BAREKENG: Journal of Mathematics and Its Applications
September 2025 Volume 19 Issue 3 Page 2083-2096
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss3pp2083-2096

IMPACT OF FEATURE SELECTION ON DECISION TREE AND RANDOM FOREST FOR CLASSIFYING STUDENT STUDY SUCCESS

Firdaus Amruzain Satiranandi Wibowo [1*], Heri Retnawati[2], Muhammad Lintang Damar Sakti[3], Asma Khoirunnisa[4], Angella Ananta Batubara[5], Miftah Okta Berlian[6], Zulfa Safina Ibrahim[7], Jailani[8], Sumaryanto[9], Lantip Diat Prasojo[10]

[1,2,3,4,5,6,8]*Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Negeri Yogyakarta*
[9,10]*Department of Sports Science, Faculty of Sports and Health Sciences, Universitas Negeri Yogyakarta*
[7]*Department of Educational Research and Evaluation, Faculty of Postgraduate, Universitas Negeri Yogyakarta*
Jln. Colombo No.1, Yogyakarta, 55281, Indonesia

*Corresponding author's e-mail: * **firdausamruzain.2021@studnent.uny.ac.id***

Article History:

Received: 30th December 2024
Revised: 6th February 2025
Accepted: 12th April 2025
Published: 1st July 2025

Keywords:

Classification;
Decision Tree;
Machine Learning;
Random Forest;
Selection Feature.

ABSTRACT

The advancement of technology has a profound impact on the field of education. Education plays a crucial role in enhancing quality of life, particularly in higher education, where one of the key parameters is student success. This study investigates the influence of feature selection on the performance of machine learning models, particularly Decision Tree and Random Forest, in classifying student academic success. Utilizing a dataset of 19,061 students, the research aims to identify significant variables impacting classification outcomes. Feature selection was conducted using LASSO regression, resulting in a refined dataset of critical predictors. To address data imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied, improving the representation of underrepresented classes. Both Decision Tree and Random Forest models were trained on balanced datasets, with performance evaluated using accuracy, precision, recall, and F1-score metrics. The Random Forest model demonstrated superior accuracy (96.41%) compared to the Decision Tree (67.15%), as well as higher AUC values. Model interpretability was enhanced using SHAP (SHapley Additive exPlanations). This study underscores the utility of advanced machine learning techniques in educational analytics, paving the way for data-driven decision-making to support student achievement.

How to cite this article:

F. A. S. Wibowo, H. Retnawati, M. L. D. Sakti, A. Khoirunnisa, A. A. Batubara, M. O. Berlian, Z. S. Ibrahim, Jailani, Sumaryanto and L. D. Prasojo., "IMPACT OF FEATURE SELECTION ON DECISION TREE AND RANDOM FOREST FOR CLASSIFYING STUDENT STUDY SUCCESS," *BAREKENG: J. Math. & App.*, vol. 19, no. 3, pp. 2083-2096, September, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

The swift advancement of technology has significantly influenced a wide range of sectors, particularly in the field of education [1]. Technological advancements have significantly shaped various aspects of our daily lives and are essential in contemporary society. The integration of technology offers numerous advantages and conveniences in both professional environments and educational settings, ultimately enhancing efficiency and accessibility [2]. In addition to enhancing the quality of education at the high school and university levels, there is a pressing need for comprehensive reports that can effectively support the advancement of student academic achievement [3]. Despite the progress made in educational technology, there remains a significant gap in the development of comprehensive analytical tools and systems that are essential for enhancing student achievement [4]. The importance of addressing this issue cannot be overstated, as there is an urgent need for intelligent, data-driven strategies aimed at gaining a deeper understanding of student academic outcomes and enhancing overall performance.

This information is derived from a comprehensive analysis of course grades, student achievement indexes, and various other data points, all of which contribute valuable insights aimed at enhancing student academic performance [5]. The landscape of higher education is increasingly shaped by market standards, including considerations of profitability and quality. As a result, competitiveness has emerged as a critical factor in demonstrating institutional excellence, particularly within universities [6]. The competitiveness of a higher education institution significantly influences students' propensity to continue their studies at that institution. Consequently, institutions that excel in competitiveness are more likely to ensure their long-term sustainability, particularly when compared to those that do not keep pace with their peers in the academic landscape [7]. One key factor that reflects the quality of a higher education institution is the rate of student success, which is characterized by high graduation rates and low dropout rates. This concept is further underscored by the overall success metrics. Student success is primarily measured by the timely completion of degree programs, whereas student failure is assessed through the number of individuals who discontinue their education [8]. The challenges faced by students often stem from internal factors that significantly impact their academic performance. These internal factors include elements such as interest, motivation, and overall health. Conversely, external factors—such as the influence of peers, the family environment, and the availability of facilities and infrastructure—also play a crucial role in shaping the learning experience. It is essential for educators to understand and consider both internal and external factors to effectively enhance the quality of student learning and improve academic outcomes [9].

