JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9772
Hal 1075-1085
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Comparison of Logistic Regression and Random Forest Performance in Student Dropout Prediction based on Multi Source Data

Sartika Lina Mulani Sitio*, Sunardi, Abdul Fadlil

Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[1] Department of Informatics Engineering, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1]2536083044@webmail.uad.ac.id, [2]sunardi@mti.uad.ac.id, [3]fadlil@mti.uad.ac.id
Email Corresponding Author: 2536083044@webmail.uad.ac.id
Submitted **04-05-2026;** Accepted **09-06-2026;** Published **30-06-2026**

Abstract

The high rate of student dropouts is one of the important challenges in higher education because it can affect academic quality, learning effectiveness, and the performance of educational institutions. This condition encourages the need for a prediction system that is able to identify students at risk of dropouts early so that preventive measures can be taken appropriately. This study aims to compare the performance of Logistic Regression and Random Forest algorithms in predicting student dropout based on multi-source. The dataset consists of 4,424 student data with 34 attributes covering academic, demographic, socioeconomic, and academic administration aspects. The research stages include data preprocessing, target transformation into binary classification, feature scaling, data sharing using an 80:20 scheme, and handling class imbalances using the Synthetic Minority Oversampling Technique (SMOTE). Furthermore, a modeling process was carried out using Logistic Regression and Random Forest algorithms to predict the risk of student dropout. Model evaluation was carried out using accuracy, precision, recall, F1-score, and Area Under Curve Receiver Operating Characteristic (AUC-ROC). The results showed that Random Forest performed better than Logistic Regression with an accuracy of 0.884, precision of 0.842, recall of 0.785, F1-score of 0.812, and AUC-ROC of 0.930. Meanwhile, Logistic Regression obtained an accuracy of 0.871, precision of 0.780, recall of 0.835, F1-score of 0.806, and AUC-ROC of 0.928. These results show that Random Forest is more effective in handling complex relationships in multi-source data for student dropout predictions.

Keywords: Student Dropout Prediction; Machine Learning; Multi-Source Data; Random Forest; Logistic Regression

1. INTRODUCTION

The problem of student dropouts is one of the important issues in the implementation of higher education, which, until now, is still a concern for various academic institutions. The high number of students who are unable to complete their studies on time not only has an impact on the efficiency of the education system but also affects the quality of graduates and the image of the educational institution itself. Various studies show that students' decisions to quit their studies are influenced by various factors, both academic, social, and economic [1]. In addition, the development of information technology, especially the use of Learning Management System (LMS) based learning data, has opened up opportunities to identify student behavior patterns that have the potential to drop out earlier [2]. However, in practice, there are still many universities that have not been able to utilize the data optimally to support the academic decision-making process.

The problem of student dropouts is not only caused by a single factor, but is the result of the interaction of various factors, such as academic, socioeconomic, and student behavior. Previous research has shown that factors such as academic grades, attendance, economic conditions, and the level of student involvement in learning activities have a significant influence on students' decisions to quit their studies [3]. In addition, other studies also confirm that dropouts are a phenomenon that is difficult to predict because they are influenced by various complex and often subjective factors. This shows that conventional approaches in identifying students at risk of dropouts still have limitations, especially in processing large and complex data [4].

Along with the development of technology, Machine Learning-based approaches are starting to be widely used as a solution in predicting the risk of student dropouts. This method is considered to be able to process large amounts of data and identify hidden patterns that are difficult to detect manually. Some studies have shown that the Random Forest algorithm has a pretty good ability to generate accurate predictions in the case of college dropouts [5]. On the other hand, Logistic Regression is also still widely used as a baseline model because of its ability to provide a clear interpretation of the relationships between variables [6]. Therefore, the two methods are relevant to compare in order to gain a more comprehensive understanding of the performance and characteristics of each model in the context of dropout prediction.

Research related to student dropout predictions has been widely conducted in recent years with various approaches. the integration of Machine Learning with Explainable Artificial Intelligence (XAI) is not only able to improve model performance, but also provides an explanation of the factors that affect the risk of dropout [7]. Furthermore, previous research demonstrated that the integration of ensemble learning methods with SMOTE techniques and XAI approaches such as SHAP and LIME, which have been proven to improve model accuracy and transparency [8]. Furthermore, previous research showed that XGBoost can achieve excellent performance in student dropout prediction while maintaining model interpretability trough the use of SHAP explanations [9]. In addition, that Random Forest had a good ability to detect students at risk of dropouts with a high recall rate, as well as being able to identify key factors that influenced them [3]. Futhermore, the application of data imbalance handling techniques, particularly SMOTE, has been demonstrated to significantly enhance the predictive performance of classification models [10]. Meanwhile, the Random Forest model to the transformer-based approach, where the transformer model shows higher performance, but Random Forest still excels in terms of interpretability [11].

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9772
Hal 1075-1085
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Although these studies have made significant contributions, there are still some gaps that need to be studied further. Most studies tend to focus on the use of complex models to improve accuracy, but not many have done systematic comparisons between simple models and complex models in the context of multi-source data. In addition, the balance between model performance and ease of interpretation is still a challenge that has not been fully answered. Therefore, research is needed that is able to examine both aspects at the same time. Based on this description, this study aims to compare the performance of the Logistic Regression and Random Forest algorithms in predicting the risk of student dropout based on multi-source data. This research is expected to provide a clearer picture of the advantages and limitations of each method, as well as produce a model that can be used as a basis for the development of decision support systems in the academic environment.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This research was carried out through several systematic stages, starting from the collection of student datasets based on multi-source data. The data is then processed through a preprocessing stage that includes cleaning, missing value handling, encoding, and normalization. Furthermore, the data was divided into training data and test data with a ratio of 80:20, and class imbalance was handled using SMOTE on the training data. To ensure the stability of the model, the k-fold cross-validation method is used. At the modeling stage, Logistic Regression and Random Forest algorithms were used. The performance of both models was evaluated using accuracy, precision, recall, F1 score, and ROC AUC metrics. The results of the evaluation were then compared to determine the best model in predicting the risk of student dropout, as shown in Figure 1.

