Comparative Analysis of Machine Learning Algorithms with RFE-CV for Student Dropout Prediction

Sekar Gesti Amalia Utami*[1], Haryono Setiadi[2], Arif Rohmadi[3]

[1,2,3]Informatics, Universitas Sebelas Maret, Indonesia

Email: [1]sekargestiau@student.uns.ac.id

Received : May 7, 2025; Revised : Jun 12, 2025; Accepted : Jun 13, 2025; Published : Jun 23, 2025

Abstract

The high dropout rate of students in higher education is a problem faced by educational institutions, impacting quality assessments and accreditation evaluations by BAN-PT. This study aims to develop an early prediction model of potential dropout students using demographic data with a learning analytics approach. Five classification algorithms are used in this research, namely Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), and Support Vector Machine (SVM). The dataset used consists of undergraduate student data of Sebelas Maret University in 2013 (n=2476) which is processed through preprocessing techniques, resampling with *SMOTE*, and validation using *K-Fold Cross-Validation*. The results showed that the RF model gave the best performance with an accuracy of 96.01%, followed by LGBM (95.26%), DT (91.24%), LR (83.68%), and SVM (83.19%). The use of the *Recursive Feature Elimination with Cross-Validation* (RFE-CV) feature selection method was able to improve the efficiency of the model by reducing the number of features without significantly degrading performance. The best feature selection was obtained when using 75% features, which provided an optimal balance between the number of features and model accuracy. The most contributing features include IPS_range (Semester GPA range), parents' income, students' regional origin, as well as several other demographic factors. This study contributes to the development of early warning systems in higher education by providing accurate predictive models and identifying key risk factors.

Keywords : *Educational Data Mining, Machine Learning, RFE-CV, Student Dropout*

1. INTRODUCTION

The high dropout rate of students in higher education is an issue that has become a major concern in a number of countries, including Indonesia [1]. Based on Higher Education Statistics 2022, 4% of the total 9,320,410 students in Indonesia experienced dropout (PDDIKTI, 2022). The National Higher Institution Database (PDDIKTI) (2020) explains that the dropout rate is influenced by various factors such as the inability to fulfill academic requirements, dropping out of college for personal reasons, and the decision to resign. This phenomenon not only impacts individual students by limiting their academic and career opportunities but also affects the assessment of institutional quality, which is a crucial factor in the accreditation evaluation process conducted by the National Accreditation Board for Higher Education (BAN-PT) [2].

Universitas Sebelas Maret (UNS) as one of the public universities in Indonesia also faces challenges related to the high student dropout rate. Data from Smartin Universitas Sebelas Maret shows that until the end of 2023, there are 50,508 registered students who are at risk of dropout, with a percentage reaching 8.4% of the total undergraduate students. This issue requires special attention due to its impact on educational institutions and broader socio-economic aspects. Early identification of students at risk of dropout is an important step to enable timely preventive interventions [3].

Early prediction of potential students at risk of dropout can support strategic decision-making in higher education to improve student retention [4]. Various machine learning approaches have been applied in predicting student dropout with varying degrees of success. The Naive Bayes (NB) algorithm demonstrated classification ability with an accuracy of up to 89% in a study of [5], with advantages in efficient computation and interpretability of results. Meanwhile, XGBoost demonstrated superior performance with an AUC of 0.978 in the study [3], demonstrating its ability to handle complex data. K-Nearest Neighbors (KNN) has also been implemented with an accuracy of 78.20% after the dataset reduction process [6], utilizing its advantages in data adjacency analysis. While these algorithms show promising performance, each has limitations that need to be addressed. NB tends to be less than optimal on data that has high correlation between variables as well as the independence assumption that is not always fulfilled in education data [7]. XGBoost, despite its high accuracy, is often complex to implement and requires intensive parameter customization [8]. KNN is sensitive to data scale and less efficient for high-dimensional datasets commonly encountered in student behavior analysis [9], [10].

In the context of student dropout prediction, algorithms such as RF, DT, and LGBM have specific challenges. RF tends to be computationally intensive with a large number of features from the student's academic data [11]. DT is at risk of overfitting on student data that has varied characteristics [12]. Meanwhile, regression-based algorithms such as LR are difficult to capture complex non-linear patterns in student academic behavior [13]. The most significant challenge is the imbalance of class distribution, where the number of students who drop out is usually much less than the students who stay, so the prediction model tends to be biased. Therefore, a more comprehensive approach is needed to improve the performance of prediction models, one of which is to combine several machine learning algorithms with the RFE-CV feature selection method. RFE-CV offers a solution to these problems through systematic and objective feature selection. This technique works by gradually eliminating irrelevant features and evaluating the model performance at each iteration through cross-validation [14], [15]. By integrating RFE-CV into the five selected machine learning algorithms, this research can overcome the problem of high dimensionality in student data, reduce computational complexity, and improve prediction accuracy. In addition, the implementation of the SMOTE *(*Synthetic Minority Over-sampling Technique*)* technique to address class imbalance will ensure the model is able to predict dropout cases more accurately.

The novelty in this research lies in the implementation of RFE-CV for feature selection optimization which is expected to improve model performance and reduce computational complexity. Unlike previous studies, this research integrates RFE-CV-based feature selection into multiple machine learning models using Indonesian student data to improve predictive accuracy while maintaining computational efficiency. The use of the UNS local dataset provides a specific context that has not been widely explored in previous studies. This approach allows a more comprehensive comparative analysis compared to previous studies that generally only focus on 2-3 algorithms. This research is expected to contribute to the development of an early warning system for student dropout prevention. The resulting predictive model not only helps identify students at risk of dropout, but also provides a deeper understanding of the factors that influence the risk. The findings of this research can serve as a basis for educational institutions in designing more effective interventions to improve student retention, as well as encourage the utilization of technology and data analysis in the decision-making process in higher education.

This research addresses the following questions.

- RQ1: How is the implementation and performance comparison of RF, DT, LR, LGBM, and SVM algorithms in predicting potential student dropouts?
- RQ2: How effective is the use of RFE-CV feature selection in improving the performance of each machine learning algorithm for student dropout prediction?

- RQ3: What are the characteristics and patterns of the features selected by RFE-CV in influencing the accuracy of student dropout prediction?

This study aims to develop a predictive model for student dropout by integrating RFE-CV-based feature selection into five machine learning algorithms using Indonesian higher education data, and to evaluate its effectiveness in both accuracy and feature interpretability.

Various previous studies have utilized DT, RF, LR, LGBM, and SVM algorithms for student dropout prediction, with all related studies summarized in Table 1. Research by Flores et al. (2022)[5] compared several student dropout prediction models using the class balancing method with the SMOTE technique. This research uses RF, Random Tree (RT), J48, REPTree, JRIP, OneR, Bayes Net, and NB algorithms. Research by Niyogisubizo et al. (2022 developed a student dropout prediction model with a two-layer ensemble machine learning approach that utilizes RF, XGBoost, Gradient Boosting (GB), Feed-forward Neural Networks (FNN), and Stacking ensemble algorithms. Research by Jiménez et al. (2023)[16] conducted research on predicting student dropout in universities in Peru by applying RF, DT, Neural Network (NN), and SVM algorithms. Another study by Krüger et al. (2023)[17] focused on using DT, LR, RF, AdaBoost, and XGBoost algorithms in predicting student dropout.