Implementing data mining is a significant step toward addressing this challenge. Data mining is a process that involves extracting valuable knowledge from vast quantities of data. This technique is primarily utilized to identify previously unknown patterns and to transform raw data into understandable and actionable insights [10]. Machine learning refers to the scientific investigation of algorithms and statistical models that enable computer systems to perform specific tasks autonomously, without the need for explicit programming [11]. There is a significant need for advanced machine learning techniques to assist in simplifying processes following data mining. Among these techniques, supervised machine learning stands out as one of the most widely utilized and effective approaches within the field [12]. Machine learning encompasses two primary categories of problems: classification and regression. In the classification framework, the objective is to accurately predict a class label from a set of predefined options [13]. The implementation of data mining techniques within the educational sector has garnered significant interest in recent years. Educational Data Mining (EDM) refers to the application of conventional data mining methods to address challenges and enhance decision-making processes in the realm of education [14].

Accurate predictions of student academic achievement necessitate a comprehensive understanding of the various factors and features that influence educational outcomes [15]. In this context, machine learning methods effectively develop models that align input data with anticipated target values when addressing a supervised optimization problem. Among the prevalent models utilized for classification in machine learning, Decision Trees and Random Forests stand out. Decision Trees serve as tree-structured classification models that offer clarity and accessibility, making them suitable for users with varying levels of expertise. These models can be efficiently generated from data and are rooted in one of the oldest and most established techniques for learning discriminative models, with origins in the field of statistics. On the other hand, Random Forest is an ensemble method that consists of multiple Decision Tree classifiers [16]. This approach enhances prediction accuracy by aggregating the outputs of individual trees, thereby leveraging the strengths of each. There are several approaches to introducing randomness in decision tree construction methods. Random forests serve as effective tools for making predictions regarding both nominal (classification) and

numeric (regression) target attributes. Recognized as one of the most proficient prediction models, random forests leverage ensemble learning to enhance prediction accuracy. The Feature Selection method is a widely adopted technique for feature reduction, often employed alongside classification efforts. This approach facilitates an improvement in feature quality prior to training with various classification algorithms, including Naïve Bayes, K-Nearest Neighbors, and Support Vector Machines, among others. It is important to note that Feature Selection methods each possess different biases in feature selection, much like the various classifiers themselves. Therefore, it is essential to recognize that using a specific combination of feature selection methods and classifiers may not yield generalizable results, as the effectiveness of certain feature selection methods can be significantly influenced by the classifier in use. The prior research presented an enhanced Random Forest classifier specifically developed for predicting student performance [17]. This model demonstrated a high level of accuracy; however, effective feature selection and hyperparameter tuning are critical for optimizing performance.

This study makes a significant contribution to the educational data mining field by implementing Decision Tree and Random Forest algorithms to classify student academic achievement. These models are selected due to their accuracy, interpretability, and suitability for educational classification tasks. In addition, this research explores the effect of various Feature Selection techniques to enhance model performance by reducing irrelevant attributes and improving predictive quality. Since each classifier and feature selection method carries specific biases, this study carefully evaluates their combinations to identify optimal pairings. Prior studies have shown the effectiveness of enhanced Random Forest classifiers, yet few have investigated the joint impact of feature selection and parameter optimization in academic prediction tasks. This study utilizes data from a reputable university in Yogyakarta, offering valuable insights for improving educational strategies and decision-making.

