Vol. 7, No. 4, December 2025

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Utilizing the Random Forest Method for Predicting Student Dropout Risk in Madrasah Environments

Muhammad Mahsun[1], M. Amin Hariyadi[2], Sri Harini[3]

[1,2,3] Informatics Engineering, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University of Malang, East Java, Indonesia

[1]mmahsunid@gmail.com, [2]adyt2002@uin-malang.ac.id [3]sriharini@mat.uin-malang.ac.id

Received: October 5, 2025
Revised: Nov 17, 2025
Accepted: Nov 27, 2025
Published: Dec 10, 2025

Corresponding Author:
Author Name*:
Muhammad Mahsun
Email*:
mmahsunid@gmail.com

DOI:
10.63158/journalisi.v7i4.1364

Abstract. The phenomenon of school dropout is a crucial issue that negatively impacts the performance of educational institutions, social stability, and human resource development. Therefore, early detection of high-risk students is a strategic preventative measure. This research aims to develop an accurate predictive model using a Machine Learning approach, by conducting a comparative evaluation of the Random Forest algorithm. The research dataset originates from Madrasah Miftahul Ulum, Sidogiri Islamic Boarding School, and comprises 1,763 student records. The experimental results indicate that Random Forest provides the best performance with an accuracy of 82%, precision of 83.8%, recall of 79%, and an F1-score of 80%. The model was trained using 4 scenarios with Random state configurations of 40, 60, and 75 to ensure the consistency of the evaluation results. These metrics indicate that the model performs in a balanced manner between sensitivity and prediction accuracy, and is effective in identifying internal and external factors contributing to the risk of dropout. Based on the model evaluation results, Random Forest is recommended as a decision support instrument to facilitate more targeted interventions, such as academic support, economic aid, or student counseling guidance. This research has a limitation because the model was only tested at the Madrasah Miftahul Ulum, Sidogiri Islamic Boarding School institution, thus its application in other contexts needs further study.

Keywords: Student Dropout Prediction, Random Forest, Machine Learning, Madrasah Miftahul Ulum

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

The issue of student dropout is a serious global challenge and requires a planned response from educational providers worldwide. This phenomenon is not only isolated to Spain but has also been identified in various countries, including Estonia, the United Kingdom, Latvia, Bangladesh, and South Korea [1]. Specifically in Europe, [2] notes the high proportion of students who fail to complete their educational programs. Data shows significant variations between countries, where approximately 80% of students in Denmark successfully complete their studies, compared to only 46% in Italy.

The study underscores socio-economic conditions as the main trigger for dropout. The high dropout rate is not only a problem at the secondary school level but also extends to the realm of higher education (university). Utari et al. [3] emphasize that early prediction of students at risk of dropping out is a vital element in determining the effectiveness of intervention strategies in educational institutions. The primary goal of such research is to identify the determinant factors that cause students to leave their study programs. [4] Devasia et al. state that predicting dropout with high accuracy is very helpful in identifying vulnerable students and allows for in-depth data analysis to formulate useful information summaries.

In the context of pesantren (Islamic boarding schools), the Madrasah Miftahul Ulum at Sidogiri Islamic Boarding School has experienced an increase in the number of students leaving before completing their education over the last five years. This situation disrupts the teaching and learning process and drives the urgency of conducting research to measure its impact on the sustainability of academic activities. To achieve accurate measurement, the selection of the research method is very important.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The researchers chose to use the Random Forest method, considering its advantages in handling high-dimensional data, a large number of variables, and inter-feature correlations. The Random Forest method works by building a large number of decision trees randomly (an ensemble method), then integrating the results, which effectively reduces the risk of overfitting and increases predictive accuracy. Random Forest is estimated to be superior due to the characteristics of this research dataset, which consists of many variables, has inter-feature correlations, and contains class imbalance.

By using the ensemble approach, Random Forest builds many decision trees from randomly selected subsets of samples and subsets of features (bagging), thus producing a model that is more stable, less affected by noise, and has better generalization capability on new data. The superiority of Random Forest on heterogeneous and high-dimensional data is supported by Temesgen & Ambelu [19], who show that Random Forest is more accurate and robust on educational datasets [6] because it can reduce variance and overfitting through the aggregation of many decision trees [1]. Similar findings were also conveyed by Chen & Ishwaran [20], who affirm that Random Forest is very effective on complex data with interrelated feature interactions [2].