Table 1. Related Studies

No	Author (Year)	Algorithm	Research Result	Limitations
1	Flores et al. (2022) [5]	RF, RT, J48, REPTree, JRIP, OneR, Bayes Net, and NB with SMOTE	RF achieved the highest accuracy (97%).	Feature selection was not implemented
2	Niyogisubizo et al. (2022) [3]	RF, XGBoost, GB, FNN, and Stacking ensemble	The stacking ensemble achieved the highest accuracy (92.18%).	The study utilized a small dataset and did not examine key features
3	Lottering et al. (2020) [6]	DT, SVM, NB, RF, KNN	SVM performed best on the original dataset, while KNN was the best after feature selection (78.20%).	The dataset was restricted to a single institution
4	Jiménez et al. (2023) [16]	RF, DT, NN, SVM	RF demonstrated the best performance (AUC 0.9623).	The study was confined to a specific geographical context
5	Krüger et al. (2023) [17]	DT, LR, RF, AdaBoost, XGBoost	XGBoost achieved the best results with an AUC-PR of 89.5%, Precision of 95%, and Recall of 93%.	Feature selection was not a primary focus of the study

Previous research predominantly uses datasets from developed countries that have education systems with different characteristics compared to Indonesia. Research conducted by Flores et al. (2022)[5] and Jiménez et al. (2023)[16] show that the RF algorithm is able to provide the most accurate dropout prediction results compared to other methods. Jiménez et al. (2023)[16] also identified that academic factors, such as semester taken, as well as socio-economic factors, such as financing methods, have an influence on student dropout. However, they still analyzed the two factors separately and have not explored how the interaction between academic and socio-economic factors can affect dropout risk, especially in the context of developing countries with different educational characteristics and student welfare. In addition, most of the previous studies focused more on comparing model performance without considering computational efficiency through systematic feature selection. The approach used by Niyogisubizo et al. (2022)[3] used ensemble techniques to improve prediction accuracy, but they did not explicitly apply systematic feature selection methods to optimize computational efficiency and reduce model complexity. Therefore, this study overcomes these limitations by applying an RFE-CV-based feature selection method that enables the selection of features that have the most influence on dropout prediction.

2. METHOD

The research flow is illustrated in Figure 1, comprising six main stages: (1) data collection, (2) data preprocessing, (3) feature selection using RFE-CV, (4) K-fold cross-validation, (5) modeling, and (6) evaluation and interpretation. This study employs five algorithms, selected based on their strengths and characteristics in handling educational data. RF and LGBM were chosen due to their ability to manage the complexity of non-linear data. DT was selected for its high interpretability. Meanwhile, LR and SVM were included to compare the performance of linear models with ensemble models in the educational domain.

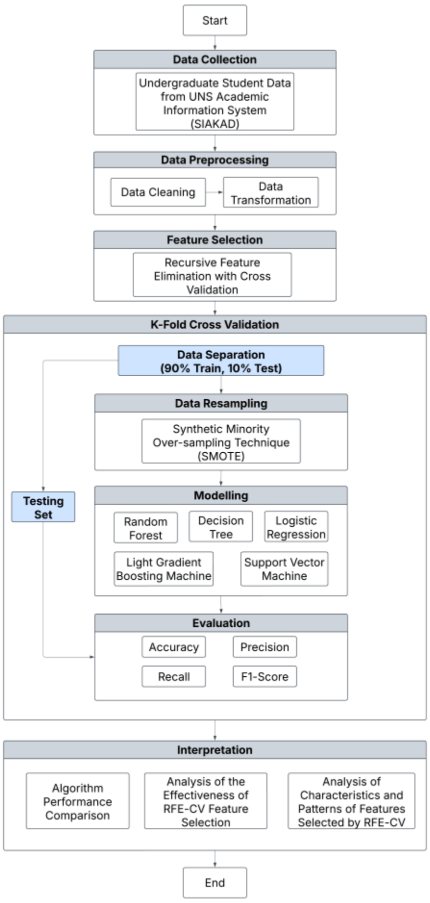
Start
Data Collection
Undergraduate Student Data from UNS Academic Information System (SIAKAD)
Data Preprocessing
Data Cleaning
Data Transformation
Feature Selection
Recursive Feature Elimination with Cross Validation
K-Fold Cross Validation
Data Separation (90% Train, 10% Test)
Data Resampling
Synthetic Minority Over-sampling Technique (SMOTE)
Testing Set
Modelling
Random Forest
Decision Tree
Logistic Regression
Light Gradient Boosting Machine
Support Vector Machine
Evaluation
Accuracy
Precision
Recall
F1-Score
Interpretation
Algorithm Performance Comparison
Analysis of the Effectiveness of RFE-CV Feature Selection
Analysis of Characteristics and Patterns of Features Selected by RFE-CV
End

Figure 1. Research Flowchart

The five algorithms were also selected for their compatibility with RFE-CV, as tree-based models (RF, LGBM, DT) provide feature importance scores, while LR and linear SVM yield coefficients for feature selection. Parameter optimization was performed using random search (for DT, RF, LGBM) and grid search (for LR and SVM) with 5-fold cross-validation, focusing on optimizing the F1-score to address class imbalance in the dataset. Model validation was conducted using 10-fold cross-validation with stratified sampling to ensure consistent class distribution across folds. The dataset was split into 90% for training and 10% for testing, strengthening the validity of the results by providing an independent subset for final evaluation.

2.1 Data Collection

This study utilized data from regular undergraduate (non-transfer) students of Universitas Sebelas Maret from the 2013 cohort, obtained from the academic platform SIAKAD UNS. The dataset consists of 2,476 students, comprising 2,267 non-dropout students and 209 dropout students. The dataset includes various attributes categorized into academic and non-academic features.

2.2 Data Preprocessing

The data preprocessing stage aims to prepare the dataset for subsequent processes. This stage involves two main steps.

- Data Cleaning: Includes the removal of irrelevant columns, handling of missing values, and standardization of data values.
- Data Transformation: Involves data normalization, encoding of categorical variables, and format conversion to ensure compatibility with the machine learning algorithms employed.

2.3 Feature Selection

Following data preprocessing, feature selection was performed using the RFE-CV method to identify the optimal subset of features. The RFE-CV process includes the following steps [14], [18]. The process began by initializing the model using all available features. Next, the relevance score of each feature was calculated, followed by the iterative elimination of the least important features. At each iteration, the model's performance was evaluated using cross-validation to ensure robustness. The optimal number of features was determined based on the configuration that yielded the best performance. RFE-CV was implemented on five algorithms (RF, DT, LGBM, LR, SVM) to analyze the consistency of important features across models. The optimal parameters used for each of the five machine learning algorithms are presented in Table 2.

Table 2. Machine Learning Algorithm Parameters

Algorithm	Parameters	Values
RF	n_estimators	100
	min_samples_split	10
	min_samples_leaf	10
	max_features	sqrt
	max_depth	30
	class_weight	balanced
	bootstrap	False
DT	min_samples_split	5
	min_samples_leaf	5
	max_features	0.6
	max_depth	30
	criterion	gini
	class_weight	balanced
	ccp_alpha	0.01
LGBM	num_leaves	20
	max_depth	5
	learning_rate	0.2
	n_estimators	50
	min_child_samples	10
	subsample	1.0
	colsample_bytree	1.0
	reg_lambda	0
	reg_alpha	1

Algorithm	Parameters	Values
LR	C	0.01
	penalty	l2
	solver	liblinear
	class_weight	balanced
SVM	C	1
	kernel	linear
	class_weight	balanced

The parameter values listed in Table 2 represent the results of the tuning process that provided the best performance based on evaluation scores such as the F1-score during cross-validation. Each algorithm has its own characteristics and important parameters are adjusted. In the Random Forest algorithm, tuning was performed on the number of trees (n_estimators), the maximum depth of the tree (max_depth), and setting the class weight distribution (class_weight). For Logistic Regression and SVM, parameters such as C (regularization), as well as the penalty or kernel, were adjusted to prevent the model from overfitting or underfitting.