2. RESEARCH METHODS

2.1 Machine Learning

Machine Learning, often referred to as ML, is a branch of computer science focused on developing algorithms and statistical models that enable computer systems to perform tasks without the need for explicit programming [18]. Researchers are actively exploring ways to advance artificial intelligence (AI) towards achieving human-like capabilities. Machine Learning involves the analysis of patterns and the formulation of conclusions based on data. The algorithms utilized in this field generate mathematical models derived from sample data, commonly referred to as training data. The application of these techniques is integral to the ongoing development of machine learning and AI technologies. Ultimately, these systems validate the algorithms and programs that operate on computer infrastructure, contributing to more efficient and intelligent decision-making processes [19]. To ensure the successful application of a data mining solution, it is essential to approach it as a comprehensive process rather than merely a collection of tools or techniques. By carefully evaluating the outcomes at each stage of the SEMMA process, one can effectively address new questions that arise from previous results. This iterative approach allows for a return to the exploration phase, facilitating further refinement of the data and enhancing the overall analysis [20] as shown as in the Figure 1.

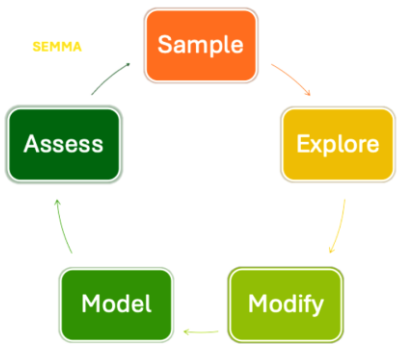

Figure 1. SEMMA Structure

From the **Figure 1**. The SEMMA methodology consists of five distinct stages. The first stage, Sample, involves extracting a subset from a larger dataset, ensuring that it is sufficiently representative to contain significant information while remaining manageable for efficient analysis. In the second stage, Explore, users investigate the dataset for unexpected patterns and anomalies, which aids in gaining a deeper understanding of the data. The third stage, Modify, focuses on the creation, selection, and transformation of variables that will be central to the model development process. In the Model stage, the emphasis is placed on identifying combinations of variables that consistently predict the desired outcomes. Finally, the Assess phase entails evaluating the utility and reliability of the insights gained throughout the data mining process. This structured approach enables effective data management and enhances the quality of predictive modeling efforts.

2.2 Classification

Classification is a fundamental process involved in the development of models or functions that effectively describe and differentiate between various data classes or concepts. These models are constructed through a thorough analysis of a designated set of training data, which comprises data objects with established class labels. Once established, the model serves as a valuable tool for predicting the class labels of previously unclassified data objects **[21]**. Classification is an effective approach for managing large datasets. It encompasses two primary methodologies: supervised learning, where a model is trained on labeled data, and unsupervised learning, which involves identifying patterns in unlabeled data as shown as in the **Figure 2**.

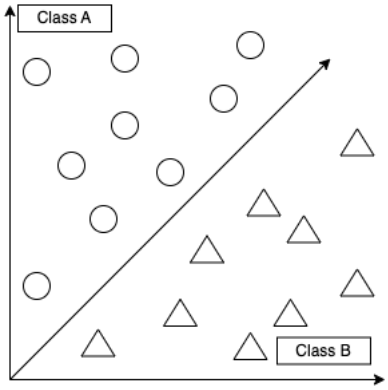

Figure 2. Classification Concept

Figure 2 illustrates a fundamental concept of classification in machine learning. In this representation, two distinct classes, Class A (represented by circles) and Class B (represented by triangles), are visualized within a two-dimensional feature space. Each axis likely corresponds to a particular feature or variable relevant to the classification problem. The objective of a classification algorithm is to find a decision boundary that best separates the two classes. This boundary may be linear or non-linear depending on the complexity of the data and the model used. A linear boundary would form a straight line, while a more flexible model may create a curved or irregular boundary to accurately capture the separation between the classes.

2.3 Random Forest

Random Forest (RF) is an advanced methodology designed to enhance accuracy in the generation of attributes for each node through a randomized process **[22]**. The Random Forest algorithm, introduced by Leo, is a sophisticated method designed to enhance accuracy in generating attributes for each node through a randomized process. This approach allows for improved performance and reliability in predictive modeling **[23]**. Random Forest (RF) is composed of a set of decision trees that work collaboratively to classify data into distinct categories. The classification process involves evaluating the nodes within each tree and ultimately reaching various leaf nodes to derive a conclusive result **[24]**. The process of constructing a decision tree utilizing the Random Forest (RF) method closely resembles that of the Classification and Regression Tree (CART) approach; however, it is important to note that no pruning is performed in the Random Forest methodology. At each internal node of the decision tree, the Gini index is employed to determine the selection of features. The calculation of the Gini Index value can be conducted as follow on **Equation (1)**:

$$Gini(S_i) = 1 - \sum_{i=0}^{c-1} p_i^2 \quad (1)$$

The variable p_i represents the relative frequency of class C_i within the set. C_i denotes the class for $i = 1, \cdots, c-1$, and c is the total number of identified classes. The quality of the split for the feature k into the subset S_i can be assessed by the count of samples that belong to class C_i. This assesment is further quantified through the calculation of Gini impurity measures derived from the resulting subset. The relevant data can be evaluated using the **Equation (2)**:

$$Gini_{split} = \sum_{i=0}^{k-1} \left(\frac{n_i}{n}\right) Gini(S_i) \quad (2)$$