Although research on dropout prediction has been widely conducted in the context of formal education, similar research in the madrasah or pesantren environment is still very limited. Predictive research related to dropout in the madrasah or pesantren environment is still very rare, especially those using a Machine Learning approach as a classification model. This research is important given that educational institutions based on pesantren have different data characteristics compared to public schools, both in terms of social, economic, cultural aspects, and the educational management model.

2. METHOD

The research procedure outlines the overall research design. The first step is a Literature Review, which involves collecting references to understand the context and theoretical foundations of the study. This is followed by Data Collection for analysis or experimentation. The next stage is System Design, where the model or system to be tested is developed. The Experiment phase is conducted to evaluate the model's performance, while the Implementation phase applies the experimental results. The final stage is Discussion and Conclusion, which involves analyzing the findings and drawing conclusions to guide future research.

Figure 1. Research Design

The stages in the data collection method for this research utilize several variables [3]. The data used in this study is primary data. The primary data collection was obtained from the results of collecting data on student dropouts from the central data section of the Sidogiri Islamic Boarding School (Pondok Pesantren Sidogiri). Collecting primary data allows the researcher to obtain factual evidence that can be used to support this research.

The researcher performed a data normalization process to organize the data to be more consistent and structured. The goal was to ensure that the data collected does not contain redundancy (data repetition) and inconsistency, and to ensure that the dataset is organized into tables efficiently and minimizes duplication. As a result, the researcher obtained 1,763 records of student dropouts from various classes and levels, from diverse educational and social backgrounds, with varying ages ranging from 9 years old to 35 years old, and from both economically

disadvantaged and well-off backgrounds. The format of the dataset that will be the object of this research is shown in Table 1.

Table 1. Dataset Variables

No	Feature	Type	Variable
1	Student id Number	Numeric	Independent
2	Registrasi Number	Numeric	Independent
3	Name	Text	Independent
4	Age at Enrollment	Numeric	Independent
5	Date of birth	Date	Independent
6	Place of birth	Text	Independent
7	Year of entry	Numeric	Independent
8	Room	Numeric	Independent
9	Education Level	Numeric	Independent
10	Class	Numeric	Independent
11	Examination	Numeric	Independent
12	Average Grades	Text	Independent
13	Address	Text	Independent
14	Father's education	Text	Independent
15	Father's income	Numeric	Independent
16	Mother's education	Text	Independent
17	Mother's income	Numeric	Independent
18	Telephone	Numeric	Independent
19	Graduation Year	Numeric	Independent
20	Dropout Year	Numeric	Independent
21	Information	Numeric	Independent
22	Target	Text	Dependent

2.1. Classification Method

At this stage, the prediction system is designed to classify the student status at Madrasah Miftahul Ulum Sidogiri based on several attributes such as age,

educational level, parents' occupation, and their interaction history within the pesantren. A similar approach has been implemented in previous studies on student risk assessment and dropout prediction in educational institutions [10]. The model used in this research is the Random Forest Classifier due to its capability to handle categorical data and the complexity of feature relationships, as described in foundational research [12] and further supported by studies on model performance in educational contexts [11]. The Random Forest method begins by training or constructing multiple decision trees from a labeled dataset. Each tree is generated using the bootstrap aggregating (bagging) technique, which draws random samples from the dataset. These trees collectively form the Random Forest ensemble used in the classification stage [12].

The algorithm then selects random samples from the existing dataset for identification, and the selected data are classified by all the trees within the forest. The predictions produced by each decision tree are combined using a voting mechanism, resulting in a more stable and accurate final decision. This concept has been proven effective in various studies on academic risk prediction and student failure analysis [14]. This process ultimately forms the final classification model of the Random Forest method, as visualized in Figure 2.