2.4 K-Fold Cross-Validation and Data Resampling

To ensure valid evaluation and address class imbalance, the following techniques were applied:

K-Fold Cross-Validation (k=10): This technique involves training and testing the model k times, where each fold is used once as a test set while the remaining k−1 folds serve as the training set [19], [20]. With k=10, the dataset is divided into 10 folds with 9 folds for training and 1 fold for testing alternately.

SMOTE: This technique was used to handle class imbalance [21]. This resampling technique was applied only to the training data in each fold to avoid data leakage, using the following formula [22].

$$Z = X_0 + w(X - X_0) \qquad (1)$$

where w is a random variable in the range [0,1]. This approach ensures that the model can learn patterns in the minority class (dropout) without being biased toward the majority class.

2.5 Modeling

After resampling, the modeling phase was carried out using five machine learning algorithms (RF, DT, LGBM, LR, and SVM) for student dropout prediction. In this stage, the models were trained and evaluated using K-Fold Cross-Validation in each fold to assess their performance.

2.5.1 DT

DT constructs a tree-shaped model where each internal node represents an attribute, branches represent decision rules, and leaf nodes represent predicted class labels [23]. One of the attribute selection methods in DT is Information Gain, based on the concept of Entropy [24], [25]. The equation is as follows [25].

$$Entropy(p_1, \dots, p_n) = -\sum_{i=1}^{n} p_i \, log \, log \, p_i \ (2)$$

where p represents the probability of a certain class label being found. Information Gain is calculated by subtracting the initial entropy from the average entropy after a split by a particular attribute, as shown in the following equation [26].

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \qquad (3)$$

By using this approach, the DT selects the attribute with the highest Information Gain to split the data, thereby forming an optimal decision tree for separating the target classes.

2.5.2 RF

RF is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting [27]. It builds several decision trees using random subsets of features and samples [28], [29]. In the process of model development and performance optimization of RF, several important parameters used include *n_estimators*, *max_depth*, *max_features*, and *bootstrap* [30], [31].

2.5.3 LGBM

LGBM is a gradient boosting framework based on decision trees [32]. It uses a leaf-wise growth strategy instead of the level-wise strategy used in other boosting algorithms, selecting nodes that minimize loss most effectively [33], [34]. The objective function is as follows [35].

$$L(\theta) = \sum_{i=1}^{n} l\big(y_i, f(x_i, \theta)\big) + \Omega(\theta) \qquad (4)$$

The objective function $L(\theta)$ is defined as the function to be minimized based on the model parameters θ, where $l\big(y_i, f(x_i, \theta)\big)$ represents the loss function that measures the difference between the predicted output $f(x_i; \theta)$ and the actual output y_i for each training sample *i*. Additionally, Ω(θ) is the regularization term added to the objective function to prevent overfitting and enhance the model's generalization ability.

2.5.4 LR

LR is a classification algorithm used to estimate the probability that an instance belongs to one of two classes (binary classification) [36], [37]. LR can be used to predict the probability of a new event being classified as positive or negative based on its feature values [36]. Mathematically, the LR model is expressed as follows [38].

$$p(X) = \frac{e^{(B_0+B_1x)}}{1+e^{(B_0+B_1x)}} \qquad (5)$$

In the formula 5, p represents the predicted class probability, B_0 is the constant coefficient, B_1 is the regression coefficient, and X is the independent variable, which refers to the feature used for prediction.

2.5.5 SVM

SVM is a machine learning algorithm used for classification and regression tasks [39]. SVM works by determining an optimal hyperplane that separates data in a high-dimensional feature space, maximizing the margin between different classes [40]. This algorithm transforms the input space into a higher-dimensional feature space using a kernel function, allowing SVM to handle data with complex non-linear relationships [41]. The basic equation for linear SVM classification is expressed as [42].

$$w \,.\, x + b = 0 \qquad (6)$$

In the formula 6, w is the weight vector that defines the direction and slope of the hyperplane, x is the input vector that represents the feature data, and b is the bias term, which adjusts the position of thehyperplane.

For non-linear cases, SVM uses a kernel function to transform the input space, with the general equation [43].

$$f(x_i) = \sum_{n=1}^{N} \alpha_n \cdot y_n \cdot K(x_n, x_i) + b \qquad (7)$$

In this formula 7, f(x) represents the decision function used for classification, α_n is the Lagrange coefficient, y_n is the class label for the data point x_n, with values of +1 or -1, $K(x_n, x_i)$ is the kernel function applied to project the data into a higher-dimensional space (such as Gaussian, polynomial, or radial basis function kernels), and b is the bias term.

2.6 Evaluation and Interpretation

After the modeling process, the next stage is the evaluation of each algorithm used, namely RF, DT, LGBM, LR, SVM using specific evaluation metrics. Model evaluation was conducted using Accuracy [44], Precision [45], Recall [46], and F1-Score [47] which are calculated using the following equations.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad (8)$$

$$Precision = \frac{TP}{TP+FP} \qquad (9)$$

$$Recall = \frac{TP}{TP+FN} \qquad (10)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (11)$$

The confusion matrix is used for more detailed analysis of false positives and false negatives, providing deeper insight into the prediction errors of the model.

The interpretation stage includes the following aspects.

Algorithm Performance Comparison: This evaluation aims to compare the performance of the five algorithms employed in the study.

Effectiveness Analysis of RFE-CV Feature Selection: This analysis seeks to assess the effectiveness of RFE-CV in enhancing model performance and efficiency.

Analysis of Selected Feature Characteristics and Patterns: This exploration is conducted to identify patterns and relationships between the selected features and the model's predictions.

3. RESULT

The implementation steps of the research methodology up to the interpretation results of the existing research. The methodology used includes Data Collection, Data Preprocessing, Feature Selection, Modeling, Evaluation, and Interpretation.

3.1 Data Collection

The dataset collected originates from the SIAKAD academic system at Universitas Sebelas Maret (UNS) for regular undergraduate (S1) students of the 2013 cohort (non-transfer students). A total of 2,476 raw data entries were obtained, containing 65 student biodata features (e.g., GPA, province code, gender, housing status, and others), along with 1 target feature with two classes: "NOT DROP OUT" and "DROP OUT". Several identified attributes of the dataset are presented in Table 3.