In the context of Random Forest ensembles, n_i represent the number of samples within the subset S_i after the data has been partitioned, while n denotes the total number of samples at a specified node. For instance, $\{h(x, \Theta_k), k = 1, ...\}$ where $\{\Theta_k\}$ constitutes an independet and identially distributed (iid) random vector. Each decision tree employs a majority vote mechanism to determine the class with the highest average occurrence. For Random Forests, the upper limit can be quantified individually, assuming independence between the respective components **[25]**.

2.4 Decision Tree

A decision tree is a structured model wherein each non-leaf node, also referred to as an internal node, corresponds to a specific decision, while the leaf nodes typically represent an outcome or class label. Each internal node conducts a test on one or more attribute values, leading to two or more branches. These branches are associated with potential decision values and are designed to be mutually exclusive and collectively exhaustive, ensuring that each possible outcome is represented through a distinct link. Decision trees are recognized as one of the most effective methods for decision-making. This approach provides a systematic framework for organizing choices and examining all potential outcomes associated with a given option. Within the context of decision trees, several key terms are utilized: the Decision Node, which represents the features employed to make decisions; the Root Node, which is the top decision node; the Leaf Node, signifying the output of the decision; and the Subtree, which consists of the various branches or sections of the tree.

Figure 3. Decision Tree Concept

From the **Figure 3** above This figure encapsulates how decision trees operate by transforming raw data into human-readable rules. Their interpretability and structured approach make decision trees a popular choice for various machine-learning tasks, although they can be prone to overfitting if not carefully pruned or regulated.

The ID3 algorithm effectively partitions data into two distinct groups based on their attributes by utilizing a metric known as entropy. Entropy serves as an indicator of the level of randomness within a dataset. This partitioning process is conducted systematically, progressing from top to bottom. The calculations for entropy can be outlined as follows in **Equation (3)**:

$$\boldsymbol{Entrophy(S) = -\sum_{i=1}^{n} p_i \times log_2(p_i)} \quad (3)$$

where:

S : Set of cases
n : Number of participants
$\boldsymbol{p_i}$: number of cases in the-$\boldsymbol{i}$ partition

2.5 Feature Selection

Feature selection is an essential process in scenarios where there is a large volume of features, particularly when some of them may be noisy or redundant. In the context of classification tasks, supervised feature selection is frequently employed [26]. The presence of class labels allows for the application of algorithmically driven supervised feature selection, which can effectively identify discriminative features that enable the differentiation of various categories [27]. In this research, researcher using Lasso regression for feature selection. Lasso regression serves as an effective method for feature selection, as it allows for the reduction of coefficient estimates associated with less significant variables. This capability enhances the model's performance by focusing on the most impactful predictors.

$$\min_{\beta_0, \beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where:

β_0 : Intercept
β_j : Regression coefficient for feature-j
x_{ij} : Value of the-j feature for data-i
n : Number of samples
p : Number of features

The formula presented includes several key components, specifically the loss function and LASSO regularization:

1. The *loss function* employed in this context is the Mean Square Error (MSE). This function quantifies the deviation of the predicted value $\hat{y}_i$ from the actual value y_i.
2. Lasso Regulatization:

   The penalty applied in this context is the total sum of the absolute values of all regression coefficients, denoted as $|\beta_j|$. The parameter λ serves as a regularization factor. When λ is set to 0, the model corresponds to standard linear regression, while a larger λ results in an increased number of coefficients $|\beta_j|$ being reduced to zero.

