes. *Jurnal Permata Indonesia*, *14*(2), 132–138. https://doi.org/10.59737/jpi.v14i2.276

Zeniarja, J., Salam, A., & Ma'ruf, F. A. (2022). Seleksi Fitur dan Perbandingan Algoritma

Klasifikasi untuk Prediksi Kelulusan Mahasiswa. *Jurnal Rekayasa Elektrika*, *18*(2), 102–108. https://doi.org/10.17529/jre.v18i2.24047