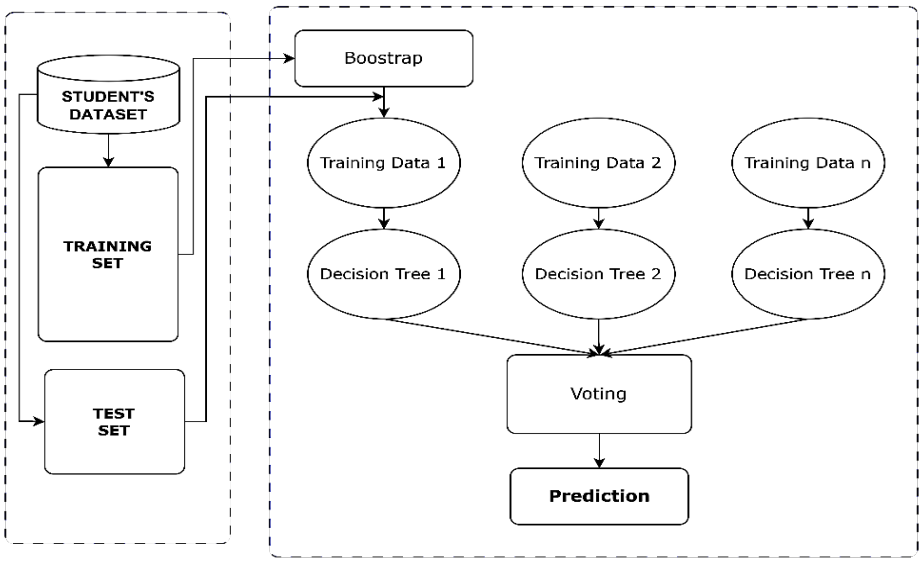

Figure 2. Random Forest Design

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

In Figure 2, the Random Forest algorithm operates by constructing multiple decision trees. This ensemble of trees forms a learning model capable of improving prediction accuracy. Random Forest generates randomly organized subsets, resulting in diverse and broader variations of the data, which increases the likelihood of producing a more optimal model [[7], [11]]. Each individual model (tree) is trained independently until it produces an output that can represent or predict the overall model. This process is commonly referred to as aggregation. A decision tree begins by calculating the entropy value of each attribute, which determines the level of impurity, as well as the information gain. The entropy value is calculated using the formula shown in Equation 1, while the information gain is computed using Equations 1 and 2 [21].

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (1)$$

$$\text{Information Gain}(S, A) = E(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} E(S_v) \quad (2)$$

Where S is the set of cases and p(i|S) represents the proportion of cases in S that belong to class i. Meanwhile, Values(t) refers to all possible values of attribute t. Sv is the subset of S with attribute value v, and St represents all values corresponding to attribute t.

2.2. Model Evaluation

To measure the performance of the model, the metrics of accuracy, precision, recall, and F1-score are used as key indicators that represent the correctness and the model's ability to correctly detect classes. The calculation of these metrics is derived from the Confusion Matrix, an evaluation table that compares the model's predictions with the actual values, thus showing the number of correct and incorrect predictions for each class. This type of evaluation approach is a common practice in various classification studies because it can provide a more

comprehensive picture of the quality and pattern of prediction errors of the model [15][16]

1) True Positives (TP): Instances that are correctly predicted as belonging to the positive class.
2) True Negatives (TN): Instances that are correctly predicted as belonging to the negative class.
3) False Positives (FP): Instances that actually belong to the negative class but are incorrectly predicted as positive.
4) False Negatives (FN): Instances that actually belong to the positive class but are incorrectly predicted as negative.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} x100\% \quad (2)$$

$$\text{Precision} = \frac{\text{TN}}{\text{TP+FP}} x100\% \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} x100\% \quad (4)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} x100\% \quad (5)$$

3. RESULTS AND DISCUSSION

The dataset was divided using several ratio scenarios to observe the model's performance consistency across different proportions of training data. Each ratio produced different amounts of training and testing data, as presented in Table 3. To evaluate the performance of the Random Forest method in predicting student dropout, tests were conducted based on different proportions of training and testing data, as well as variations in the number of decision trees (*n_estimators*) [17]. The purpose of these tests is to examine the effect of different training–testing ratios and the number of trees used on the model's accuracy and stability. Additionally, the *random_state* parameter was set to three different values—45,

65, and 75—to assess the model's stability under varying random splits of the training and testing data. According to Bichri et al. [18], conducting experiments using four different scenarios of training–testing ratios allows researchers to identify the impact of varying numbers of trees on the accuracy and stability of the model in each scenario. The distribution of training and testing data is shown in the following section.