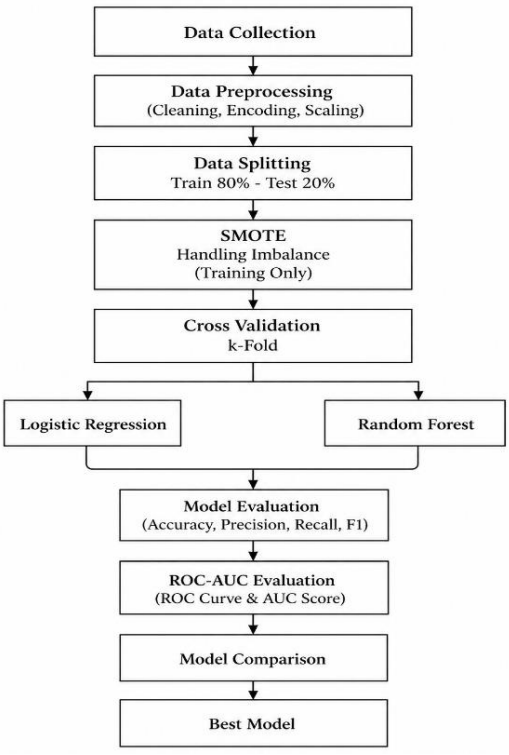

Figure 1. Research Stages

2.2 Data Collection

Data collection in this study used a dataset obtained from Kaggle, namely the Student Dropout and Academic Success Dataset. The dataset consists of 4,424 student data with 34 predictor attributes and 1 target attribute used to represent the academic condition of students. The available data covers various aspects, such as demographic, academic, social economic, and academic administration data so that it can be categorized as multi-source data. See **Table 1** for the attributes used include Marital Status (X_1), Application Mode (X_2), Application Order (X_3), Course (X_4), Daytime/Evening Attendance (X_5), Previous Qualification (X_6) Previous Qualification Grade (X_7), Nationality (X_8), Mother's Qualification (X_9), Father's Qualification (X_{10}) , Mother's Occupation (X_{11}), Father's Occupation (X_{12}), Admission Grade (X_{13}), Dislocated (X_{14}), Educational Special Needs (X_{15}), Debtor (X_{16}), Tuition Fees Up to Date (X_{17}), Gender (X_{18}), Scholarship Holder (X_{19}), Age and Enrollment (X_{20}), International (X_{21}), Curricular Units 1st Sem (Credited) (X_{22}), Curricular Units 1st Sem (Enrolled) (X_{23}), Curricular Units 1st Sem (Evaluations) (X_{24}), Curricular Units 1st Sem (Approved) (X_{25}), Curricular Units 1st Sem (Grade) (X_{26}), Curricular Units 2nd Sem (Credited) (X_{27}),

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9772
Hal 1075-1085
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Curricular Units 2nd Sem (Enrolled) (X_{28}), Curricular Units 2nd Sem (Evaluations) (X_{29}), Curricular Units 2nd Sem (Approved) (X_{30}), Curricular Units 2nd Sem (Grade), (X_{31}) Unemployment Rate (X_{32}), Inflation Rate (X_{33}), and GDP (X_{34}). Meanwhile, the target attributes in the dataset consist of three categories, namely Dropout, Enrolled, and Graduate. Dropout status indicates students who do not continue their studies, Enrolled indicates students who are still actively undergoing lectures, while Graduate indicates students who have successfully completed their studies.

Table 1. Sample Data [12]

No	X_1	X_2	X_3	....	X_{33}	X_{34}	Target
1	1	8	2	....	1.4	1.74	Dropout
2	1	6	11	....	-0.3	0.79	Graduate
3	1	1	5	....	1.4	1.74	Dropout
4	1	8	15	....	-0.8	-3.12	Graduate
5	1	9	1	....	1.4	1.74	Enrolled
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
4424	1	5	15	.....	3.7	-1.7	Graduate

2.3 Data Preprocessing

The data preprocessing stage is carried out to improve the quality of the data before it is used in the modeling process. This process includes cleaning the data from missing or inconsistent values, transforming categorical data into numerical data, and normalizing it to make the data uniform. This stage is important because education data generally comes from various heterogeneous sources, so it requires adjustments so that it can be processed optimally by machine learning algorithms. In addition, preprocessing also plays a role in improving the accuracy and stability of the resulting model [9][12]. In the preprocessing stage, it is carried out with 3 stages, namely cleaning which is used to change the column when there is an error, then encoding is carried out, namely for the conversion of categorical targets to binary (1 is dropout, 0 is non dropout).

2.4 Data Splitting

In this study, the training and testing process was carried out to build and evaluate the ability of machine learning models in predicting student dropout risk. The dataset that has gone through the preprocessing stage is then divided into two parts, namely training data and data testing using the train test split method. The data distribution is carried out in an 80:20 proportion, where 80% of the data is used as data training and the remaining 20% is used as data testing. Data splitting is done by stratified splitting to keep the distribution of dropout and non-dropout classes balanced across both data groups [13][14]. Splitting is done by stratify is y so that the proportion of classes remains balanced in both sets.

2.5 SMOTE

SMOTE is used to address the problem of class imbalance that is common in dropout datasets, where the number of students who do not drop out is much higher than the number of dropouts. This technique works by producing synthetic data in a minority class so that the distribution of data becomes more balanced. The use of SMOTE has been proven to improve the model's ability to detect students at risk of dropouts and reduce bias against the majority class [15]. Data imbalance itself is a major challenge in dropout prediction that can degrade model performance if not handled properly [16].

2.6 Cross Validation

Cross-validation is a model evaluation technique used to test the stability and reliability of the model. A commonly used method is k-fold cross-validation, where the trained data is divided into several parts, then the model is trained and tested alternately on each of these parts. This approach helps to produce more representative evaluations and reduces the risk of overfitting, as the model does not rely on a single data source [17]. In addition, the use of cross-validation is also widely applied in dropout prediction research to ensure that the resulting model has consistent performance [18].