Table 3. Description of Student Data

Code	Attribute	Values	Description
jenis_kelamin	Student Gender	[0, 1]	0: Male 1: Female

Code	Attribute	Values	Description
status_rumah	Student Housing Status	[1, 2, 3, 4, 5]	1: Parent's House 2: Relative's House 3: Dormitory/Boarding 4: Own House 5: Others
kewarganegaraan	Student Nationality	[1, 2, 3]	1: Native Indonesian 2: Indonesian Descent 3: Foreign National
penguasaan_teks_asing	Foreign Text Comprehension	[1, 2, 3, 4, 5]	1: Fully Understand 2: Easy to Understand 3: Fairly Understand 4: Somewhat Difficult 5: Not Understand at All
…	…	…	…
STATUS DROP OUT	Student Dropout Status	[0, 1]	0: NOT DROP OUT 1: DROP OUT

In this study, dropout status is defined based on the "STATUS DROP OUT" attribute in the dataset. A student is considered to have dropped out if they are labeled "DROP OUT", while students who are still active or have graduated are grouped under "NOT DROP OUT". Dropout status is specifically defined as students who have been registered for more than seven years without a recorded graduation date. Based on the dataset, the number of non-dropout students significantly exceeds the number of dropout students, with 2,267 and 209 individuals respectively.

3.2 Data Preprocessing

Following data collection, the Data Preprocessing stage was conducted, consisting of two main steps.

3.2.1 Data Cleaning

The data cleaning process was carried out to ensure the quality of the dataset prior to analysis. The steps involved in this process include the removal of irrelevant columns, handling of missing values, and standardization of values across certain features. After the cleaning process, only 35 out of the original 66 attributes were retained, as shown in Table 4.

Table 4. List of Relevant Features

Relevant Features				
mhs_provinsi	mhs_kabupaten	jenis_kelamin	status_rumah	kewarganegaraan
penguasaan_teks_asing	agama	status_marital	hobi	riwayat_pdk_ayah
riwayat_pdk_ibu	pekerjaan_ayah	pekerjaan_ibu	jalur_masuk	rata_un
beasiswa	jurusan_smta	sumber_biaya	wali_provinsi	wali_kabupaten
prestasi	pemberi_beasiswa	penghasilan_ibu_range	penghasilan_ayah_range	penyakit_diderita
kegiatan_mahasiswa	fakultas	nama_prodi_asli	IPS1	IPS2
IPS3	IPS4	total_kredit_capai_smt4	gapyear	STATUS DROP OUT

The selected features are those deemed to have significant potential in contributing to the prediction of student dropout status. These include demographic variables such as jenis_kelamin (*gender)*, status_rumah (*housing status)*, and kewarganegaraan (*citizenship)*, educational and socio-economic background variables such as riwayat_pdk_ayah (*father's educational history)*, pekerjaan_ibu (*mother's occupation)*, and penghasilan_ayah_range (*father's income range)*, and academic performance indicators including IPS1 – IPS4 (*GPA1 through GPA4)*, total_kredit_smt4 (*total credits earned by the fourth semester)*, and *gap year* status. Furthermore, the number of data entries was reduced from 2,476 to 2,473.

3.2.2 Data Transformation

The data transformation process included the elimination of duplicate entries, handling of missing values, and normalization of data to ensure consistency in feature representation. After this process, the number of available records was reduced from 2,473 to 2,463 entries, all of which were free of missing values. One of the results of this process is shown in Table 5.

Table 5. Data Transformation Results Based On Specific Ranges

Feature	Values	Description
Semester GPA (1–4)	{1, 2, 3, 4, 5}	1 = 0–1.50 2 = 1.51–2.00 3 = 2.01–2.50 4 = 2.51–3.50 5 = 3.51–4.00
National Exam Average Score	{1, 2, 3, 4, 5}	1 = 0–4.50 2 = 4.51–5.50 3 = 5.51–6.50 4 = 6.51–7.50 5 = 7.51–10.0
Total Credits up to Semester 4	{1, 2, 3, 4, 5}	1 = 0–20 credits 2 = 21–40 credits 3 = 41–60 credits 4 = 61–80 credits 5 = 81–96 credits

This data transformation aims to simplify numerical data into categorical forms that are more easily interpreted by predictive models. After this process, the original columns can be removed to avoid data redundancy.

3.3 Feature Selection with RFE-CV

Following the data preprocessing stage, the next step involves feature selection to identify the most relevant and informative subset of features for predicting student dropout using RFE-CV. The process begins by utilizing all available features, totaling 34. In each iteration, the model evaluates the importance of each feature using 5-fold cross-validation with the F1-score as the evaluation metric. The least contributive feature is eliminated iteratively until the remaining features yield the highest accuracy. The RFE-CV procedure is conducted using five algorithms: RF, DT, LGBM, LR, and SVM. The results of the RFE-CV feature selection for each algorithm are presented in Table 6.

Table 6. Comparison Of Feature Selection Results Using RFE-CV

Algorithm	Number of Features	Top 5 Features	Bottom 3 Features
RF	30	IPS_range, mhs_provinsi, fakultas	gapyear, sumber_biaya, status_marital
DT	20	jurusan_smta, IPS_range, beasiswa	hobi, agama, penguasaan_teks_asing
LGBM	29	sumber_biaya, IPS_range, nama_prodi	gapyear, status_marital, kewarganegaraan
LR	34	[all features]	[none]
SVM	23	mhs_provinsi, IPS_range, gapyear	jalur_masuk, hobi, nama_prodi

The results in Table 6 indicate that the *IPS_range* (Semester GPA range) feature consistently appears as an important feature across all algorithms, underscoring the critical role of academic performance in predicting dropout risk. Demographic features such as *mhs_provinsi* (student's province of origin) and financial features such as *sumber_biaya* (source of tuition funding) also emerge as strong predictors in several algorithms, suggesting the influence of geographic background and financial support. Additionally, the *gapyear* (study gap year) feature frequently appears among the top features, potentially reflecting the impact of a study gap prior to university enrollment on dropout risk. Conversely, features such as *hobi* (hobbies), *status_marital* (marital status), and *jalur_masuk* (admission path) tend to be less influential, as evidenced by their appearance among the bottom three features in several algorithms. Overall, these findings highlight the predominance of academic and financial factors over personal characteristics in predicting student dropout. The variation in important features across algorithms also suggests the potential benefit of model ensemble approaches for more comprehensive insights. The consistent appearance of *IPS_range* (Semester GPA range) across all algorithms reinforces the importance of supporting academic performance to reduce dropout risk.

3.4 Model Performance Comparison

In the modeling stage, each machine learning model (RF, DT, LGBM, LR, and SVM) was trained using the training set and tested on the test set, utilizing the full set of 34 preprocessed features. The modeling process employed 10-fold cross-validation. Following model training, the five machine learning models were evaluated using standard performance metrics, including Accuracy, Precision, Recall, and F1-Score. The performance results using the complete feature set are presented in Table 7.

Table 7. Model Performance Results With All Feature Set

Model	Accuracy	Precision	Recall	F1-Score
RF	**96.01%**	**97.80%**	**94.13%**	**95.93%**
DT	91.24%	89.17%	93.89%	91.47%
LGBM	95.26%	95.92%	94.55%	95.23%
LR	83.68%	83.03%	84.67%	83.84%
SVM	83.19%	81.70%	85.54%	83.57%

Table 7 presents the comprehensive performance results of all five models. Overall, ensemble-based algorithms, particularly RF and LGBM, consistently emerged as the top performers across all metrics, highlighting the superior capability of ensemble methods in this prediction task. Specifically, the results in Table 7 highlight the superiority of ensemble-based algorithms, with RF achieving the highest F1-Score at 95.93%, followed closely by LGBM at 95.23%. These ensemble methods

DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4695

consistently outperform non-ensemble approaches such as DT with 91.47%, as well as regression-based models like LR at 83.84% and SVM at 83.57%. The strength of Random Forest lies in its ability to handle data complexity and reduce overfitting through the bootstrap aggregating process. Meanwhile, LGBM comes close to RF's performance due to its efficiency in handling large-scale data using optimized gradient boosting techniques. The performance results of each model are also visualized in the chart shown in Figure 2.