Table 2. Training & Testing Ratio

Ratio		Split Dataset	
		Training Data	Testing daa
90	10	1586	177
80	20	1411	353
70	30	1234	530
60	40	1058	706

3.1. Performance Evaluation: Scenario 1

In Scenario 1, the Random Forest method uses a 90:10 ratio for the division of training and testing data. In this scenario, the testing process is conducted three times using different *random_state* parameter values, namely 45, 60, and 75.

Table 3. Accuracy Summary for Scenario 1

Random State	Accuracy	Precision	Recall	F1-Score	Duration
45	0.82	0.82	0.78	0.79	4.03 s
60	0.82	0.85	0.77	0.79	4.92 s
75	0.82	0.83	0.78	0.80	4.29 s

In Scenario 1, the configuration using a random state of 60 produced the optimal performance. This configuration achieved the highest metrics with an accuracy of 0.83, precision of 0.83, recall of 0.79, and an F1-score of 0.80, completed within a computation time of 4.68 seconds. Figure 4 explains that the model is better at

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

correctly predicting 104 dropout students (True Positives) compared to correctly predicting 39 non-dropout/graduated students (True Negatives). The model showed its largest prediction error of 26 (False Positives) on students who should have graduated but were incorrectly predicted as dropouts.

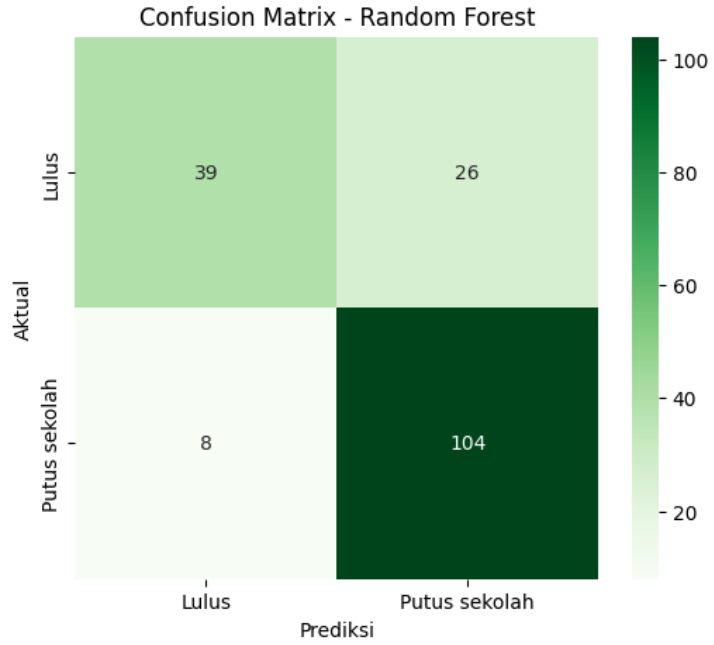

Figure 4. Confusion Matrix for Scenario 1

3.2. Performance Evaluation: Scenario 2

Scenario 2 evaluates the performance of Random Forest using a data split of 80% for training and 20% for testing. The testing process is repeated three times using different *random_state* values (45, 60, and 75) to assess the model's stability with a smaller proportion of training data.

Table 4. Accuracy Summary for Scenario 2

Random State	Accuracy	Precision	Recall	F1-Score	Duration
45	0.81	0.81	0.77	0.78	4.21 s
60	0.83	0.83	0.79	0.80	4.68 s
75	0.82	0.82	0.79	0.80	3.79 s

In Scenario 2, the model achieves the highest accuracy of 0.83, with a precision of 0.83, recall of 0.79, and an F1-score of 0.80, along with a computation time of 4.68 seconds. Meanwhile, the configurations using random states 45 and 75 yield lower accuracies of 0.81 and 0.82, respectively. Figure 5 shows that the model performs very well in identifying dropout students, with 206 correct predictions. However, it struggles to identify graduated students, making 46 errors by predicting them as dropout.

