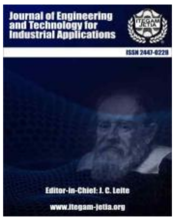
Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.11n.55, p.86-95. September/October, 2025.
DOI: https://doi.org/10.5935/jetia.v11i55.1910


Journal of Engineering and Technology for Industrial Applications
ISSN 2447-0228
Editor-in-Chief: J. C. Leite
www.itegam-jetia.org

RESEARCH ARTICLE OPEN ACCESS

MULTI-STAGE FEATURE SELECTION FOR OPTIMIZING STUDENT DROPOUT PREDICTION

Arif Mudi Priyatno[1], Yunia Ningsih[2], Rizqon Jamil Farhas[3], Fahmi Iqbal Firmananda[4], Resy Kumala Sari[5], and Aryadi[6]

[1,3,4,5,6] Universitas Pahlawan Tuanku Tambusai, Riau, Indonesia
[2] Universitas Trisakti, Jakarta, Indonesia

[1]http://orcid.org/0000-0003-3500-3511, [2]http://orcid.org/0009-0005-6388-1233, [3]http://orcid.org/0009-0002-1092-9596
[4]http://orcid.org/0009-0001-9744-6580, [5]http://orcid.org/0000-0003-3138-3502, [6]http://orcid.org/0009-0009-0495-2264

Email: arifmudi@universitaspahlawan.ac.id, yunia@trisakti.ac.id, rizqonjamil@universitaspahlawan.ac.id, fahmiiqbalfirmananda@universitaspahlawan.ac.id, resikumalasari@universitaspahlawan.ac.id, aryadi@universitaspahlawan.ac.id

ARTICLE INFO

Article History
Received: April 22, 2025
Revised: May 25, 2025
Accepted: September 15, 2025
Published: September 30, 2025

Keywords:
Multi-Stage Feature Selection,
Student Dropout Prediction,
Boruta,
XGBoost,
Educational Data.

ABSTRACT

The high rate of college dropouts is a significant challenge in higher education. Dropout prediction requires an accurate model and is supported by a selection of relevant features. This study proposes a step-by-step feature selection framework to improve prediction accuracy, consisting of three stages, namely Variance Threshold, Mutual Information, and Boruta. The classification model is built using the Extreme Gradient Boosting (XGBoost) algorithm, with evaluation through Stratified 10-fold Cross-Validation. The dataset used includes 4,423 student data that reflects academic, demographic, and socioeconomic information. A total of 18 features were confirmed to be relevant by Boruta. XGBoost models trained on selected features show high performance, with an accuracy of 90.77%, precision of 92.07%, recall of 83.68%, and an F1-score of 87.63%. These results show that the integration of filter and wrapper approaches in the feature selection process effectively improves the performance of the dropout prediction model. This framework is able to filter out important features and produce a more stable and efficient classification model in the context of higher education.

I. INTRODUCTION

The student dropout rate is one of the serious challenges in the world of higher education in various parts of the world, including Indonesia [1], [2], [3], [4]. This phenomenon not only affects an individual's academic achievement, but also poses significant social and economic implications for educational institutions and society at large [5]. When a student decides to drop out of college, the institution loses out on the potential of qualified graduates, while the individual in question may face career barriers or additional economic burdens. Therefore, efforts to identify and predict potential dropouts early on are critical so that appropriate interventions can be provided before such decisions occur.

In recent years, machine learning-based approaches have been widely applied in the field of education to help the decision-making process, including in predicting student dropouts [6], [7], [8], [9]. Accurate predictive models can provide in-depth insights into the risk factors that cause students to drop out of the education system. However, the effectiveness of this model is highly dependent on the quality of the data and features used in the training process [10], [11]. The many attributes in educational data, such as academic, demographic, and behavioral information, often result in a curse of dimensionality problem, which is when the large number of features is not proportional to the amount of data, so that it can reduce the performance of the prediction model.

Another challenge that is often faced in building a dropout prediction model is the existence of irrelevant or redundant features [12], [13]. Features that do not provide significant information about the target variable, or features that are highly correlated with each other, can cause overfitting the model and worsen generalizations in new data [14]. In this context, the feature selection process is crucial as an effort to filter out features that are really relevant and have a significant contribution to the prediction process [15], [16]. By selecting the right features, the data dimension can be optimally reduced, while improving the computational efficiency and accuracy of the model.

Various feature selection methods have been developed, both filters [17], wrappers [18], and embedded [19]. However, most previous studies have relied on only one approach in the feature selection process. For example, the use of Mutual Information is often used to measure the dependencies between features and target variables [20], while Boruta is used to evaluate the importance of features iteratively based on the decision tree [21]. The use of a single method tends to have limitations in comprehensively filtering features, as they do not consider the combination of information value, variability, and feature stability.

To overcome these limitations, this research proposes a multi-stage feature selection approach to improve the accuracy of student dropout predictions. This approach consists of three stages that are systematically designed. In the first stage, initial screening is carried out using the Variance Threshold to remove features with low variability that tend to be uninformative, then followed by Mutual Information to assess the strength of dependency between each feature and the target. This stage aims to eliminate features that are irrelevant in the first place, while maintaining features with high information value. The second stage involves the Boruta method as a Random Forest-based wrapper technique, which is tasked with evaluating the importance of each feature to the target variable in more depth. Boruta has the advantage of detecting important features by considering the interaction between features in an exploratory manner. Thus, this stage is focused on selecting the key features that have the most influence on dropout predictions, while also filtering out features that may have escaped initial selection but do not have a significant contribution in the overall context of the model. After the optimal features are obtained through the two stages of selection, the third stage is carried out by building a prediction model using Extreme Gradient Boosting (XGBClassifier), which is known as one of the high-performance ensemble-based classification algorithms. XGBoost was chosen for its ability to handle large-scale data and effectively handle class imbalances. The model was then evaluated using relevant metrics to measure the accuracy and stability of predictions on student dropout data.