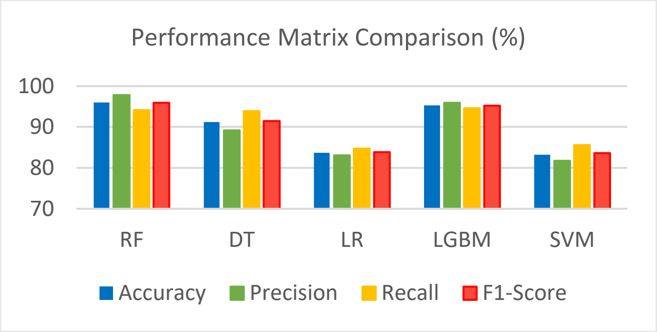

Figure 2. Model Performance Chart with All Features

Based on Figure 2, the RF model achieved the highest accuracy, followed by LGBM, indicating that ensemble models tend to be more accurate in capturing patterns from the complete set of features used. In contrast, SVM and LR rank the lowest, suggesting their limitations in handling non-linear relationships and complex data. These findings indicate that in the context of the dropout prediction problem, which involves complex and non-linear patterns, ensemble approaches are more effective than linear models. In addition, the execution times of each algorithm are compared in Figure 3.

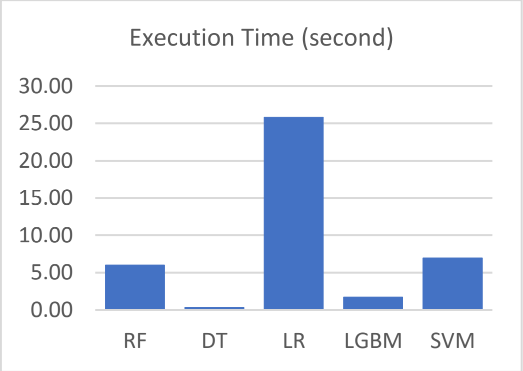

Figure 3. Model Execution Time Chart with All Features

As shown in Figure 3, DT shows the fastest execution time at 0.35 seconds, indicating good computational efficiency for real-time scenarios or devices with limited resources. LGBM follows with a relatively efficient time of 1.74 seconds, despite being an ensemble-based method. On the other hand, LR recorded the highest execution time at 25.82 seconds, most likely due to the complexity of optimization calculations on high-dimensional data. SVM and RF also showed relatively high execution times of 6.95 seconds and 6.02 seconds respectively, which may be influenced by the model structures and parameters used. Model evaluation was further extended by utilizing confusion matrices to gain deeper insights into false positive and false negative rates. The confusion matrix of the best-performing algorithm (RF) is presented in Figure 4.

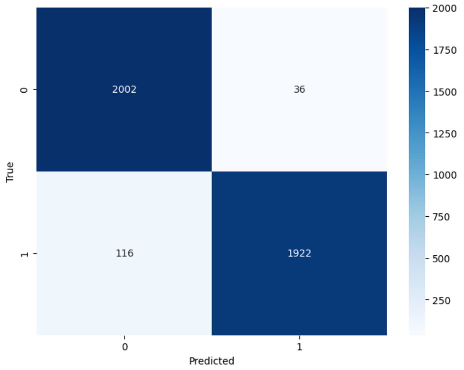

Figure 4. Confusion Matrix RF

Based on Figure 4, the confusion matrix for the RF model shows a strong performance, recording 1,922 true positives and 2,002 true negatives, with only 36 false positives and 116 false negatives. In addition, the confusion matrix for the lowest-performing algorithm (LR) is shown in Figure 5.

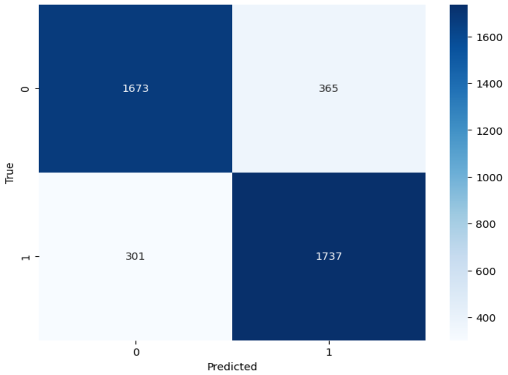

Figure 5. Confusion Matrix LR

As shown in Figure 5, LR produced 1,737 true positives and 1,673 true negatives, with a significantly higher number of false positives (365). This substantial discrepancy in false positive rates suggests that RF is more effective in accurately identifying students at genuine risk of dropping out.

3.5 Analysis of the Effectiveness of RFE-CV Feature Selection

In the modeling stage, each machine learning model was evaluated using the selected features derived from RFE-CV feature selection across five different algorithms: RF, DT, LGBM, LR, SVM. Additionally, evaluations were conducted using subsets of the selected features, specifically 25%, 50%, and 75% of the features obtained through RFE-CV for each algorithm. The best performance using RFE-CV (achieved by RF) is presented in Table 8, with feature subsets of 25% (7 features), 50% (15 features), 75% (22 features), and 100% (30 features).

Table 8. Model Performance Results Using Selected Features from RFE-CV with RF

Model	Feature Percentage	Accuracy	Precision	Recall	F1-Score
RF	25%	88.47%	87.12%	90.29%	88.67%
	50%	95.22%	95.88%	**94.51%**	95.19%
	75%	95.81%	97.16%	94.37%	95.74%
	100%	**95.96%**	**97.64%**	94.20%	**95.89%**
DT	25%	87.86%	86.82%	89.29%	91.20%
	50%	90.99%	88.57%	**94.13%**	88.03%
	75%	**91.04%**	**88.79%**	93.94%	91.26%
	100%	90.95%	**88.79%**	93.82%	**91.29%**

Model	Feature Percentage	Accuracy	Precision	Recall	F1-Score
LGBM	25%	87.98%	86.31%	90.28%	88.25%
	50%	93.97%	93.38%	**94.64%**	94.01%
	75%	94.89%	95.11%	**94.64%**	94.87%
	100%	**95.30%**	**95.99%**	94.54%	**95.26%**
LR	25%	73.50%	69.48%	83.89%	75.99%
	50%	78.91%	77.94%	80.63%	79.26%
	75%	81.47%	81.17%	81.93%	81.55%
	100%	**82.99%**	**82.45%**	**83.82%**	**83.13%**
SVM	25%	76.55%	73.04%	84.24%	78.22%
	50%	79.51%	78.04%	82.13%	80.03%
	75%	81.30%	80.84%	82.06%	81.44%
	100%	**83.09%**	**81.60%**	**85.46%**	**83.48%**

The RF model achieved optimal performance with the highest F1-Score of 95.89% using 100% of the features. However, the most significant performance improvement occurred when the number of features increased from 25% to 50%, with a gain of 6.52% (from 88.67% to 95.19%). In contrast, increasing the features from 75% to 100% resulted in only a 0.15% improvement. A similar trend is observed in the LGBM model, where the difference between using 75% (94.87%) and 100% (95.26%) of the features is relatively small. This indicates that the last 25% of features contribute weakly to predictive power and are likely redundant. Most of these features relate to demographic information such as hobbies and marital status. On the other hand, the DT model exhibits relatively stable performance across all feature subsets, while LR and SVM display a more linear trend of improvement with increasing feature size, suggesting that linear models require a more comprehensive representation of features.