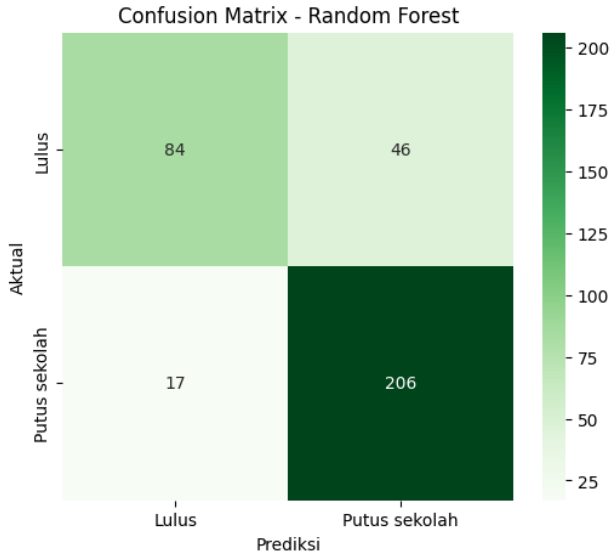

Figure 5. Confusion Matrix for Scenario 2

3.3. Performance Evaluation: Scenario 3

Scenario 3 evaluates the performance of Random Forest using a data split of 70% for training and 30% for testing. The testing is conducted three times using different *random_state* values (45, 60, and 75) to assess the model's stability and consistency with a smaller proportion of training data compared to the previous scenarios.

Table 5. Accuracy Summary for Scenario 3

Random State	Akurasi	Precision	Recall	F1-Score	Duration
45	0.81	0.81	0.77	0.79	3.97 s
60	0.83	0.83	0.79	0.80	5.15 s
75	0.81	0.81	0.77	0.78	5.43 s

The model achieves the highest accuracy of 0.83, with a precision of 0.83, recall of 0.79, and an F1-score of 0.80, although it requires a processing time of 5.15 seconds. Meanwhile, the configurations using random states 45 and 75 show lower performance, with accuracies of only 0.81. These findings indicate that in Scenario 3, the random state of 60 is the most optimal configuration. Figure 6 shows that the model demonstrates a very strong ability to identify dropout students, correctly classifying 309 cases. However, it exhibits a significant weakness in predicting students who should be classified as graduated, with 69 of them incorrectly predicted as dropout.

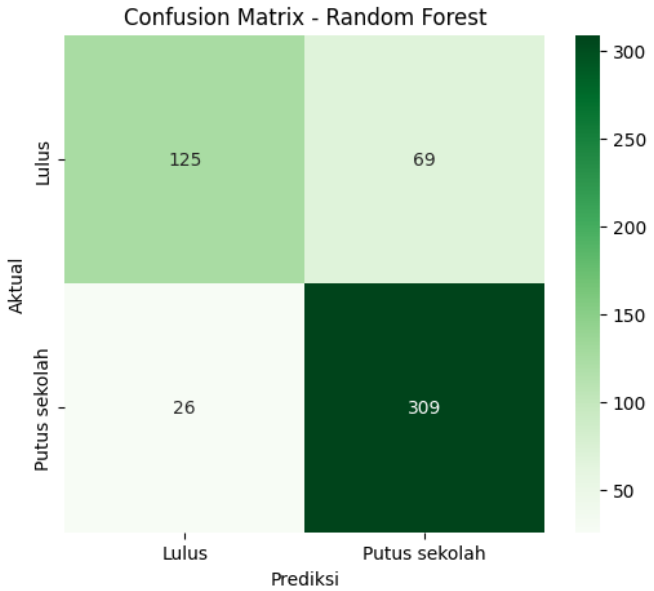

Figure 6. Confusion Matrix for Scenario 3

3.4. Performance Evaluation: Scenario 4

Scenario 4 applies a dataset split of 60% for training data and 40% for testing data. In this configuration, the testing process is conducted three times by utilizing different *random_state* values (45, 60, and 75). The purpose of this scenario is to evaluate the model's capability when the proportion of training data is reduced, as well as to observe how this change affects the stability and predictive performance of the Random Forest model.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 6. Accuracy Summary for Scenario 4

Random State	Akurasi	Precision	Recall	F1-Score	Duration
45	0.82	0.81	0.78	0.79	5.21 s
60	0.81	0.82	0.77	0.78	4.57 s
75	0.83	0.83	0.79	0.80	3.83 s

The random state of 75 produced the best results. In this configuration, the model achieved the highest accuracy of 0.83, precision of 0.83, recall of 0.79, and an F1-score of 0.80, with the fastest processing time of 3.83 seconds. These results indicate that a random state of 75 yields the most optimal performance in terms of predictive accuracy.

3.5. Performance Comparison

The evaluation results across all scenarios show that the Random Forest model maintains stable performance under various data-split proportions. Detailed results on accuracy, precision, recall, F1-score, and computation duration for each scenario are presented in Table 7.