2.7 Logistic Regression

Logistic Regression is a statistical method used to model and predict a binary dependent variable that consists of two possible outcomes. The method estimates the probability of an event occurring based on a set of predictor variables that may influence the outcome. In developing the model, the regression coefficients are estimated using the Maximum Likelihood Estimation (MLE) approach to identify the parameter values that best fit the observed data [19]. Unlike linear regression, which produces continuous numerical outputs, Logistic Regression is specifically designed for classification problems involving dichotomous target variables, such as dropout and non-dropout. Therefore, it is widely applied in predictive studies because it effectively captures the relationship between independent variables and the probability of a particular event. In general, a Logistic Regression model with $p$independent variables can be expressed as follows [20][21]:

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9772
Hal 1075-1085
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0 + \beta_1 X_1 + \beta_2 + \cdots + \beta_p X_p)}} \quad (1)$$

where P(Y = 1) is Probability of students experiencing dropouts, and β_0, β_1,..., β_p are regression coefficients. There is a linear model hidden within the logistic regression model. The natural logarithm of the ratio of P(Y = 1) to (1 P(Y = 1)) gives a linear model in X_i [20][21]:

$$g(x) = \text{In } (\frac{P\ (Y=1)}{1-P(Y-1)}) \quad (2)$$

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...+ \beta_p X_p \quad (3)$$

The function g(x) is defined to satisfy the properties of a linear regression framework. The independent variables are allowed to include a mixture of continuous and categorical attributes. In this study, Logistic Regression is used as a baseline model that functions as a comparison to Random Forest. The use of Logistic Regression aims to evaluate the extent to which linear models that have a high level of interpretability are still able to produce competitive performance in predicting student dropout risk compared to more complex ensemble learning models.

2.8 Random Forest

Random Forest is an ensemble learning method designed to improve the performance of classification models by combining a number of decision trees in one prediction system. This algorithm applies the bootstrap aggregating (bagging) technique, which is to build each tree using a randomly selected data sample from the training dataset as well as a different subset of features in each tree formation process. This approach allows for the creation of diversity between trees so that it can reduce model variance and improve generalization capabilities to new data. The final prediction is obtained through the process of aggregating prediction results from all decision trees that have been built. In the process of forming a decision tree, Random Forest uses an impurity measure to determine the best attributes to use as a separator. The two most commonly used criteria are Gini Index and Entropy. Both measures serve to evaluate the level of homogeneity of data on a node, so that the model can select the separation that results in the most optimal class distribution. The lower the impurity value produced, the better the quality of the data separation performed. The formula for the Gini index is [8]

$$Gini = 1 - \sum_{i=1}^{c} (P_i)^2 \quad (4)$$

Where P_i =Relative Frequency

Unlike the Gini index, Entropy is more mathematical intensive as it contains a logarithmic function. For entropy, the equation is:

$$\text{Entropy} = \sum_{i=1}^{c} - P_i * Log_2(P_i)^2 \quad (5)$$

In this study, Random Forest was applied to predict student dropout risk using multi-source data consisting of academic, demographic, socioeconomic, and academic administration attributes obtained from the Student Dropout and Academic Success Dataset. The input stage uses 34 predictor attributes ($X_1 - X_{34}$), including Admission Grade (X_{13}), Debtor (X_{16}), Tuition Fees Up to Date (X_{17}), Scholarship Holder (X_{19}), Curricular Units 1st Sem (Approved) (X_{25}), Curricular Units 1st Sem (Grade) (X_{26}), Curricular Units 2nd Sem (Approved) (X_{30}), and Curricular Units 2nd Sem (Grade) (X_{31}). Meanwhile, the target variable consists of two classes, namely Dropout ($Y = 1$)and Non-Dropout ($Y = 0$).

At the process stage, Random Forest constructs multiple decision trees using bootstrap sampling, where each tree is trained on a randomly selected subset of training data. Furthermore, a random subset of features is considered at each node split, allowing different trees to learn different patterns from the dataset. The optimal split is selected by minimizing the prediction error within each node, resulting in more homogeneous data partitions. After all decision trees are generated, each tree produces an individual prediction, and the final classification is determined through a majority voting mechanism. The output generated by Random Forest includes student classification into Dropout and Non Dropout categories, dropout probability scores, and feature importance values. Feature importance analysis is useful for identifying the most influential attributes affecting student dropout risk. Previous studies have shown that Random Forest performs well in educational data mining because it can process heterogeneous and multi source datasets effectively while maintaining high prediction performance [22]. **Figure 2** shows the division of the dataset into training data and testing data before the modeling process using Machine Learning algorithms. Of the total 4,424 student data, 3,539 data (80%) were used as training data to train models, while 885 data (20%) were used as data testing to evaluate model performance. In addition, the distribution of target classes in the training data consisted of 2,402 Non Dropout data (class 0) and 1,137 Dropout data (class 1), while in the testing data there were 601 Non Dropout data and 284 Dropout data.

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9772
Hal 1075-1085
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

```
Shape data training : (3539, 34)
Shape data testing  : (885, 34)

Distribusi target di training set:
Target_encoded
0    2402
1    1137
Name: count, dtype: int64

Distribusi target di testing set:
Target_encoded
0    601
1    284
Name: count, dtype: int64
```

Figure 2. Distribution Target Use Random Forest

2.9 Model Evaluation

The model evaluation stage is carried out to measure the extent to which the model is able to accurately predict the risk of student dropout. These evaluations generally use several metrics, such as accuracy, precision, recall, and F1-score, to provide a more comprehensive picture of the model's performance. The use of more than one metric is necessary because dropout datasets are usually unbalanced, so accuracy alone is not enough to objectively represent the performance of the model. Previous research has also emphasized the importance of using various evaluation metrics to obtain more comprehensive results in the classification of dropouts [13].