The implementation of RFE-CV and its varying impacts on algorithm performance are presented in Figure 6.

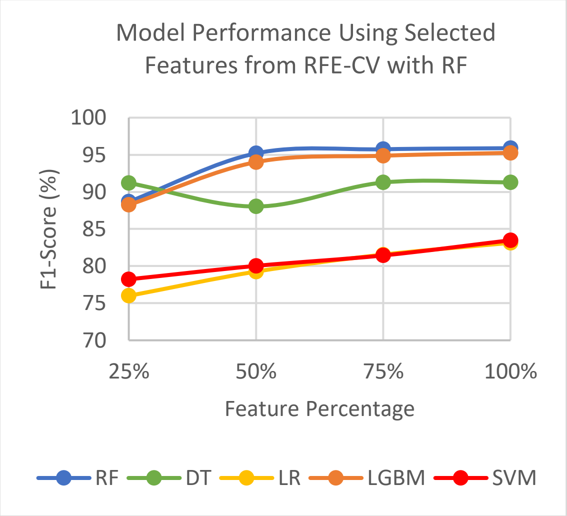

Figure 6. Model Performance Using Selected Features from RFE-CV with RF

The graph in Figure 6 shows a consistent trend that ensemble models such as RF and LGBM outperform all variations of feature subsets. This strengthens the evidence that these two algorithms are not only accurate but also stable, even with a reduction in features. Meanwhile, the performance of the LR and SVM models gradually improves as the number of features increases, but they are unable to match the performance of the ensemble models. In addition, the execution times of each algorithm using selected features of RFE-CV RF are compared in Figure 7.

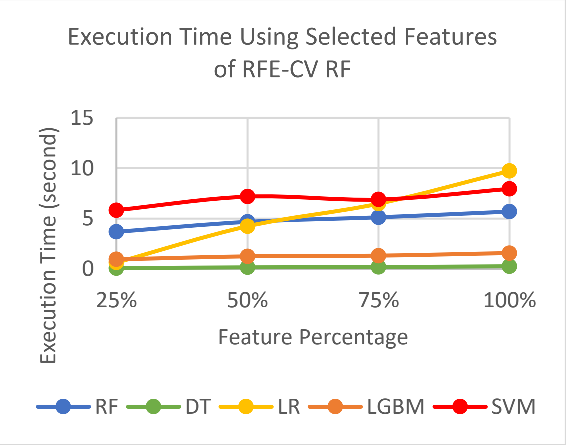

Figure 7. Execution Time Graph Using Selected Features from RFE-CV with RF

Based on Figure 6 and Figure 7, the use of 75% selected features in the RF model results in an F1-Score of 95.74%, only 0.19% lower than the model using all features, while improving computational efficiency by 14.8%. LGBM demonstrates a similar pattern. These findings suggest a trade-off between model complexity and predictive performance, where using the top 75% of features can maintain near-optimal accuracy while enhancing computational efficiency. This phenomenon aligns with the parsimony principle in machine learning, which emphasizes that simpler models with comparable accuracy are preferable due to their better generalization to new data.

3.6 Analysis of Characteristics and Patterns of RFE-CV Selected Features

Each algorithm ranked features based on their contribution to the prediction model. Features that consistently appear across multiple algorithms demonstrate higher significance in the context of student dropout prediction. The important features identified by all algorithms are presented in Table 9.

Table 9. List Of Important Features Identified By All Algorithms

Feature	RF	DT	LGBM	LR	SVM
IPS1_range	V	V	V	V	V
IPS2_range	V	V	V	V	V
IPS4_range	V		V	V	V
fakultas	V		V	V	
gapyear		V		V	V
IPS3_range	V		V	V	
mhs_provinsi	V			V	V
nama_prodi_asli	V		V	V	
beasiswa		V			
jurusan_smta		V			
sumber_biaya			V		

Based on Table 9, several key features were consistently identified by multiple algorithms, particularly the *IPS_range* (Semester GPA range) variables (IPS1_range, IPS2_range, IPS3_range, IPS4_range), *mhs_provinsi* (student's province of origin), and *sumber_biaya* (source of tuition funding). These features can be categorized into three main dropout risk predictors: (1) academic indicators, (2) economic factors, and (3) demographic factors. Conversely, features such as *marital_status*, *hobbies*, and *religion* were frequently eliminated during the RFE-CV selection process, indicating their low relevance to dropout risk in the context of UNS. This contrasts with findings from European universities [48], where students' social engagement significantly influenced dropout risk. Such differences are likely attributable to cultural characteristics and higher education systems in Indonesia, which place a greater emphasis on formal academic achievement rather than social and personal aspects of student life.

4. DISCUSSIONS

The superior performance of the RF algorithm, which achieved an accuracy of 96.01% and an F1-score of 95.93%, demonstrates its ability to capture non-linear patterns between academic factors such as *IPS_range* (Semester GPA range) and socio-economic variables such as *sumber_biaya* (source of tuition funding), both of which significantly contribute to dropout prediction. Moreover, the confusion matrix analysis shows a tendency for false negatives, where at-risk students are misclassified as non-dropouts. These misclassifications are predominantly found among students who initially demonstrated stable academic performance during the first and second semesters but experienced a decline in the third and fourth semesters. This finding highlights the importance of predictive approaches that are not only static but also account for the longitudinal dynamics of academic performance. The consistent identification of *IPS_range* (Semester GPA range) across all algorithms supports the academic integration theory in student retention models [49], wherein academic achievement serves as the primary foundation for educational continuity. Conversely, the emergence of *mhs_provinsi* (student's province of origin) as a significant feature opens a new line of discussion regarding the role of geographical background in influencing student adaptation, which has not been widely addressed in the context of higher education in Indonesia.

From a practical standpoint, these findings provide clear guidance for the development of student retention policies at higher education institutions, particularly at Universitas Sebelas Maret (UNS). This research contributes significantly to the application of data-driven decision making in academic information systems for student retention, advancing the field of educational informatics through the development of predictive frameworks that can transform traditional reactive approaches into proactive intervention strategies. The dominance of *IPS_range* as a dropout predictor underscores the need to strengthen student retention strategies through the implementation of real-time academic monitoring systems integrated with the Learning Management System (LMS), enabling early detection of academic decline. The proposed predictive framework could be integrated into UNS's existing academic information system (SIAKAD) or Learning Management System (LMS) as an early warning system for student support services, enabling proactive interventions. For instance, the system could automatically flag students at high risk based on their current academic performance, triggering alerts to academic advisors or counselors to initiate targeted support programs, such as personalized academic counseling. Additionally, the significance of features such as parental income and students' province of origin suggests the need for more adaptive and individualized intervention approaches. These could include mentoring programs or targeted coaching for students from regions with higher dropout risk, as well as improvements in financial aid schemes to ensure better targeting. However, the implementation of such predictive systems faces institutional challenges, such as fragmented data across faculties and varying levels of technological infrastructure readiness. Therefore, a phased implementation strategy is recommended, beginning with study programs that exhibit the highest dropout rates, before being scaled up more broadly across the institution. The proposed predictive framework could be integrated into UNS's early warning system for student support services, enabling proactive interventions. Beyond UNS, similar approaches could be adapted by other institutions, particularly in regions with comparable socio-economic and educational profiles, to enhance student retention policies at a national level.