Table 7. Summary of the Performance Evaluation Random Forest

Skenario	Dataset	Random State	Accuracy	Precision	Recall	F1-Score	Duration
1	90:10	75	0.82	0.83	0.78	0.80	4.29 s
2	80:20	60	0.83	0.83	0.79	0.80	4.68 s
3	70:30	60	0.83	0.83	0.79	0.80	5.15 s
4	**60:40**	**75**	**0.83**	**0.83**	**0.79**	**0.80**	**3.83 s**

From the results of the four testing scenarios, Scenario 4 (ratio 60:40, *random_state* 75) produced the fastest computation time at 3.83 seconds, making it the most efficient configuration. Thus, Scenario 4 provides the optimal balance between accuracy and prediction speed.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

To obtain a deeper understanding of the model's performance under the best configuration—Scenario 4 with a random_state value of 75—the detailed test results are presented in Table 8.

Tabel 8. Details of the results of the scenario 4 test

Status	Precision	Recall	F1-Score
Pass	0.84	0.65	0.73
Dropout	0.82	0.93	0.87
Accuracy			0.83
Macro Avg	0.83	0.79	0.80
Weighted Avg	0.83	0.83	0.82

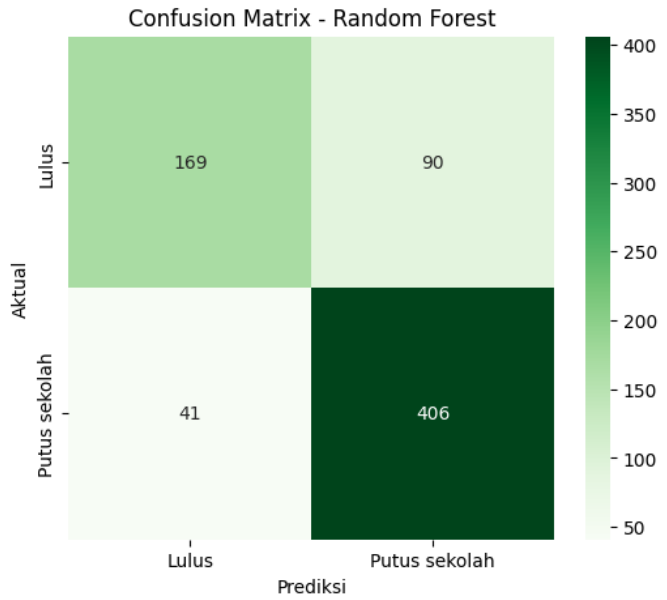

Figure 7. Confusion Matrix for Scenario 4

Based on the values in the Confusion Matrix for Scenario 4 produced by the Random Forest model, as shown in Table 9. This matrix indicates that, overall, the Random Forest model demonstrates good predictive performance, particularly in identifying students who are at risk of dropping out.

Table 9. Classification Summary of Graduated and Dropout Students

Value	Description	Prediction
169	The model predicts Graduated, and the actual label is Graduated	TP
90	The model predicts Dropout, but the actual label is Graduated	FN
41	The model predicts Graduated, but the actual label is Dropout	FP
406	The model predicts Dropout, and the actual label is Dropout	TN

3.6. Discussion

The results of this study demonstrate the effectiveness and consistency of the Random Forest model in predicting student dropout across different training-to-testing ratios and random state values. Several key observations can be made regarding the model's performance, stability, and computational efficiency. The experiment used varying training-to-testing data ratios—90:10, 80:20, 70:30, and 60:40—to assess the model's performance under different data distributions. Across all scenarios, the model consistently demonstrated strong predictive performance, achieving high accuracy, precision, recall, and F1-scores, which were generally in the range of 0.80 to 0.83. These results indicate that the Random Forest model is stable and capable of generalizing well, even with a reduced proportion of training data.