2.10 ROC AUC Evaluation

In addition to the basic metrics, the evaluation was also conducted using the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) values. ROC describes the relationship between true positive rate and false positive rate at various thresholds, while AUC shows the model's ability to distinguish between dropout and non-dropout classes. An AUC value close to 1 indicates the model's good performance in classification. The use of ROC-AUC is considered important, especially in unbalanced data, as it is able to provide a more stable evaluation than a single metric such as accuracy [16]. In addition, several studies also use ROC AUC as the main indicator in assessing the quality of dropout prediction models [9].

2.11 Model Comparison

The model comparison stage is carried out to determine the method that has the best performance based on the evaluation results that have been obtained. In this study, a comparison was made between Logistic Regression and Random Forest by considering all evaluation metrics, including ROC AUC. This process is important to see the advantages of each model, both in terms of accuracy and generalization capabilities. Some studies have shown that different models can provide varying performance depending on the characteristics of the data, so comparative analysis is needed to determine the most appropriate model [23][24].

3. RESULTS AND DISCUSSION

3.1 Target Class Distribution Before Encoding

The distribution of target classes before the encoding process shows the initial distribution of the amount of data in each category of student status, namely Graduate, Dropout, and Enrolled, which is used to understand the balance of data before modeling. The results of the target class distribution before encoding can be seen in **Figure 3**.

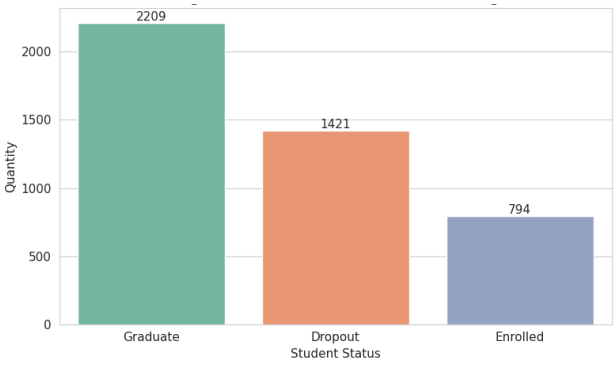

Figure 3. Target Class Distribution Before Encoding

Figure 3 explains that the dataset has three main categories, namely Graduate, Dropout, and Enrolled, with a total of 2209, 1421, and 794 data points, respectively. This distribution shows that the Graduate class dominates the dataset, followed by the Dropout, while the Enrolled class has the least number. This condition indicates a class imbalance, although not too extreme, where the proportion of data is uneven between classes. This imbalance has the potential to

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9772
Hal 1075-1085
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

affect the performance of machine learning models, especially in recognizing minority classes such as Enrolled. Therefore, special handling, such as oversampling techniques (e.g., SMOTE) or adjustment of model evaluation, is needed to keep the prediction results accurate and not biased towards the majority class.

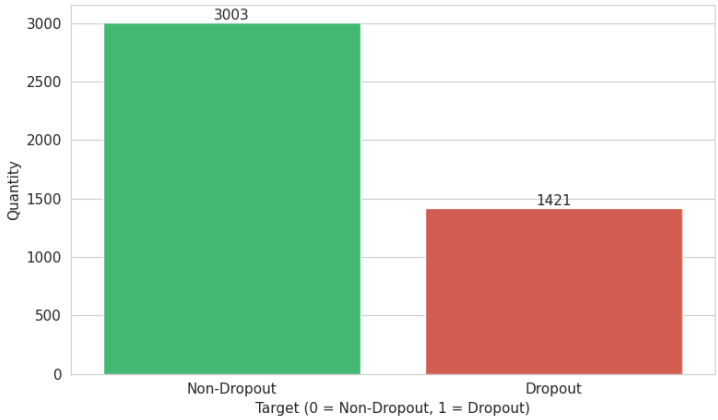

Figure 4. Target Class Distribution After Encoding

Figure 4 is the result of the visualization of the target distribution after the binary encoding process. The dataset used has been transformed from three classes (Graduate, Dropout, and Enrolled) to two classes, namely Non Dropout and Dropout, which aims to simplify the classification problem into binary forms so that it is easier to analyze and model. In Figure 3, the Non-Dropout Class, which is a combination of Graduate and Enrolled, has a total of 3003 data, while the Dropout class has 1421 data.

3.2 Feature Scaling

Feature scaling is the process of normalizing or standardizing feature values to be on the same scale, so that machine learning models can work more optimally and are not biased towards features with greater value. The results of feature scaling can be seen in Table 2.

Table 2. Feature Scaling

	Marital Status	Application Mode	Application Order	Course	Daytime/evening attendance
Count	4424.000	4424.000	4424.000	4424.000	4424.000
Mean	-0.000	-0.000	-0.000	-0.000	0.000
std	1.000	1.000	1.000	1.000	1.000
Min	-0.295	-1.111	-1.315	-2.055	-2.586
25%	-0.295	-1.111	-0.554	-0.900	0.350
50%	-0.295	0.210	-0.554	0.023	0.350
75%	-0.295	0.965	0.207	0.716	0.350
max	7.960	2.097	5.536	1.639	0.350

Table 2 shows that all features, such as Marital status, Application mode, Application order, Course, and Daytime/evening attendance, have been successfully normalized. This is indicated by a mean value close to 0 and a standard deviation (std) of 1 on each feature, which is characteristic of the standardization method (StandardScaler). In addition, the minimum and maximum values indicate that the data has been spread in a more balanced range without the presence of a certain scale dominance of certain features. This condition is important because it ensures that each feature has an equal contribution to the modeling process, so that the machine learning algorithm is not biased towards features with greater value. Thus, the results of this scaling show that the data is ready to be used for the modeling stage with more optimal and stable performance.

3.3 SMOTE

SMOTE is applied only to data training and not to data testing, to prevent data leakage and maintain the validity and objectivity of the model evaluation process. The results of SMOTE can be seen in Figure 5.