II. MATERIALS AND METHODS

This research proposes a multi-stage feature selection framework that is systematically designed to improve the performance and generalization ability of the prediction model of college dropouts. This approach is formulated to progressively screen features, with the goal of eliminating irrelevant, redundant, or low-variance attributes prior to the model training process. By combining filter- and wrapper-based selection methods, this approach is expected to achieve a balance between computational efficiency, model accuracy, and interpretability of results.

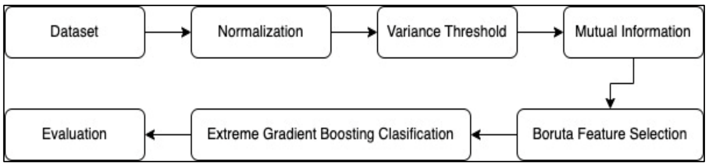


Figure 1: Proposed research framework.
Source: Authors, (2025).

The proposed research framework is presented in Figure 1. The process begins with the acquisition and pre-processing of the dataset, where all features are normalized using standardization techniques to equalize the scale between variables. The first feature selection stage is performed using the Variance Threshold to remove features with very low levels of variation, which generally do not have a significant contribution to the classification process. The remaining features are then evaluated using Mutual Information, a filter-based method that measures statistical dependencies between features and target variables.

In the second stage, the features of the initial selection results were filtered again using the Boruta Feature Selection method, which is a wrapper-based approach built on the Random Forest algorithm. This stage aims to retain the most significant and consistent features in explaining the target variables, while eliminating features that, although they pass the initial selection, have a weak predictive contribution overall.

After obtaining an optimal subset of features, the prediction model was developed using the Extreme Gradient Boosting (XGBoost) algorithm, which is widely known for its high predictive capabilities and its ability to handle high-dimensional and unbalanced data. The model's performance was rigorously evaluated using the Stratified K-Fold Cross-Validation technique, with key metrics used including accuracy, precision, recall, and F1-score. The design of this methodology aims to produce a balanced and reliable assessment of the effectiveness of the proposed approach. Each stage in this framework is described in detail in the following sub-sections.

II.1 DATASET

The dataset used in this study was obtained from the UCI Machine Learning Repository with the title Student Dropout and Academic Success. This dataset contains a total of 4,423 entries, where each entry represents a student with a diverse academic, demographic, and socioeconomic background. The target variables in this dataset are divided into two main categories, namely students who successfully complete their studies (Academic Success) and students who have dropped out of study (Dropout). Of the total available data, 3,003 students were classified as successful completions, while 1,421 students were categorized as dropouts, indicating a moderate class imbalance but still relevant for the development of a classification model [22].

This dataset includes 36 features that represent various aspects that are suspected to affect students' academic success. These features can be grouped into three main categories, namely demographic, academic, and socioeconomic. Demographic characteristics include attributes such as marital status, nationality, gender, age at the time of enrollment, and parental education level. Academic characteristics include students' grades and academic track record, such as entrance scores, the number of courses taken and completed in the first and second semesters, and final semester grades. Meanwhile, socioeconomic features reflect external conditions such as the unemployment rate, inflation rate, and gross domestic product (GDP).

Each attribute in the dataset contributes differently to the analysis process. For example, academic features such as entrance scores and the number of approved courses can directly reflect a student's academic performance. On the other hand, variables such as maternal education or father's occupation provide insights related to the influence of family background on educational outcomes. In addition, macroeconomic features such as unemployment rates and inflation allow for the exploration of the influence of the external environment on the likelihood of students continuing or terminating their studies.

II.2 NORMALIZATION

Before the feature selection process is performed, all numerical attributes in the dataset are normalized using the StandardScaler method. This normalization aims to equalize the scale between features so that each variable has a balanced contribution in the distance calculation process as well as in algorithms that are sensitive to the data scale [23]. This is especially important, especially when features have very different ranges of values, which can lead to dominance by larger-scale features in the model's calculations. The StandardScaler method works by transforming each feature value to have a mean of zero and a standard deviation of one. This transformation is carried out based on Equation 1 [24].

$$x_{new} = \frac{x_i - \mu}{\sigma} \tag{1}$$

Where x_{new} is a new value as a result of transformation. x_i is the original value, μ is an average feature, and σ is a standard deviation from the feature.

II.3 VARIANCE THRESHOLD

The first stage of feature selection was carried out using the Variance Threshold method, which functions to remove features that have very low variability [25]. Features with low variance tend to be constant across the entire data, so they don't provide meaningful information in distinguishing the target class. Therefore, features like this can be eliminated without sacrificing the accuracy of the model. Mathematically, the variance of a feature x against n the sample can be calculated with Equation 2. Where x_i is a feature to-i, $\overline{x}$ is an average feature, $var(x)$ expresses the variance of features.

$$var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2}$$

In this study, the threshold of variance was set at 0.01, which means that only features with a sufficiently high variation were retained for the next stage of analysis. The variance threshold algorithm is shown in Algorithm 1.

Algorithm 1: Algoritma Variance Threshold.

Input	Dataset of normalization results X and variance threshold values
Output	A subset of var(x) features that have a variance value higher than the threshold
Step	1. Calculate the variance of each feature in dataset X using Equation 2.
	2. Compare each variance value with the threshold value.
	3. Select features that have a variance value greater than the threshold.
	4. Create a new subset of var(x) that contains only selected features.