In the present research, Lasso regression is applied following the implementation of Decision Tree and Random Forest methodologies. This technique incorporates a penalty term into the linear regression framework, which serves to discourage excessively large coefficients [28]. The advantages of Lasso regression in addressing the challenges of this study stem from its ability to manage data sparsity. By introducing a penalty term, it effectively eliminates unstable variables, leading to more robust conclusions. Furthermore, Lasso regression demonstrates strong adaptability. It prevents the direct estimation of the coefficient for the target matrix, enhancing the modeling process. This approach also allows for improved interpretability of decision trees by incorporating only the most significant attributes within the linear models associated with the leaves [29]. The integration of Lasso regression with Decision Tree algorithms presents a conservative approach to power classification. This methodology effectively selects variables based on distinct criteria, allowing for conclusions to be drawn from two independent perspectives. By applying Lasso regression to Random Forest, the coefficients of certain predicted values from individual trees can be compressed to zero. Subsequent research indicates that this combination significantly enhances accuracy. The method subsequently identifies non-zero coefficients as the features essential for the classification process. Furthermore, the Random Forest algorithm has the capability to minimize the size of trees within the forest; however, it is important to specify the number of aggregation trees in advance.

3. RESULTS AND DISCUSSION

3.1 Variable Distribution

The data collected were derived from students' academic records, surveys administered to students, and an assessment of their socio-economic background conditions. This research presents a comparative analysis of the Machine Learning Decision Tree and Random Forest methodologies in the classification of student learning success. The dataset utilized encompasses a variety of variables pertinent to academic achievement, including Grade Point Average (GPA), gender, entry route, father's education, group participation, and other relevant factors influencing academic performance. Through feature selection employing the LASSO Regression technique, a total of eleven significant variables were identified. These include Faculty (specific category), Entry Path (specific category), Gender, Student Special Needs, Father's Education (specific category), Father's Income (specific category), Type of Student Collaboration, Age, and multiple GPA metrics (GPA.1 to GPA.6 across the entire semester). Each of these variables plays a crucial role in enhancing the accuracy of student success classification. In the **Figure 4**, we can see the distribution on the student study success with the label 0 is failure and the label 1 is success.

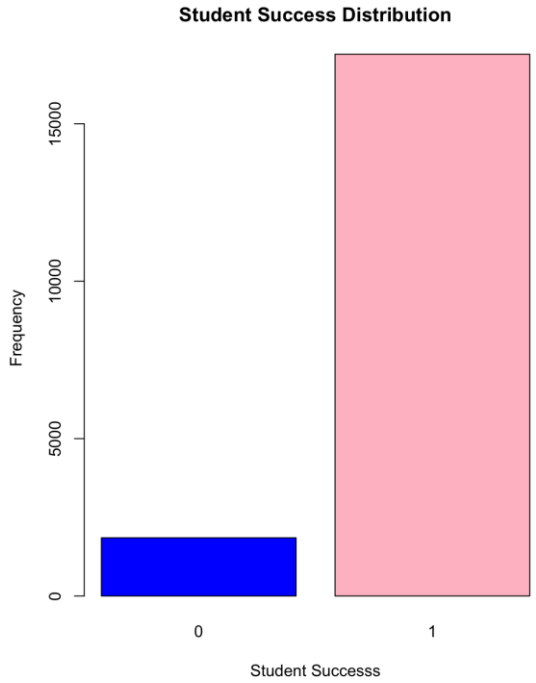

Figure 4. Student Success Distribution Bar Plot

The data presented in **Table 1** highlights the distribution of faculties and gender in relation to student success in their studies. It is evident that the number of students who have successfully completed their programs varies significantly across different faculties. The Faculty of Educational Sciences and Psychology (FIPP) has achieved the highest number of successful graduates, totaling 3,089 students. This is followed by the Faculty of Mathematics and Natural Sciences (FMIPA), which recorded 2,500 successful students. In contrast, the Faculty of Social, Legal, and Political Sciences (FISHIPOL) produced the lowest number of successful graduates, with a total of 1,364. Moreover, the Faculty of Engineering (FT) had the highest number of unsuccessful students, totaling 1,094, followed closely by the Faculty of Educational Sciences and Psychology (FIPP) with 995. The Faculty of Sports and Health Sciences (FIKK) recorded the lowest number of successful students among the faculties. This data contributes to understanding the academic performance across various fields of study.