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9772
Hal 1075-1085
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

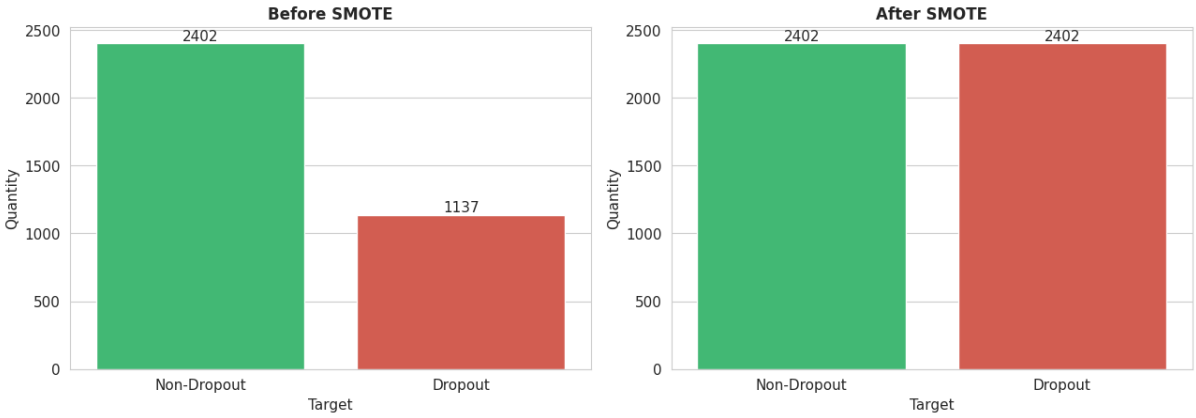

Figure 5. Comparison Class Before and After Smote

Figure 5 shows a comparison of the distribution of classes before and after the implementation of SMOTE. It can be seen that before SMOTE, the dataset had a significant class imbalance, where the number of Non Dropout data was as much as 2402, which was more dominant than the Dropout class, which only amounted to 1137. This condition has the potential to cause the model to tend to be biased towards the majority class and less optimal in recognizing patterns in the minority class. After the implementation of SMOTE, the distribution of classes became balanced, with the number of Non-Dropout and Dropout classes being equal to 2402. This change occurs because SMOTE produces synthetic data on the minority class so that the proportion is equal to the majority class. Thus, the main difference before and after SMOTE lies in the balance of data distribution, where after SMOTE the model has a better chance of studying the characteristics of the two classes in a balanced manner, so that it is expected to improve prediction performance, especially in detecting students at risk of dropout.

3.4 Cross Validation (K-Fold)

K-Fold Cross Validation is used to evaluate model performance more robustly by dividing the data into sections and conducting training and testing alternately. In this study, k = 5 (5-fold cross-validation) was used with stratification, so that each fold has a balanced class distribution according to the original data. The selection of k = 5 is based on the consideration of the balance between evaluation stability and computational efficiency, where a small number of folds is able to reduce the variance of evaluation results without significantly increasing computational time. In addition, 5-fold is also a common practice in machine learning research because it is able to provide performance estimates that are quite representative of the model's capabilities.

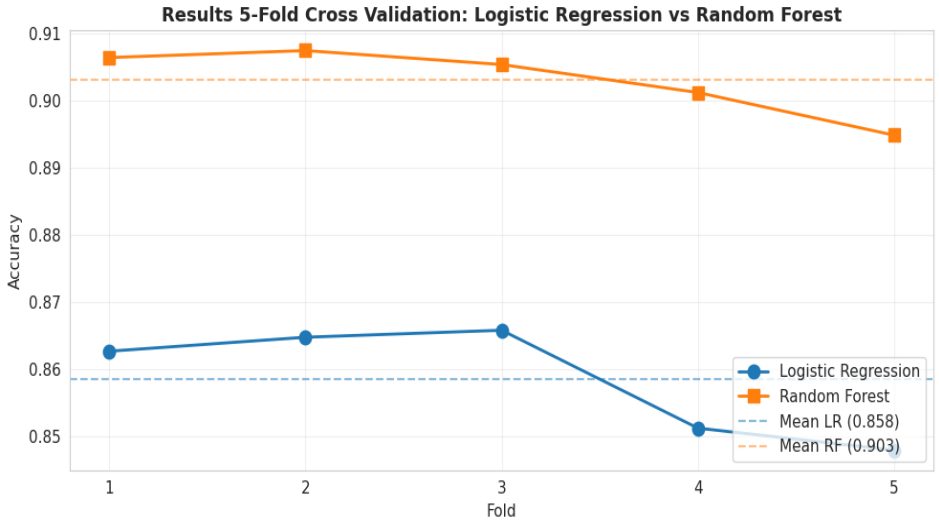

Figure 6. Cross Validation

Figure 6 shows that the Random Forest model consistently has a higher accuracy value than the Logistic Regression on each fold. The accuracy value of Random Forest is in the range of around 0.895 to 0.907, with an average of 0.903, while Logistic Regression is in the range of around 0.851 to 0.866, with an average of 0.858. In addition, Random Forest's performance also tends to be more stable despite a slight decline in the 4th and 5th folds, while Logistic Regression shows a more significant decline in the 4th fold. These results indicate that Random Forest is better able to capture complex data patterns than Logistic Regression, which tends to be linear. Thus, based on the cross-validation

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9772
Hal 1075-1085
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

evaluation, Random Forest showed better and more reliable performance to be used in predicting student dropouts in the dataset used.

3.5 Model Evaluation

Model evaluation aims to assess the extent to which the model is able to produce accurate and generalizable predictions on new data. The evaluation was carried out using several performance metrics, such as accuracy, precision, recall, and F1-score, in order to provide a more comprehensive picture of the model's performance, especially on unbalanced datasets. Here are the results obtained for the Random Forest and Logistic Regression algorithms.