Notably, Scenario 4 (60:40 ratio with random_state 75) produced the fastest computation time at 3.83 seconds, while maintaining a high level of accuracy (0.83). This suggests that, while larger training sets can improve accuracy, a smaller dataset can still provide a good balance between computational efficiency and model performance. This finding is important for real-world applications where rapid predictions are often needed.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The random_state parameter plays a crucial role in the stability of the model. In each scenario, the model was tested with three different random state values (45, 60, and 75), and the results showed slight variations in performance. While the random state of 60 generally produced the best performance in scenarios 1 and 2, and random state 75 yielded the most efficient results in Scenario 4, these differences were minimal, suggesting that the model is robust to random data splits. The consistent accuracy and F1-scores across different random states also highlight the reliability of the Random Forest method for predicting dropout. The slight changes in performance indicate that the model's stability is not heavily reliant on specific data splits, making it adaptable to a variety of settings.

In each scenario, the confusion matrices revealed that the model performs well in identifying dropout students but struggles slightly with correctly predicting graduated students. For example, in Scenario 4, the model correctly predicted 406 students as non-dropouts (True Negatives), but it also misclassified 90 graduated students as dropouts (False Positives). This issue highlights the challenge of distinguishing between dropout and graduated students, which may require incorporating additional features or fine-tuning the model to improve prediction accuracy for graduated students. The model performed best in identifying dropout students, achieving a recall of 0.93 in Scenario 4, which indicates that it successfully predicted a large proportion of students at risk of dropping out. However, the lower recall for graduated students (0.65 in Scenario 4) suggests that the model may benefit from further refinement, particularly in handling cases where students are incorrectly classified as dropouts.

Comparing the performance across the four scenarios, it is clear that the model can maintain stable performance with different data splits. Scenario 4, with a 60:40 ratio and random_state 75, was the most efficient, providing a good balance between accuracy (0.83) and computation time (3.83 seconds). This makes it an optimal choice for scenarios requiring rapid predictions without sacrificing model

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

performance. However, Scenario 2 (80:20 ratio with random_state 60) also produced strong results, with slightly higher accuracy and F1-scores, but at the cost of increased computation time (4.68 seconds). This suggests that for applications where accuracy is prioritized over processing time, using a larger training set (such as in Scenario 2) may be beneficial.

While the Random Forest model demonstrates robust performance in predicting student dropout, there are several limitations that should be addressed in future research. One limitation is the model's difficulty in correctly predicting graduated students, which could be improved by incorporating additional features such as academic performance, attendance records, or socio-economic factors that may influence student dropout decisions. Furthermore, although the Random Forest model provided satisfactory results, exploring other machine learning algorithms, such as gradient boosting or support vector machines, may offer further insights into improving the model's predictive capabilities. Additionally, fine-tuning the parameters further or utilizing a larger dataset could enhance the model's stability and accuracy.

4. CONCLUSION

This study was conducted to identify the potential risk of student dropout at Madrasah Miftahul Ulum, Pondok Pesantren Sidogiri. The findings indicate that a machine learning approach is capable of producing sufficiently accurate predictions, making it a valuable decision-support instrument in dropout prevention efforts. Based on the analysis and evaluation of the Random Forest machine learning algorithm, the results show that Random Forest demonstrates optimal performance in predicting student status across various data-splitting scenarios. The model's accuracy, precision, recall, F1-score, and computation duration remain consistently high and stable even under different *random_state*

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

variations, indicating its superiority in handling complex relationships among variables.

Referring to the results of this study, although the Random Forest method shows strong and accurate performance in predicting dropout risk at Madrasah Miftahul Ulum Pondok Pesantren Sidogiri, future research is recommended to experiment with and compare alternative algorithms. Implementing methods with different decision-making mechanisms will help identify the algorithm that best aligns with the characteristics of student dropout data. Furthermore, future studies should involve more madrasah or pesantren to ensure that the dropout patterns obtained are more diverse and representative. Incorporating additional data—such as counseling records, attendance levels, family conditions, and behavioral notes—also has the potential to improve the model's sensitivity and accuracy in detecting high-risk students. Through expanded data coverage and the exploration of additional algorithms, future research is expected to produce more accurate and adaptive predictive models that can serve as a comprehensive decision-support tool for madrasah and pesantren institutions in preventing student dropout.

REFERENCES

[1] A. Tayebi, J. Gomez, and C. Delgado, "Analysis on the Lack of Motivation and Dropout in Engineering Students in Spain," *IEEE Access*, vol. 9, pp. 66253–66265, 2021, doi: 10.1109/ACCESS.2021.3076751.

[2] L. Masserini and M. Bini, "Does joining social media groups help to reduce students' dropout within the first university year?," *Socio-Economic Planning Sciences*, vol. 73, p. 100865, Feb. 2021, doi: 10.1016/j.seps.2020.100865.