5. CONCLUSION

This study develops an early prediction model for identifying students at risk of dropping out using five machine learning algorithms, combined with a RFE-CV approach for feature selection. The results show that the RF algorithm achieves the best performance, with an accuracy of 96.01% and an F1-score of 95.93%, outperforming other algorithms such as LGBM, DT, LR, and SVM. These findings reinforce the evidence that ensemble methods are more effective in capturing complex patterns in educational data compared to single models. The implementation of RFE-CV has proven to significantly enhance model efficiency. Using the top 75% of selected features, RF achieves an F1-score of 95.74% while reducing computational time by up to 14.8%, indicating that this feature selection technique is suitable for application in the context of educational data mining. Further analysis identifies three main categories of dropout predictors: (1) academic performance (Semester GPA range), (2) economic factors (source of tuition funding), and (3) demographic factors (students' province of origin). These findings affirm that dropout is a multidimensional phenomenon that requires a holistic approach to prevention. Future work may focus on deploying this predictive model into a real-time academic monitoring system integrated with institutional data infrastructure, such as the Learning Management System (LMS) or Academic Information System (SIAKAD), to facilitate proactive and timely interventions for student retention. Additionally, further research could explore the integration of qualitative data and behavioral patterns to enhance the model's interpretability and address new emerging risk factors.

REFERENCES

[1] Nurmalitasari, Z. Awang Long, and M. Faizuddin Mohd Noor, "Factors Influencing Dropout Students in Higher Education," *Educ. Res. Int.*, vol. 2023, pp. 1–13, Feb. 2023, doi: 10.1155/2023/7704142.

[2] N. L. Ratniasih, "PENERAPAN ALGORITMA K-NEAREST NEIGHBOUR (K-NN) UNTUK PENENTUAN MAHASISWA BERPOTENSI DROP OUT," *J. Teknol. Inf. Dan Komput.*, vol. 5, no. 3, Oct. 2019, doi: 10.36002/jutik.v5i3.804.

[3] 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.

[4] C. F. De Oliveira, S. R. Sobral, M. J. Ferreira, and F. Moreira, "How Does Learning Analytics Contribute to Prevent Students' Dropout in Higher Education: A Systematic Literature Review," *Big Data Cogn. Comput.*, vol. 5, no. 4, p. 64, Nov. 2021, doi: 10.3390/bdcc5040064.

[5] V. Flores, S. Heras, and V. Julián, "Comparison of Predictive Models with Balanced Classes for the Forecast of Student Dropout in Higher Education," in *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Social Good. The PAAMS Collection*, vol. 1472, F. De La Prieta, A. El Bolock, D. Durães, J. Carneiro, F. Lopes, and V. Julian, Eds., in Communications in Computer and Information Science, vol. 1472. , Cham: Springer International Publishing, 2021, pp. 139–152. doi: 10.1007/978-3-030-85710-3_12.

[6] R. Lottering, R. Hans, and M. Lall, "A model for the identification of students at risk of dropout at a university of technology," in *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, Durban, South Africa: IEEE, Aug. 2020, pp. 1–8. doi: 10.1109/icABCD49160.2020.9183874.

[7] A. Kelly and M. A. Johnson, "Investigating the Statistical Assumptions of Naïve Bayes Classifiers," in *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA: IEEE, Mar. 2021, pp. 1–6. doi: 10.1109/CISS50987.2021.9400215.

[8] Y. Qiu, J. Zhou, M. Khandelwal, H. Yang, P. Yang, and C. Li, "Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration," *Eng. Comput.*, vol. 38, no. S5, pp. 4145–4162, Dec. 2022, doi: 10.1007/s00366-021-01393-9.

[9] M. Pagan, M. Zarlis, and A. Candra, "Investigating the impact of data scaling on the k-nearest neighbor algorithm," *Comput. Sci. Inf. Technol.*, vol. 4, no. 2, pp. 135–142, Jul. 2023, doi: 10.11591/csit.v4i2.p135-142.

[10] S. Huang, M. Huang, and Y. Lyu, "A novel approach for sand liquefaction prediction via local mean-based pseudo nearest neighbor algorithm and its engineering application," *Adv. Eng. Inform.*, vol. 41, p. 100918, Aug. 2019, doi: 10.1016/j.aei.2019.04.008.

[11] Q. Wang, T.-T. Nguyen, J. Z. Huang, and T. T. Nguyen, "An efficient random forests algorithm for high dimensional data classification," *Adv. Data Anal. Classif.*, vol. 12, no. 4, pp. 953–972, Dec. 2018, doi: 10.1007/s11634-018-0318-1.

[12] R. Garcia Leiva, A. Fernandez Anta, V. Mancuso, and P. Casari, "A Novel Hyperparameter-Free Approach to Decision Tree Construction That Avoids Overfitting by Design," *IEEE Access*, vol. 7, pp. 99978–99987, 2019, doi: 10.1109/ACCESS.2019.2930235.

[13] S. Ray, "A Quick Review of Machine Learning Algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India: IEEE, Feb. 2019, pp. 35–39. doi: 10.1109/COMITCon.2019.8862451.

[14] Y. ElNakieb *et al.*, "The Role of Diffusion Tensor MR Imaging (DTI) of the Brain in Diagnosing Autism Spectrum Disorder: Promising Results," *Sensors*, vol. 21, no. 24, p. 8171, Dec. 2021, doi: 10.3390/s21248171.

[15] M. Awad and S. Fraihat, "Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems," *J. Sens. Actuator Netw.*, vol. 12, no. 5, p. 67, Sep. 2023, doi: 10.3390/jsan12050067.

[16] O. Jiménez, A. Jesús, and L. Wong, "Model for the Prediction of Dropout in Higher Education in Peru applying Machine Learning Algorithms: Random Forest, Decision Tree, Neural Network and Support Vector Machine," in *2023 33rd Conference of Open Innovations Association (FRUCT)*, Zilina, Slovakia: IEEE, May 2023, pp. 116–124. doi: 10.23919/FRUCT58615.2023.10143068.

[17] J. G. C. Krüger, A. D. S. Britto, and J. P. Barddal, "An explainable machine learning approach for student dropout prediction," *Expert Syst. Appl.*, vol. 233, p. 120933, Dec. 2023, doi: 10.1016/j.eswa.2023.120933.

[18] A. Chowdhury *et al.*, "Ultrasound classification of breast masses using a comprehensive Nakagami imaging and machine learning framework," *Ultrasonics*, vol. 124, p. 106744, Aug. 2022, doi: 10.1016/j.ultras.2022.106744.

[19] J. Kolluri, V. K. Kotte, M. S. B. Phridviraj, and S. Razia, "Reducing Overfitting Problem in Machine Learning Using Novel L1/4 Regularization Method," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, Tirunelveli, India: IEEE, Jun. 2020, pp. 934–938. doi: 10.1109/ICOEI48184.2020.9142992.

[20] V. Lumumba, D. Kiprotich, M. Mpaine, N. Makena, and M. Kavita, "Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models," *Am. J. Theor. Appl. Stat.*, vol. 13, no. 5, pp. 127–137, Oct. 2024, doi: 10.11648/j.ajtas.20241305.13.