3.5.1 Random Forest

Figure 7 shows the results of the model evaluation conducted using random forest.

```
=== RANDOM FOREST - EVALUATION RESULTS ===
Accuracy  : 0.8836 (88.36%)
Precision : 0.8415
Recall    : 0.7852
F1-Score  : 0.8124

Classification Report:
              precision    recall  f1-score   support

 Non-Dropout       0.90      0.93      0.92       601
     Dropout       0.84      0.79      0.81       284

    accuracy                           0.88       885
   macro avg       0.87      0.86      0.86       885
weighted avg       0.88      0.88      0.88       885
```

Figure 7. Model Evaluation Random Forest

Figure 7 shows the results of the evaluation of the Random Forest model, with an accuracy value of 88.36%, which shows that the model can classify data with a fairly high level of accuracy. A precision value of 0.8415 and a recall of 0.7852 indicate that the model is quite good at identifying students at risk of dropout, although there are still some cases that are not detected (false negatives). The F1-score of 0.8124 indicates a fairly good balance between precision and recall. From the classification report, it can be seen that the model has better performance in the Non-Dropout class with a precision of 0.90 and a recall of 0.93, compared to the Dropout class with a precision of 0.81 and a recall of 0.79. This suggests that models still tend to be more accurate in predicting the majority class. However, overall, the macro average and weighted average values in the range of 0.86 - 0.88 show that the Random Forest model has a fairly stable and reliable performance in predicting student dropouts.

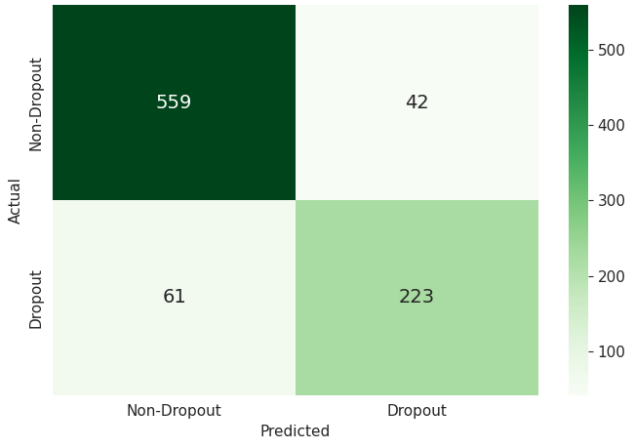

Figure 8. Confusion Matrix Random Forest

Figure 8 shows the results of the confusion matrix in the Random Forest model. It can be seen that the model is able to classify the data quite well. A total of 559 Non Dropout data were correctly predicted (true negative), while 223 Dropout data were also correctly recognized (true positive). However, there are still prediction errors, namely 42 Non-Dropout data that are incorrectly classified as Dropout (false positive) and 61 Dropout data that are not detected or classified as Non-Dropout (false negative). These results show that the model has a better ability to recognize Non-Dropout classes than Dropouts, although the performance in detecting Dropouts is also quite good. Overall, the distribution of values in this confusion matrix supports the results of previous evaluations that Random Forest has a high

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9772
Hal 1075-1085
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

and relatively balanced performance in predicting both classes, although there is still room for improvement, especially in reducing errors in the Dropout class.

3.5.2 Logistic Regression

Figure 9 shows the results of the model evaluation conducted using losgistic regression.

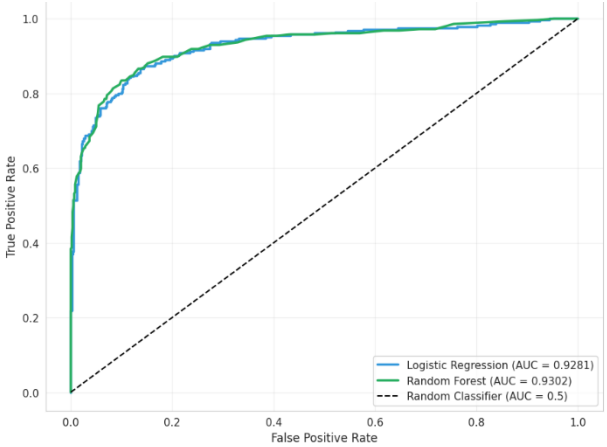

Figure 9. Model Evaluation: Logistic Regression

Figure 9 shows the results of the evaluation of the Logistic Regression model; an accuracy value of 87.12% was obtained, which shows that the model is able to perform classification quite well. A precision value of 0.7796 and a recall of 0.8345 indicate that the model is relatively better at detecting dropout classes (high recall), although there are still errors in positive predictions (lower precision). The F1-score of 0.8061 reflects a fairly good balance between precision and recall. From the classification report, it can be seen that the model has a higher performance in the Non Dropout class with a precision of 0.92 and a recall of 0.89, while in the Dropout class, it obtained a precision of 0.78 and a recall of 0.83, which shows that the model's ability to recognize students at risk is quite good, even though it is not optimal. The macro average and weighted average values in the range of 0.85 – 0.87 show that the model has a fairly stable performance overall, but is still slightly lower than the Random Forest model in terms of prediction accuracy and balance.

```
=== LOGISTIC REGRESSION - EVALUATION RESULTS ===
Accuracy  : 0.8712 (87.12%)
Precision : 0.7796
Recall    : 0.8345
F1-Score  : 0.8061

Classification Report:
              precision    recall  f1-score   support

 Non-Dropout       0.92      0.89      0.90       601
     Dropout       0.78      0.83      0.81       284

    accuracy                           0.87       885
   macro avg       0.85      0.86      0.85       885
weighted avg       0.87      0.87      0.87       885
```

Figure 10. Confusion Matrix Logistic Regression

Figure 10 shows the results of the confusion matrix in the Logistic Regression model. It can be seen that the model is able to classify most of the data quite well. A total of 534 Non-Dropout data were correctly predicted (true negative), while 237 Dropout data were also correctly recognized (true positive). However, there were prediction errors in the form of 67 Non-Dropout data classified as Dropout (false positive) and 47 Dropout data that were not detected or predicted as Non-Dropout (false negative). These results show that Logistic Regression has a fairly good ability to detect dropout classes, even though the number of false negatives is smaller than false positives, which indicates that the model is relatively sensitive to dropout cases. Nonetheless, compared to Random Forest, this model tends to generate more errors in the Non Dropout class, so its overall performance is slightly lower, although it still shows a fairly good balance in the classification of the two classes.