Source: Authors, (2025).

II.4 MUTUAL INFORMATION

After the features with low variance are eliminated, the next stage in the feature selection process is to measure the relevance between each feature and the target variable using the Mutual Information (MI) method [26]. Mutual Information is an effective filter-based feature selection technique to identify both linear and non-linear relationships between free and target variables. Conceptually, Mutual Information measures how much knowledge of feature variable X can reduce uncertainty about target Y. The higher the MI value between a feature and a target, the more information the feature contributes to the target's prediction. The Mutual Information equation between two discrete variables X and Y is defined in Equation 3.

$$MI(X,Y) = \sum_{x\epsilon X}\sum_{y\epsilon Y} p(x,y) \,.\, log\, (\frac{p(x,y)}{p(x)\,.p(y)}) \tag{3}$$

Where $p(x,y)$ is the combined probability between x and y. $p(x)$ and $p(y)$ is the marginal probability of x and y. In this research, the MI value was calculated for each feature against the dropout target. Features with an MI score above the median of the entire MI score are retained for the next selection stage. This approach aims to filter out features that provide significant information to the target, while maintaining a reasonable number of features. The Mutual Information Algorithm is shown in Algorithm 2.

Algorithm 2: Algoritma Mutual Information.

Input	A subset of features of the Variance Threshold (X_{var}) and target variable and
Output	A subset of features X_{mi} with a Mutual Information value above the median value
Step	1. Calculate the value of Mutual Information between each feature $x_i \in X_{var}$ and targets y
	2. Store MI scores in an array $S = \{MI(x1, y), MI(x2, y), \ldots, MI(xn, y)\}$
	3. Calculate the median value of the MI score: median(S).
	4. Select features with a median > MI score
	5. Form a new feature subset X_{mi} for the advanced selection stage.

Source: Authors, (2025).

II.5 BORUTA FEATURE SELECTION

After the features are filtered based on relevance using Mutual Information, the next stage is to apply a wrapper-based advanced feature selection technique using the Boruta algorithm [27]. Boruta is designed to identify all features relevant to the target by comparing the importance of each original feature against the randomly generated shadow features. This method provides a more exploratory and conservative approach than the usual selection method, with a primary focus on maintaining all the features that really matter, not just the best.

Boruta works by building a Random Forest model on the dataset, then assessing the importance of each feature using the feature importance score of Random Forest. To test the significance of the original feature, Boruta created a random copy of the feature (shadow features) and compared the original feature score to the shadow feature score distribution. Only features with consistently higher scores than shadow features are maintained. In general, the main logic of Boruta is that if a feature is statistically more important than a random (shadow) feature, then it is considered relevant and retained. The Boruto feature selection algorithm is shown in Algorithm 3.

Algorithm 3: Algoritma Boruto Feature Selection.

Input	A subset of features of Mutual Information results (X_{mi}), target variable y
Output	A subset of important features X_{boruto} selected by Boruto
Step	1. Add *shadow features* by randomizing the value of each original feature.
	2. Train the Random Forest model on datasets that contain both original features and shadows.
	3. Calculate *feature importance* for all features, both original and shadow.
	4. Compare the *importance score* of each original feature to the maximum score of the shadow feature.
	5. Mark the feature as "important" if the score > the shadow max significantly.
	6. Remove non-essential features and repeat the process until the convergence or iteration limit is reached.
	7. Save features marked as important to be X_{boruto}.

Source: Authors, (2025).

Boruta was chosen in this study because of its excellence in addressing the interaction between features and tolerance to multicollinearity, which is particularly important in the context of complex educational data. The feature selection process at this stage is iterative and lasts until all features can be classified as confirmed, tentative, or rejected. Only features with confirmed status are used for model training at a later stage.

II.6 CROSS-VALIDATION

To ensure the reliability and generalization of prediction models against previously unseen data, this study used the Stratified K-Fold Cross-Validation technique [28]. Cross-validation is a model validation approach that is very important in classification studies, as it allows for a thorough evaluation of model performance by systematically dividing the data into subsets (folds) of training and testing.

Stratified K-Fold was chosen because the characteristics of the target used are unbalanced, where the number of students who drop out is less than those who successfully complete their studies. Stratification ensures that each fold has the same class proportions as in the original distribution, thus avoiding bias and providing a more stable performance estimate.

In its implementation, the dataset is divided into *k* subsets (folds), then the training and testing process is carried out k times. In each iteration, one fold is used as test data, while the *remaining k-1* fold is used for training. The average of all evaluation metrics obtained from each iteration is used as an indicator of the overall performance of the model. The cross-validation algorithm is shown in Algorithm 4.

Algorithm 4: Algoritma Stratified K-Fold Cross-Validation.

Input	Final feature selection dataset X_{boruto}, Target Label Y, Number of Fold *K*
Output	The value of the evaluation metrics (Accuracy, Precision, Recall, F1-Score) of each fold and its average value
Step	1. Divide the data into *k* subsets by maintaining the proportion of classes in each subset (stratification).
	2. For each iteration to-*i* (from 1 to *k*):
	a. Use the i-fold as the test data.
	b. Use the other $k-1$ fold as training data.
	c. Train machine learning models on training data.
	d. Evaluate model performance on test data using classification metrics.
	3. Save the results of the evaluation on each iteration.
	4. Calculate the average value of each evaluation metric from all folds.

Source: Authors, (2025).

This technique was used in the study with k=10, so the model was trained and tested 10 times with different combinations of data. This provides more accurate performance estimates compared to single data sharing, and reduces the risk of overfitting a specific subset of data.