[21] S. A. Alex, J. Jesu Vedha Nayahi, and S. Kaddoura, "Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification," *Appl. Soft Comput.*, vol. 156, p. 111491, May 2024, doi: 10.1016/j.asoc.2024.111491.

[22] D. Arifah, T. H. Saragih, D. Kartini, M. Muliadi, and M. I. Mazdadi, "Application of SMOTE to Handle Imbalance Class in Deposit Classification Using the Extreme Gradient Boosting Algorithm," *J. Ilm. Tek. Elektro Komput. Dan Inform.*, vol. 9, no. 2, pp. 396–410, Jun. 2023, doi: 10.26555/jiteki.v9i2.26155.

[23] K. Gajowniczek and M. Dudziński, "Influence of Explanatory Variable Distributions on the Behavior of the Impurity Measures Used in Classification Tree Learning," *Entropy*, vol. 26, no. 12, p. 1020, Nov. 2024, doi: 10.3390/e26121020.

[24] M. Kretowski, *Evolutionary Decision Trees in Large-Scale Data Mining*, vol. 59. in Studies in Big Data, vol. 59. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-21851-5.

[25] H. Zhao, "Research on the Application of Improved Decision Tree Algorithm based on Information Entropy in the Financial Management of Colleges and Universities," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 12, 2022, doi: 10.14569/IJACSA.2022.0131284.

[26] P. Gulati, A. Sharma, and M. Gupta, "Theoretical Study of Decision Tree Algorithms to Identify Pivotal Factors for Performance Improvement: A Review," *Int. J. Comput. Appl.*, vol. 141, no. 14, pp. 19–25, May 2016, doi: 10.5120/ijca2016909926.

[27] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Syst. Appl.*, vol. 237, p. 121549, Mar. 2024, doi: 10.1016/j.eswa.2023.121549.

[28] T.-T. Huynh-Cam, L.-S. Chen, and H. Le, "Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance," *Algorithms*, vol. 14, no. 11, p. 318, Oct. 2021, doi: 10.3390/a14110318.

[29] W. Chen, S. Zhang, R. Li, and H. Shahabi, "Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling," *Sci. Total Environ.*, vol. 644, pp. 1006–1018, Dec. 2018, doi: 10.1016/j.scitotenv.2018.06.389.

[30] H. Virro, A. Kmoch, M. Vainu, and E. Uuemaa, "Random forest-based modeling of stream nutrients at national level in a data-scarce region," *Sci. Total Environ.*, vol. 840, p. 156613, Sep. 2022, doi: 10.1016/j.scitotenv.2022.156613.

[31] L. Torre-Tojal, A. Bastarrika, A. Boyano, J. M. Lopez-Guede, and M. Graña, "Above-ground biomass estimation from LiDAR data using random forest algorithms," *J. Comput. Sci.*, vol. 58, p. 101517, Feb. 2022, doi: 10.1016/j.jocs.2021.101517.

[32] Md. K. Islam, P. Hridi, Md. S. Hossain, and H. S. Narman, "Network Anomaly Detection Using LightGBM: A Gradient Boosting Classifier," in *2020 30th International Telecommunication Networks and Applications Conference (ITNAC)*, Melbourne, VIC, Australia: IEEE, Nov. 2020, pp. 1–7. doi: 10.1109/ITNAC50341.2020.9315049.

[33] D. Zhang and Y. Gong, "The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure," *IEEE Access*, vol. 8, pp. 220990–221003, 2020, doi: 10.1109/ACCESS.2020.3042848.

[34] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "A Novel Scheme for Mapping of MVT-Type Pb–Zn Prospectivity: LightGBM, a Highly Efficient Gradient Boosting Decision Tree Machine Learning Algorithm," *Nat. Resour. Res.*, vol. 32, no. 6, pp. 2417–2438, Dec. 2023, doi: 10.1007/s11053-023-10249-6.

[35] T. O. Omotehinwa, D. O. Oyewola, and E. G. Dada, "A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis," *Healthc. Anal.*, vol. 4, p. 100218, Dec. 2023, doi: 10.1016/j.health.2023.100218.

[36] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augment. Hum. Res.*, vol. 5, no. 1, p. 12, Dec. 2020, doi: 10.1007/s41133-020-00032-0.

[37] J. Phillips, E. Cripps, J. W. Lau, and M. R. Hodkiewicz, "Classifying machinery condition using oil samples and binary logistic regression," *Mech. Syst. Signal Process.*, vol. 60–61, pp. 316–325, Aug. 2015, doi: 10.1016/j.ymssp.2014.12.020.

[38] C. Starbuck, "Logistic Regression," in *The Fundamentals of People Analytics*, Cham: Springer International Publishing, 2023, pp. 223–238. doi: 10.1007/978-3-031-28674-2_12.

[39] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," in *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, Palladam, Tamilnadu, India: IEEE, Feb. 2019, pp. 24–28. doi: 10.1109/ISS1.2019.8908018.

[40] S. F. Hussain, "A novel robust kernel for classifying high-dimensional data using Support Vector Machines," *Expert Syst. Appl.*, vol. 131, pp. 116–131, Oct. 2019, doi: 10.1016/j.eswa.2019.04.037.

[41] Md. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset,"

Comput. Methods Programs Biomed. Update, vol. 4, p. 100118, 2023, doi: 10.1016/j.cmpbup.2023.100118.

[42] T. Evgeniou and M. Pontil, “Support Vector Machines: Theory and Applications,” in *Machine Learning and Its Applications*, vol. 2049, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos, Eds., in Lecture Notes in Computer Science, vol. 2049. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 249–257. doi: 10.1007/3-540-44673-7_12.

[43] R. Nariswari and H. Pudjihastuti, “Support Vector Machine Method for Predicting Non-Linear Data,” *Procedia Comput. Sci.*, vol. 227, pp. 884–891, 2023, doi: 10.1016/j.procs.2023.10.595.

[44] X. Zhou, P. Lu, Z. Zheng, D. Tolliver, and A. Keramati, “Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree,” *Reliab. Eng. Syst. Saf.*, vol. 200, p. 106931, Aug. 2020, doi: 10.1016/j.ress.2020.106931.

[45] V. Chang, M. A. Ganatra, K. Hall, L. Golightly, and Q. A. Xu, “An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators,” *Healthc. Anal.*, vol. 2, p. 100118, Nov. 2022, doi: 10.1016/j.health.2022.100118.

[46] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, p. 6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[47] V. Tsoukas, K. Kolomvatsos, V. Chioktour, and A. Kakarountas, “A Comparative Assessment of Machine Learning Algorithms for Events Detection,” in *2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, Sep. 2019, pp. 1–4. doi: 10.1109/SEEDA-CECNSM.2019.8908366.

[48] T. Baalmann, A. Brömmelhaus, J. Hülsemann, M. Feldhaus, and K. Speck, “The Impact of Parents, Intimate Relationships, and Friends on Students’ Dropout Intentions,” *J. Coll. Stud. Retent. Res. Theory Pract.*, vol. 26, no. 3, pp. 923–947, Nov. 2024, doi: 10.1177/15210251221133374.

[49] C. Davidson and K. Wilson, “Reassessing Tinto’s Concepts of Social and Academic Integration in Student Retention,” *J. Coll. Stud. Retent. Res. Theory Pract.*, vol. 15, no. 3, pp. 329–346, Nov. 2013, doi: 10.2190/CS.15.3.b.