3.6 ROC AUC

The ROC Curve (Receiver Operating Characteristic) is used to visualize the trade-off between the True Positive Rate and the False Positive Rate at various threshold values, while the AUC (Area Under Curve) serves to measure the overall quality of the classification, where the value closer to 1 indicates the better the performance of the model. The results of the ROC AUC can be seen in Figure 11.

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 3, Juni 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v13i3.9772
Hal 1075-1085
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

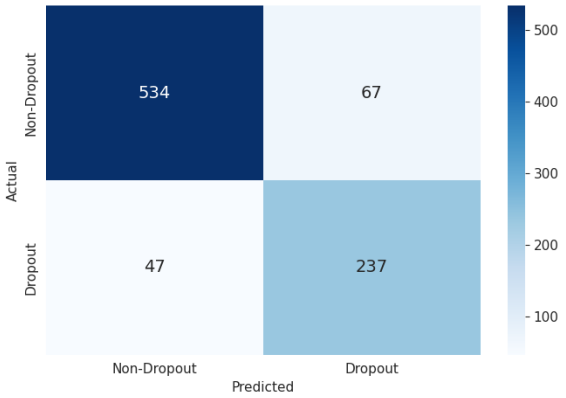

Figure 11. Results ROC AUC

Figure 11 shows that both models, namely Logistic Regression and Random Forest, have excellent classification performance because the curves are both well above the diagonal line (random classifier). The AUC values for Logistic Regression of 0.9281 and Random Forest of 0.9302 indicate that both models can distinguish between Dropout and Non-Dropout classes with a high degree of accuracy. Although the difference is relatively small, Random Forest has a slightly higher AUC value, which indicates better performance in the overall classification. In addition, the curve shape approaching the upper left corner indicates that both models have a combination of a high True Positive Rate and a low False Positive Rate. Overall, these results reinforce previous findings that Random Forest is slightly superior to Logistic Regression in predicting student dropout risk.

3.7 Comparison Visualization

The comparison model is carried out by comparing the performance of the two models based on all evaluation metrics used, so that the most optimal model can be determined in predicting the risk of student dropout. The results of the comparison of Random Forest and Logistic Regression can be seen in Figure 12.

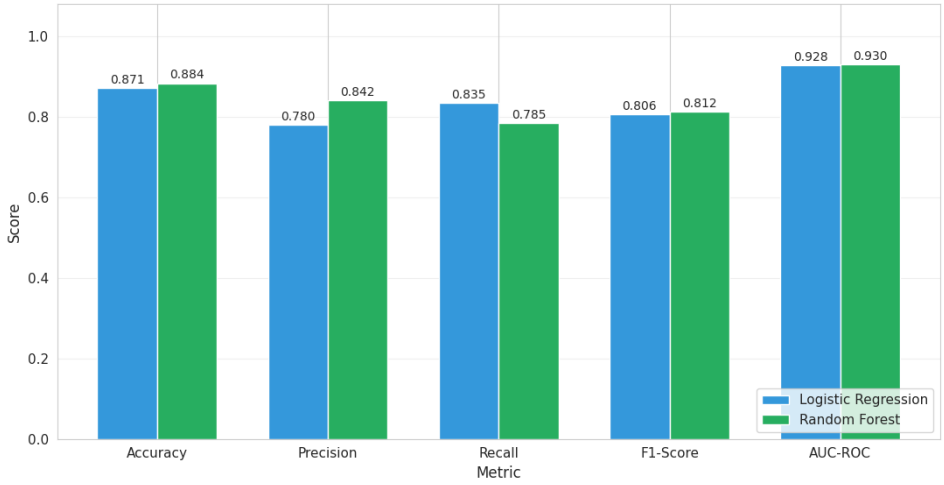

Figure 12. Comparison Visualization

Figure 12 shows the results of the model performance comparison. It can be seen that Random Forest generally performs better than Logistic Regression on almost all evaluation metrics. Random Forest has higher accuracy (0.884), precision (0.842), F1-score (0.812), and AUC-ROC (0.930) than Logistic Regression, which are 0.871, 0.780, 0.806, and 0.928, respectively. This shows that Random Forest is superior in terms of classification accuracy, the ability to accurately identify positive predictions, and the balance between precision and recall. However, Logistic Regression has a slightly higher recall value (0.835 compared to 0.785), which indicates that this model is more sensitive in detecting dropout cases. Overall, Random Forest can be considered a more optimal model because it provides more consistent and superior performance on most evaluation metrics.

4. CONCLUSION

Based on the results of the research that has been conducted, it can be concluded that the application of the machine learning method is able to provide an effective solution in predicting the risk of student dropout by utilizing multi-source data that includes academic, demographic, socioeconomic, and academic administration aspects. The data processing process through the stages of data preprocessing, target transformation, feature scaling, and handling class imbalances

using the SMOTE method has been proven to be able to improve data quality so that the model can learn data patterns more optimally. The results of the evaluation showed that both models, namely Logistic Regression and Random Forest, had a good performance in the classification of student dropout risk. However, Random Forest consistently outperformed most evaluation metrics with an accuracy score of 0.884, precision of 0.842, F1-score of 0.812, and AUC-ROC of 0.930. Meanwhile, Logistic Regression obtained an accuracy score of 0.871, precision of 0.780, F1-score of 0.806, and AUC-ROC of 0.928. On the other hand, Logistic Regression showed a higher recall value, which was 0.835 compared to Random Forest's 0.785, which showed that Logistic Regression was more sensitive in detecting students who were at risk of dropout.

REFERENCES

[1] A. Igualde-Sáez *et al.*, "University Student Dropout: A Longitudinal Dataset of Demographic, Socioeconomic, and Academic Indicators," *Data*, vol. 10, no. 10, pp. 1–12, 2025, doi: 10.3390/data10100162.