II.7 EXTREME GRADIENT BOOSTING CLASIFICATION

The classification model used in this study is Extreme Gradient Boosting (XGBoost), which is an ensemble algorithm based on gradient boosting decision trees (GBDT) which is known to have high performance and good computational efficiency [29]. XGBoost is designed to optimize the loss function gradually through the addition of sequential decision trees, where each new tree is built to correct prediction errors from the previous tree.

One of the main advantages of XGBoost is its ability to handle high-dimensional data and unbalanced class distribution, which is a common characteristic in the case of predictive college dropouts. In addition, this algorithm supports regularization (L1 and L2) which serves to reduce the risk of overfitting and improve model generalization. XGBoost iteratively minimizes loss functions with the same 4.

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))) + \Omega(f_t) \quad (4)$$

Where l is a loss function (e.g. log-loss), $\hat{y}_i^{(t-1)}$ is a prediction from the previous iteration, f_t is the decision tree that is added on the iteration to-t, and $\Omega(f_t)$ is a regularization function to control the complexity of the tree. The XGBoost model in this study was implemented with default parameters, using the histogram method to speed up the tree pruning process and evaluation based on *logloss metrics* suitable for binary classification.

II.8 EVALUATION

The evaluation of model performance was carried out using several common and representative classification metrics, namely accuracy [30], precision [31], recall [32], and F1-score [24]. Each metric was chosen to capture different aspects of prediction quality, especially in the context of the class imbalances that often occur in student dropout predictions. This approach ensures that the model's performance assessment is not biased against only the majority class. Mathematically, the evaluation metrics are defined in Equations 5-8.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1_{score} = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \quad (8)$$

Where TP (True Positive) is a correctly predicted dropout, TN (True Negative) is a correctly predicted non-dropout, FP (False Positive) is a dropout prediction when it is not, and FN (False Negative) is a prediction of not a dropout when it should be dropout.

In the context of predicting student dropouts, recall is an important metric because it represents the model's ability to accurately detect dropout cases. Meanwhile, precision describes the proportion of dropout predictions that are truly relevant. F1-score provides a balance between precision and recall, and accuracy is used as a general metric that describes the correct proportion of the prediction of the entire data.

III. RESULTS AND DISCUSSIONS

The dataset used in this study is a dataset obtained from the UCI Machine Learning Repository, with a total of 4,423 entries, each representing a student. This data includes academic, demographic, and socioeconomic information that is relevant to analyze the potential for students to drop out of college. The target variables in the dataset are classified binarily, namely 0 for graduates and 1 for dropouts.

Figure 2 shows the distribution of the number of students by target class. From the total data, as many as 3,003 students are classified as graduates, while 1,421 students are classified as dropouts. This distribution shows a significant class imbalance, but not extreme, so evaluation approaches such as Stratified K-Fold Cross Validation are used to ensure that the model remains fairly tested against both classes.

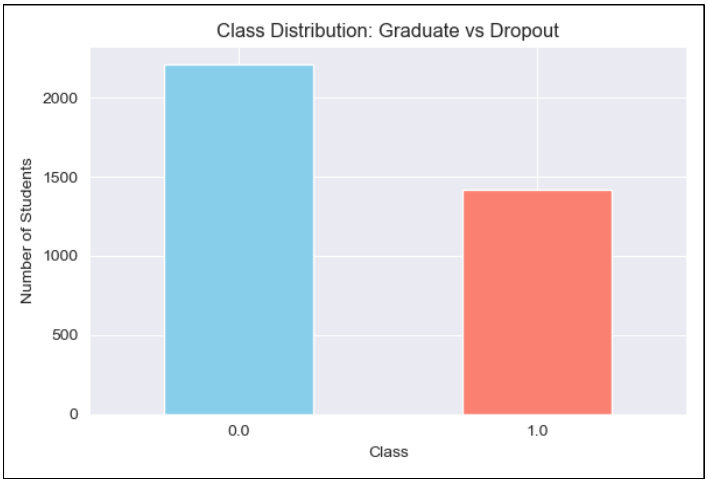


Figure 2: Distribution of classes between graduating students and dropping out of college.
Source: Authors, (2025).

To provide a preliminary overview of the characteristics of the data, Table 1 presents a descriptive statistical summary of some of the selected numerical features that are considered representative. In terms of demographics, students have an average age when registering at 23.46 years, with a standard deviation of 7.83, and an age range between 17 and 70 years. Academic features such as Admission grade have an average score of 127.29, with grade variations from 95 to 190, while the number of courses taken in the first semester reaches 26 courses, although the average is only 6.34. The number of courses approved in the same semester has an average score of 4.79. From the socioeconomic aspect, the unemployment rate at the time of enrollment showed an average score of 11.63, with variations between 7.6 to 16.2. These variables are an important basis for understanding the potential factors that can affect the success or failure of student studies.

Table 1: Initial descriptive statistics of selected numerical features in the dataset.

	Average	**Standard Deviation**	**Minimum Score**	**Maximum Value**
Age at enrollment	23.46	7.83	17	70
Admission grade	127.29	14.61	95	190
Curricular units 1st sem (enrolled)	6.34	2.57	0	26
Curricular units 1st sem (approved)	4.79	3.24	0	26
Unemployment rate	11.63	2.67	7.6	16.2
Inflation rate	1.23	1.38	-0.8	3.7

Source: Authors, (2025).

The normalization process is carried out to equalize the scale of all numerical features before the feature selection and model training stages are implemented. This transformation uses the StandardScaler approach that sets each feature to have a mean of zero and a standard deviation of one. A statistical summary of the five numerical features selected after the normalization process is shown in Table 2.