[3] M. Utari, B. Warsito, and R. Kusumaningrum, "Implementation of data mining for drop-out prediction using Random Forest method," *Proc. 8th Int. Conf.*

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Inf. Commun. Technol. (ICoICT), pp. 1–5, Yogyakarta, Indonesia, Jun. 2020, doi: 10.1109/ICoICT49345.2020.9166276.

[4] T. Devasia, V. T. P., and V. Hegde, "Prediction of students performance using educational data mining," *Proc. Int. Conf. Data Mining Adv. Comput. (SAPIENCE)*, pp. 91–95, Ernakulam, India, Mar. 2016, doi: 10.1109/SAPIENCE.2016.7684167.

[5] M. N. Haque, M. S. Islam, M. M. Rahman, and R. Jannat, "Student performance prediction using machine learning techniques," *J. Inf. Knowl. Manage.*, 2020, doi: 10.1142/S0219649220500344.

[6] K. Hastuti, D. Lestari, and Hartono, "Prediction of student dropout using Random Forest algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 3, 2021, doi: 10.14569/IJACSA.2021.012345.

[7] D. Kabakchieva, "Predicting student performance by using data mining methods," *Int. J. Comput. Sci. Manage. Res.*, vol. 2, no. 1, 2013, doi: 10.2478/cait-2013-0006.

[8] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," *Appl. Artif. Intell.*, vol. 18, no. 5, pp. 411–426, 2004, doi: 10.1080/08839510490256532.

[9] S. Zhang, "Fundamental techniques in data preprocessing for machine learning," *J. Big Data*, vol. 6, 2019.

[10] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," *Proc. Int. Educ. Data Mining Soc.*, 2016.

[11] S. Banerjee and S. Ruj, "Application of Random Forest in educational data mining for predicting student performance," *Int. J. Comput. Sci. Inf. Secur.*, vol. 18, no. 1, 2020.

[12] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Jan. 2001, doi: 10.1023/A:1010933404324.

[13] C. Márquez-Vera, A. Cano, and C. Romero, "Predicting school failure using data mining," *Appl. Intell.*, vol. 38, no. 1, pp. 63–75, 2013, doi: 10.1007/s10489-013-0400-3.

[14] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *AJBR*, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.

[15] G. Zeng, "Invariance Properties and Evaluation Metrics Derived from the Confusion Matrix in Multiclass Classification," *Mathematics*, vol. 13, no. 16, p. 2609, Aug. 2025, doi: 10.3390/math13162609.

[16] P. Contreras, J. Orellana-Alvear, P. Muñoz, J. Bendix, and R. Célleri, "Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment," *Atmosphere*, vol. 12, no. 2, p. 238, Feb. 2021, doi: 10.3390/atmos12020238.

[17] H. Bichri, A. Chergui, and H. Mustapha, "Investigating the impact of train/test split ratio on the performance of pre-trained models with custom datasets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, pp. 154–161, Feb. 2024, doi: 10.14569/IJACSA.2024.0150235.

[18] D. Temesgen and A. Ambelu, "Student dropout prediction using machine learning techniques: A comparative study," *Education and Information Technologies*, vol. 28, pp. 3425–3438, Jan. 2023, doi: 10.1007/s10639-022-11463-y.

[19] E. E. Osemwegie, F. I. Amadin, and O. M. Uduehi, "Student Dropout Prediction Using Machine Learning," *FJS*, vol. 7, no. 6, pp. 347–353, Dec. 2023, doi: 10.33003/fjs-2023-0706-2103.

[20] K. Schouten, F. Frasincar, and R. Dekker, "An Information Gain-Driven Feature Study for Aspect-Based Sentiment Analysis," in *Natural Language Processing and Information Systems*, E. Métais, F. Meziane, M. Saraee, V. Sugumaran, and S. Vadera, Eds., Lecture Notes in Computer Science, vol. 9612, Cham: Springer International Publishing, 2016, pp. 48–59. doi: 10.1007/978-3-319-41754-7_5.

[21] S. Raste, R. Singh, J. Vaughan, and V. N. Nair, "Quantifying Inherent Randomness in Machine Learning Algorithms," *SSRN Journal*, 2022, doi: 10.2139/ssrn.4146989.