[2] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technol. Soc.*, vol. 76, no. December 2023, p. 102474, 2024, doi: 10.1016/j.techsoc.2024.102474.

[3] F. da C. Silva, A. M. Santana, and R. M. Feitosa, "An Investigation Into Dropout Indicators in Secondary Technical Education Using Explainable Artificial Intelligence," *Rev. Iberoam. Tecnol. del Aprendiz.*, vol. 20, pp. 105–114, 2025, doi: 10.1109/RITA.2025.3566095.

[4] A. M. Rabelo and L. E. Zárate, "A model for predicting dropout of higher education students," *Data Sci. Manag.*, vol. 8, no. 1, pp. 72–85, 2025, doi: 10.1016/j.dsm.2024.07.001.

[5] N. T. Vo, Q. N. Le, and H. Le, "EAI Endorsed Transactions Artificial Intelligence-Driven Early Prediction of Student Dropout and Academic Outcomes in Higher Education : A Comparative Study of Advanced Machine Learning Approaches," vol. 13, no. 1, pp. 1–10, 2026, doi: 10.4108/eetinis.131.11758.

[6] M. Segura, J. Mello, and A. Hernández, "Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?," *Mathematics*, vol. 10, no. 18, pp. 1–20, 2022, doi: 10.3390/math10183359.

[7] J. G. C. Krüger, A. de S. Britto, and J. P. Barddal, "An explainable machine learning approach for student dropout prediction," *Expert Syst. Appl.*, vol. 233, no. May, p. 120933, 2023, doi: 10.1016/j.eswa.2023.120933.

[8] F. T. Johora, M. N. Hasan, A. Rajbongshi, M. Ashrafuzzaman, and F. Akter, "An explainable AI-based approach for predicting undergraduate students academic performance," *Array*, vol. 26, no. February, 2025, doi: 10.1016/j.array.2025.100384.

[9] A. Bettahi, F. Z. Belouadha, and H. Harroud, "A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education," *Algorithms*, vol. 18, no. 10, pp. 1–31, 2025, doi: 10.3390/a18100662.

[10] C. H. Cho, Y. W. Yu, and H. G. Kim, "A Study on Dropout Prediction for University Students Using Machine Learning," *Appl. Sci.*, vol. 13, no. 21, 2023, doi: 10.3390/app132112004.

[11] A. Zanellati, S. P. Zingaro, and M. Gabbrielli, "Balancing Performance and Explainability in Academic Dropout Prediction," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 2140–2153, 2024, doi: 10.1109/TLT.2024.3425959.

[12] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting Student Dropout and Academic Success," *Data*, vol. 7, no. 11, 2022, doi: 10.3390/data7110146.

[13] P. D. Atika, "A Comparative Study of Machine Learning-Based Student Dropout Risk Prediction," vol. 14, no. 225, pp. 167–174, 2026, doi: 10.33558/piksel.v14i1.12299.

[14] I. K. Nti and S. Ramanayake, "Explainable Machine Learning for Student Dropout Prediction and Tailored Interventions in Online Personalized Education," 2025, [Online]. Available: https://www.researchsquare.com/article/rs-6615052/v1

[15] N. Mduma, "Data Balancing Techniques for Predicting Student Dropout Using Machine Learning," *Data*, vol. 8, no. 3, pp. 0–14, 2023, doi: 10.3390/data8030049.

[16] G. A. Borges, C. Filipa Dos Santos Pedro, J. Cesar Santos Dos Anjos, A. Rodrigues, F. Boavida, and J. Sá Silva, "A Platform for Early Class Dropout Prediction of University Students," *IEEE Access*, vol. 13, no. July, pp. 109116–109133, 2025, doi: 10.1109/ACCESS.2025.3581751.

[17] J. O. Q. Quispe, O. C. Toledo, M. C. Toledo, E. E. C. Llatasi, and E. M. R. Saira, "Early Prediction of University Student Dropout Using Machine Learning Models," *Nanotechnol. Perceptions*, vol. 20, no. S5, pp. 659–669, 2024, doi: 10.62441/nano-ntp.v20iS5.62.

[18] B. Bennemann, B. Schwartz, J. Giesemann, and W. Lutz, "Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms," *Br. J. Psychiatry*, vol. 220, no. 4, pp. 192–201, 2022, doi: 10.1192/bjp.2022.17.

[19] A. Balboa, A. Cuesta, J. González-Villa, G. Ortiz, and D. Alvear, "Logistic regression vs machine learning to predict evacuation decisions in fire alarm situations," *Saf. Sci.*, vol. 174, no. February, 2024, doi: 10.1016/j.ssci.2024.106485.

[20] I. Kurt, M. Ture, and A. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 366–374, 2008, doi: 10.1016/j.eswa.2006.09.004.

[21] A. S. Aldila *et al.*, "EVALUATING LOGISTIC REGRESSION , SVM , KNN , AND ENSEMBLE MODELS FOR ACCURATE HEART DISEASE RISK PREDICTION," vol. 11, no. 3, pp. 949–957, 2026, doi: 10.33480/jitk.v11i3.6738.EVALUATING.

[22] D. Opazo, S. Moreno, E. Álvarez-Miranda, and J. Pereira, "Analysis of first-year university student dropout through machine learning models: A comparison between universities," *Mathematics*, vol. 9, no. 20, pp. 1–27, 2021, doi: 10.3390/math9202599.

[23] E. Melo, I. Silva, D. G. Costa, C. M. D. Viegas, and T. M. Barros, "On the Use of eXplainable Artificial Intelligence to Evaluate School Dropout," *Educ. Sci.*, vol. 12, no. 12, 2022, doi: 10.3390/educsci12120845.

[24] A. K. Sharma, L. H. Li, and R. Ahmad, *Default Risk Prediction Using Random Forest and XGBoosting Classifier*, vol. 314. Springer International Publishing, 2023. doi: 10.1007/978-3-031-05491-4_10.