Table 2: Descriptive statistics of numerical features after normalization.

	Average	Standard Deviation	Minimum Score	Maximum Value
Age at enrollment	0	1	-0.83	5.95
Admission grade	0	1	-2.21	4.29
Curricular units 1st sem (enrolled)	0	1	-2.47	7.65
Curricular units 1st sem (approved)	0	1	-1.48	6.55
Unemployment rate	0	1	-1.51	1.71

Source: Authors, (2025).

Based on the results in Table 2, all the features analyzed show an average value close to zero and a standard deviation close to one. For example, the Admission grade has a minimum score of –2.21 and a maximum of 4.29 after normalization, while the Curricular units 1st sem (enrolled) show a range between –2.47 to 7.65. This process results in a more uniform distribution of values and prevents the dominance of certain features in the calculation of variance as well as in the machine learning process. After the normalization process is carried out, the initial stage of feature selection is applied to filter out attributes that do not contribute significant information to the target variable. The first method used is the Variance Threshold, with a minimum threshold of 0.01. Features that have variance below this value are considered not to provide a significant difference between the data and are eliminated from the subsequent analysis process. Based on the results of elimination, as many as 36 features were successfully maintained from the total initial features.

Features that pass the first stage are then further evaluated using the Mutual Information (MI) method, which aims to measure the degree of statistical dependence between each feature and the target variable. MI is non-linear and does not assume linear relationships, making it suitable for datasets with complex structures like this. In implementation, the MI value is calculated for each feature and features with a score above the median value are retained. A total of 18 features were selected based on the value of Mutual Information, and the ten features with the highest scores were presented in Figure 3. These features generally come from the student's academic domain, such as the number of approved courses and the grades of the first and second semesters. These findings show that academic performance is the main indicator that distinguishes students who complete their studies and those who experience dropouts. Seleksi fitur tahap pertama ini menyaring atribut yang tidak hanya bervariasi secara statistik, tetapi juga terbukti memiliki hubungan informasional yang kuat terhadap target. Fitur-fitur yang lolos kemudian digunakan sebagai input pada tahap seleksi berikutnya menggunakan metode Boruta.

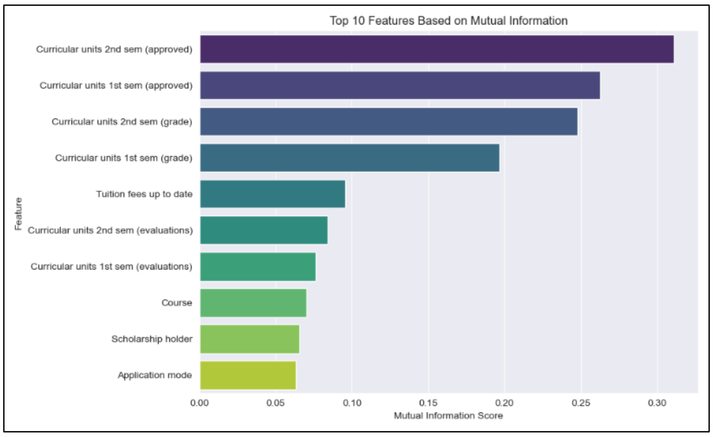


Figure 3: The top ten features are based on the value of Mutual Information against the dropout target variable.

Source: Authors, (2025).

The results of this second stage show that all the evaluated features are classified as confirmed by the Boruta algorithm. A total of 18 features were confirmed as relevant, with no features included in the tentative or rejected categories. Although these results are not common in the practice of Boruta applications, they can be methodologically justified. First, the features that enter the Boruta stage have been strictly screened through the Mutual Information approach, so that only features that have a high information contribution remain.

Second, the number of features tested is relatively limited, which increases the likelihood that each feature will show significant advantages over random features (shadow features).

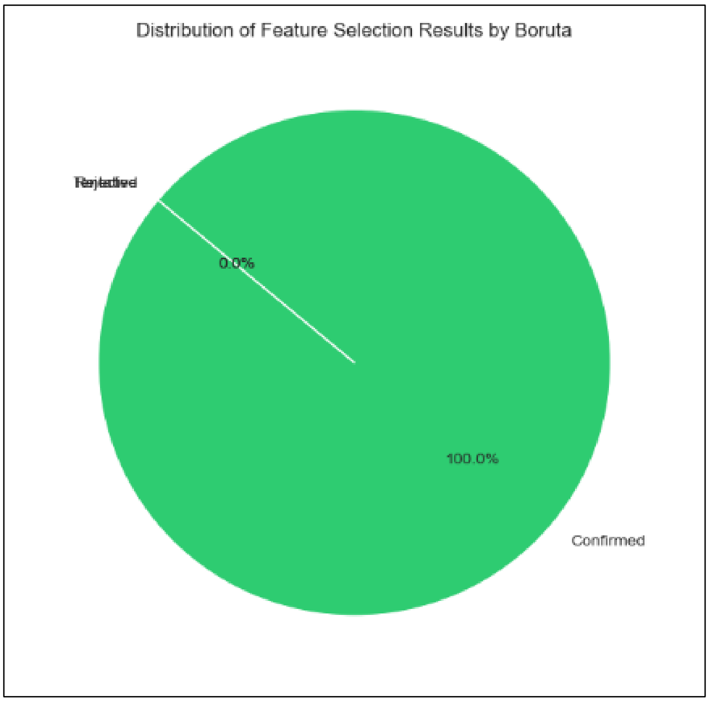


Figure 4: The distribution of feature selection results is based on the Boruta algorithm.
Source: Authors, (2025).

The distribution of feature selection results is shown in Figure 4, which shows that all features are in the confirmed category. Table 3 presents a list of these features, which include administrative indicators (such as Application mode and Course), financial status (Debtor, Tuition fees up to date), scholarship status (Scholarship holder), and various academic and demographic variables. These findings reinforce the assumption that student dropout risk is not only determined by academic performance alone, but also by administrative and socioeconomic conditions.

Table 3: The features are confirmed as relevant by the Boruta algorithm.

Feature	Selected	Tentative
Application mode	TRUE	FALSE
Course	TRUE	FALSE
Previous qualification (grade)	TRUE	FALSE
Admission grade	TRUE	FALSE
Debtor	TRUE	FALSE
Tuition fees up to date	TRUE	FALSE
Gender	TRUE	FALSE
Scholarship holder	TRUE	FALSE
Age at enrollment	TRUE	FALSE
Curricular units 1st sem (enrolled)	TRUE	FALSE
Curricular units 1st sem (evaluations)	TRUE	FALSE
Curricular units 1st sem (approved)	TRUE	FALSE
Curricular units 1st sem (grade)	TRUE	FALSE
Curricular units 2nd sem (credited)	TRUE	FALSE
Curricular units 2nd sem (enrolled)	TRUE	FALSE
Curricular units 2nd sem (evaluations)	TRUE	FALSE
Curricular units 2nd sem (approved)	TRUE	FALSE
Curricular units 2nd sem (grade)	TRUE	FALSE

Source: Authors, (2025).

These results reinforce the effectiveness of the proposed phased feature selection framework. The integration between filter-based methods in the early stages and wrapper-based validation in the advanced stages has been proven to be able to significantly reduce the dimensions of features, while retaining attributes that have a real contribution to the classification process. A subset of features obtained from this stage is then used as input for predictive model training at a later stage.

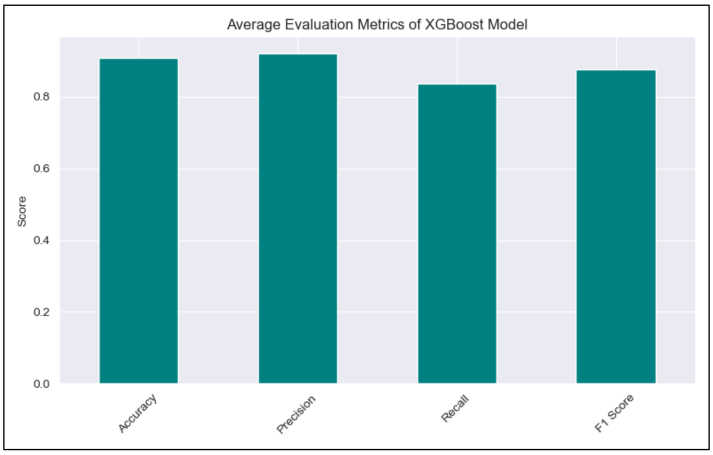


Figure 5: The average score of the XGBoost model evaluation metric is based on cross-validation.
Source: Authors, (2025).

The final classification model in this study was constructed using the Extreme Gradient Boosting (XGBoost) algorithm, with inputs in the form of features that have been selected through a gradual approach based on Variance Threshold, Mutual Information, and Boruta. To evaluate the performance of the generalization of the model, a 10-fold Stratified K-Fold Cross-Validation technique was used, which maintained a balanced proportion of classes in each iteration of training and testing.

Table 4: The average value and standard deviation of the XGBoost evaluation metric.

Metric	Average	Standard Deviation
Accuracy	0.9077	0.0128
Precision	0.9207	0.0185
Recall	0.8368	0.0318
F1 Score	0.8763	0.0186

Source: Authors, (2025).

The results of the model performance evaluation are shown in Figure 5 and summarized in Table 4. Based on an average value of 10 folds, the XGBoost model achieves an accuracy of 90.77%, with a precision of 92.07%, a recall of 83.68%, and an F1-score of 87.63%. Lower-than-precision recall metrics indicate that although the model is highly accurate in identifying dropout students, there are still a number of undetected cases of dropouts (false negatives).

However, a fairly high F1 score indicates a good balance between precision and sensitivity of the model. The stability of the metric reflected in the relatively low standard value of the deviation reinforces the validity of the model evaluation results. With this performance, the XGBoost model shows high effectiveness in classifying potential student dropouts, while showing that the step-by-step feature selection framework used contributes to the achievement of optimal classification results.

Table 5: Comparison with previous research.

Research	Model	Accuracy	Precision	Recall	F1 Score
Proposed	XGB	90.77	92.07	83.68	87.63
[24]	XGB	87.00	88.00	74.00	81.00
[33]	ANN	77.30	76.00	77.30	76.20

Source: Authors, (2025).

The results of the evaluation of this study show that the XGBoost model built through a phased feature selection approach provides the best performance compared to the previous two studies. With an accuracy of 90.77%, precision of 92.07%, and an F1-score of 87.63%, this model shows significant advantages, especially in identifying students who are at high risk of dropping out. Compared to the reseacrh [24] that also used XGBoost on the same dataset, the model in this study resulted in an improvement in all evaluation metrics, specifically the recall which increased from 74% to 83.68%, which is critical for early detection of dropout cases. This difference can be attributed to the gradual feature selection process and consistent use of Stratified K-Fold.

Meanwhile, Research [33] that used a multiclass classification approach with ANN only achieved an accuracy of 77.3% and an F1-score of 76.2%. Although ANN is known to be able to handle data complexity, these results suggest that XGBoost with optimal feature selection can excel in the context of binary dropout prediction. This comparison confirms that the design of predictive models that focus not only on algorithm selection, but also on feature selection and validation strategies, plays a key role in producing precise and reliable systems in higher education.

IV. CONCLUSIONS

This study proposes a multi-stage feature selection framework to improve the accuracy of student dropout predictions in higher education. The proposed approach consists of three main stages: initial screening using Variance Threshold and Mutual Information, validation of relevant features through the Boruta algorithm, and training of classification models using Extreme Gradient Boosting (XGBoost). The results of the experiment showed that this phased approach effectively filtered out uninformative features and retained attributes that had a significant contribution to the prediction target. Boruta successfully confirmed as many as 18 features as relevant, reflecting the importance of academic, demographic, administrative, and socioeconomic factors in determining student dropout risk.

The XGBoost model trained using selected features showed superior performance, with an accuracy of 90.77%, an F1-score of 87.63%, and a precision of 92.07%. Evaluation using 10-fold Stratified Cross-Validation yields stable metrics, indicating a good generalization of new data. The results of this study confirm that the combination of filter-based and wrapper-based feature selection methods can significantly improve classification performance in the context of predicting student dropout risk. For further research, this approach can be further developed by exploring the selection of the most optimal feature normalization method, as the proper normalization process has the potential to significantly affect model performance. Meta-learning strategies can be used to automate the selection of normalization schemes that best suit the characteristics of the educational data, thus supporting the generalization of predictive models in a more adaptive and contextual manner.

V. AUTHOR’S CONTRIBUTION

Conceptualization: Arif Mudi Priyatno, Yunia Ningsih, and Rizqon Jamil Farhas.
Methodology: Arif Mudi Priyatno, Fahmi Iqbal Firmananda, and Resy Kumala Sari.
Investigation: Arif Mudi Priyatno, and Aryadi.
Discussion of results: Arif Mudi Priyatno, Yunia Ningsih, Fahmi Iqbal Firmananda, and Rizqon Jamil Farhas..
Writing – Original Draft: Arif Mudi Priyatno.
Writing – Review and Editing: Arif Mudi Priyatno and Resy Kumala Sari.
Resources: Arif Mudi Priyatno.
Supervision: Arif Mudi Priyatno.
Approval of the final text: Arif Mudi Priyatno, Yunia Ningsih, Rizqon Jamil Farhas, Fahmi Iqbal Firmananda, Resy Kumala Sari, and Aryadi.

VI. REFERENCES

[1] O. Lorenzo-Quiles, S. Galdón-López, and A. Lendínez-Turón, “Factors contributing to university dropout: a review,” Front. Educ., vol. 8, Mar. 2023, doi: 10.3389/feduc.2023.1159864.

[2] R. Goran et al., “Identifying and Understanding Student Dropouts Using Metaheuristic Optimized Classifiers and Explainable Artificial Intelligence Techniques,” IEEE Access, vol. 12, pp. 122377–122400, 2024, doi: 10.1109/ACCESS.2024.3446653.

[3] S. Kim, E. Choi, Y.-K. Jun, and S. Lee, “Student Dropout Prediction for University with High Precision and Recall,” Appl. Sci., vol. 13, no. 10, p. 6275, May 2023, doi: 10.3390/app13106275.

[4] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, “Predicting student’s dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization,” Comput. Educ. Artif. Intell., vol. 3, p. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.

[5] É. Terrin and M. Triventi, “The Effect of School Tracking on Student Achievement and Inequality: A Meta-Analysis,” Rev. Educ. Res., vol. 93, no. 2, pp. 236–274, Apr. 2023, doi: 10.3102/00346543221100850.

[6] P. Rodríguez, A. Villanueva, L. Dombrovskaia, and J. P. Valenzuela, “A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile,” Educ. Inf. Technol., vol. 28, no. 8, pp. 10103–10149, Aug. 2023, doi: 10.1007/s10639-022-11515-5.

[7] M. N. Yakubu and A. M. Abubakar, “Applying machine learning approach to predict students’ performance in higher educational institutions,” Kybernetes, vol. 51, no. 2, pp. 916–934, Feb. 2022, doi: 10.1108/K-12-2020-0865.

[8] B. Albreiki, N. Zaki, and H. Alashwal, “A Systematic Literature Review of Student’ Performance Prediction Using Machine Learning Techniques,” Educ. Sci., vol. 11, no. 9, p. 552, Sep. 2021, doi: 10.3390/educsci11090552.

[9] L. R. Pelima, Y. Sukmana, and Y. Rosmansyah, “Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review,” IEEE Access, vol. 12, pp. 23451–23465, 2024, doi: 10.1109/ACCESS.2024.3361479.

[10] J. Kabathova and M. Drlik, “Towards Predicting Student’s Dropout in University Courses Using Different Machine Learning Techniques,” Appl. Sci., vol. 11, no. 7, p. 3130, Apr. 2021, doi: 10.3390/app11073130.

[11] R. L. S. do Nascimento, R. A. de A. Fagundes, and R. M. C. R. de Souza, “Statistical Learning for Predicting School Dropout in Elementary Education: A Comparative Study,” Ann. Data Sci., vol. 9, no. 4, pp. 801–828, Aug. 2022, doi: 10.1007/s40745-021-00321-4.

[12] H. Nguyen Thi Cam, A. Sarlan, and N. I. Arshad, “A hybrid model integrating recurrent neural networks and the semi-supervised support vector machine for identification of early student dropout risk,” PeerJ Comput. Sci., vol. 10, p. e2572, Nov. 2024, doi: 10.7717/peerj-cs.2572.

[13] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, “Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review,” IEEE Access, vol. 10, pp. 72480–72503, 2022, doi: 10.1109/ACCESS.2022.3188767.

[14] C. Aliferis and G. Simon, “Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI,” 2024, pp. 477–524. doi: 10.1007/978-3-031-39355-6_10.

[15] D. Theng and K. K. Bhoyar, “Feature selection techniques for machine learning: a survey of more than two decades of research,” Knowl. Inf. Syst., vol. 66, no. 3, pp. 1575–1637, Mar. 2024, doi: 10.1007/s10115-023-02010-5.

[16] P. Dhal and C. Azad, A comprehensive survey on feature selection in the various fields of machine learning, vol. 52, no. 4. Applied Intelligence, 2022. doi: 10.1007/s10489-021-02550-9.

[17] A. M. Priyatno and T. Widiyaningtyas, “A SYSTEMATIC LITERATURE REVIEW: RECURSIVE FEATURE ELIMINATION ALGORITHMS,” JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer), vol. 9, no. 2, pp. 196–207, Feb. 2024, doi: 10.33480/jitk.v9i2.5015.

[18] M. Zaffar, M. A. Hashmani, K. S. Savita, and S. A. Khan, “A review on feature selection methods for improving the performance of classification in educational data mining,” Int. J. Inf. Technol. Manag., vol. 20, no. 1/2, p. 110, 2021, doi: 10.1504/IJITM.2021.114161.

[19] D. Effrosynidis and A. Arampatzis, “An evaluation of feature selection methods for environmental data,” Ecol. Inform., vol. 61, p. 101224, Mar. 2021, doi: 10.1016/j.ecoinf.2021.101224.

[20] C. Kuzudisli, B. Bakir-Gungor, N. Bulut, B. Qaqish, and M. Yousef, “Review of feature selection approaches based on grouping of features,” PeerJ, vol. 11, 2023, doi: 10.7717/peerj.15666.

[21] S. M. F. D. Syed Mustapha, “Predictive Analysis of Students’ Learning Performance Using Data Mining Techniques: A Comparative Study of Feature Selection Methods,” Appl. Syst. Innov., vol. 6, no. 5, p. 86, Sep. 2023, doi: 10.3390/asi6050086.

[22] M. V. Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho, “Early Prediction of student’s Performance in Higher Education: A Case Study,” in Trends and Applications in Information Systems and Technologies, 2021, pp. 166–175. doi: 10.1007/978-3-030-72657-7_16.

[23] A. M. Priyatno, L. Ningsih, and M. Noor, “Harnessing Machine Learning for Stock Price Prediction with Random Forest and Simple Moving Average Techniques,” J. Eng. Sci. Appl., vol. 1, no. 1, pp. 1–8, Mar. 2024, doi: 10.69693/jesa.v1i1.1.

[24] A. Ridwan and A. M. Priyatno, “Predict Students’ Dropout and Academic Success with XGBoost,” J. Educ. Comput. Appl., vol. 1, no. 2, pp. 1–8, Dec. 2024, doi: 10.69693/jeca.v1i2.13.

[25] M. Saied, S. Guirguis, and M. Madbouly, “Review of filtering based feature selection for Botnet detection in the Internet of Things,” Artif. Intell. Rev., vol. 58, no. 4, p. 119, Jan. 2025, doi: 10.1007/s10462-025-11113-0.

[26] H. Zhou, X. Wang, and R. Zhu, “Feature selection based on mutual information with correlation coefficient,” Appl. Intell., vol. 52, no. 5, pp. 5457–5474, Mar. 2022, doi: 10.1007/s10489-021-02524-x.

[27] R. Iranzad and X. Liu, “A review of random forest-based feature selection methods for data science education and applications,” Int. J. Data Sci. Anal., Feb. 2024, doi: 10.1007/s41060-024-00509-w.

[28] I. Nurzari, E. Sari, D. I. Harris, A. M. Priyatno, and H. Rusnedy, “Inter-Cluster Distance-Based SMOTE Modification for Enhanced Diabetes Classification,” ITEGAM-J. Eng. Technol. Ind. Appl. (ITEGAM-JETIA, vol. 11, no. 51, pp. 190–196, 2025, doi: 10.5935/jetia.v11i51.1453.

[29] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” Artif. Intell. Rev., vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[30] A. M. Priyatno, “SPAMMER DETECTION BASED ON ACCOUNT, TWEET, AND COMMUNITY ACTIVITY ON TWITTER,” J. Ilmu Komput. dan Inf., vol. 13, no. 2, pp. 97–107, Jul. 2020, doi: 10.21609/jiki.v13i2.871.

[31] A. M. Priyatno and F. I. Firmananda, “N-Gram Feature for Comparison of Machine Learning Methods on Sentiment in Financial News Headlines,” RIGGS J. Artif. Intell. Digit. Bus., vol. 1, no. 1, pp. 01–06, Jul. 2022, doi: 10.31004/riggs.v1i1.4.

[32] M. R. A. Prasetya and A. M. Priyatno, “Dice Similarity and TF-IDF for New Student Admissions Chatbot,” RIGGS J. Artif. Intell. Digit. Bus., vol. 1, no. 1, pp. 13–18, Jul. 2022, doi: 10.31004/riggs.v1i1.5.

[33] S. A. SULAK and N. KOKLU, “Predicting Student Dropout Using Machine Learning Algorithms,” Intell. Methods Eng. Sci., vol. 44, no. 4, pp. 1519–1532, Jan. 2025, doi: 10.58190/imiens.2024.103.