Table 1. Faculty and Gender Distribution

Labels	Faculty							Gender	
	FBSB	FEB	FIKK	FIPP	FISHIPOL	FMIPA	FT	F	M
Success	1.903	1.491	1.568	3.089	1.364	2.950	1.628	4.094	9.899
Failure	812	528	472	995	607	560	1094	2.289	2.779
Sum	2.715	2.019	2.040	4.084	1.971	3.510	2.722	6.383	12.678
	19.601							19.061	

The analysis indicates that male students exhibit a higher overall study success rate, with 9,899 successful outcomes compared to 4,094 for female students. However, it is important to note that male students also demonstrate a higher rate of academic failure at 2,779, in contrast to 2,289 for their female counterparts. This variation in student success rates across different faculties suggests that the characteristics and structures of study programs may significantly influence academic outcomes. For instance, the Faculty of Education and Psychology (FIPP) reports the highest number of successful students, potentially attributable to a robust academic support system, an effective curriculum, or a conducive learning environment. Conversely, the Faculty of Engineering (FT) shows the highest failure rate, which may indicate that the academic challenges or external factors, such as student workload, are more pronounced in this faculty. Furthermore, while male students achieve greater success overall, they also appear to be more susceptible to academic failure compared to female students. These differences may be influenced by various social, economic, or cultural factors that shape how individuals' approach and respond to academic challenges. Additionally, GPA.1, GPA.2, GPA.3, GPA.4, GPA.5, and GPA.6 are utilized in this context. The Achievement Index represents a student's final score for each semester, reflecting the effectiveness of their learning process within a given academic period **[30]**.

3.2 Preprocessing

This process encompasses essential processes including data cleaning and management of missing values. In this research, data is meticulously processed by identifying and addressing the presence of empty or inconsistent entries. We implement strategies such as the deletion or imputation of missing values, as well as the conversion of categorical variables into suitable formats (e.g., factors). Furthermore, normalization of numeric data is carried out to ensure optimal performance of machine learning algorithms. Data visualization techniques, including distribution graphs and correlation matrices, are employed to gain deeper insights into data patterns. Additionally, the dataset is transformed into a factor data type, allowing for the correct labeling of numeric data using the mutate function. The dataset is also represented using dummy variables, which is achieved through the model.matrix() function.

The dataset utilized in this study consists of 19,061 student records that underwent a comprehensive data cleaning process. Following this process, the dataset remained at 19,061 records, as the initial data did not contain any missing values (NA). This confirms the integrity and completeness of the data used for analysis.

3.3 Splitting

The dataset is partitioned to establish distinct training and testing classes. In this methodology, a division of 70:30 is implemented, indicating that 70% of the data is allocated for the training class, while the remaining 30% is reserved for the testing class. It is important to note that the way data is divided can significantly influence the resulting accuracy.

3.4 Decision Tree Model

Decision trees are utilized due to their inherent simplicity and interpretability; however, they tend to overfit when applied to complex datasets. Each node within the tree is assigned a specific class label, which encompasses both the root and other internal nodes. These nodes incorporate conditions to test attributes, facilitating the separation of records based on their characteristics. The decision tree model is refined utilizing data generated through the SMOTE technique. Hyperparameters, including the Complexity Parameter (cp), are calibrated to enhance the performance of the model. The visualization of the decision tree illustrates a hierarchical structure based on the segmentation derived from features selected through LASSO regression.

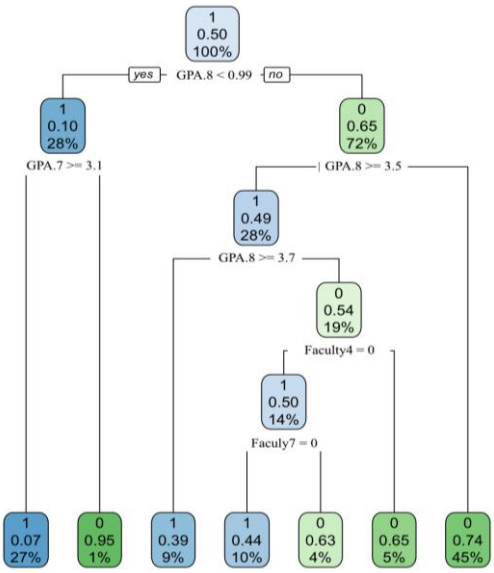

Figure 5. Node Decision Tree Plot

The graph presented in the **Figure 5** provides a visual representation of the prioritized features based on the decision-making process. Notably, variables IPK.6 (GPA on sixth semester) and IPK.7 (GPA on seventh semester) are positioned at the initial node due to their strong correlation with the target variable. The results derived from the decision tree model are as follows:

1. Confusion Matrix

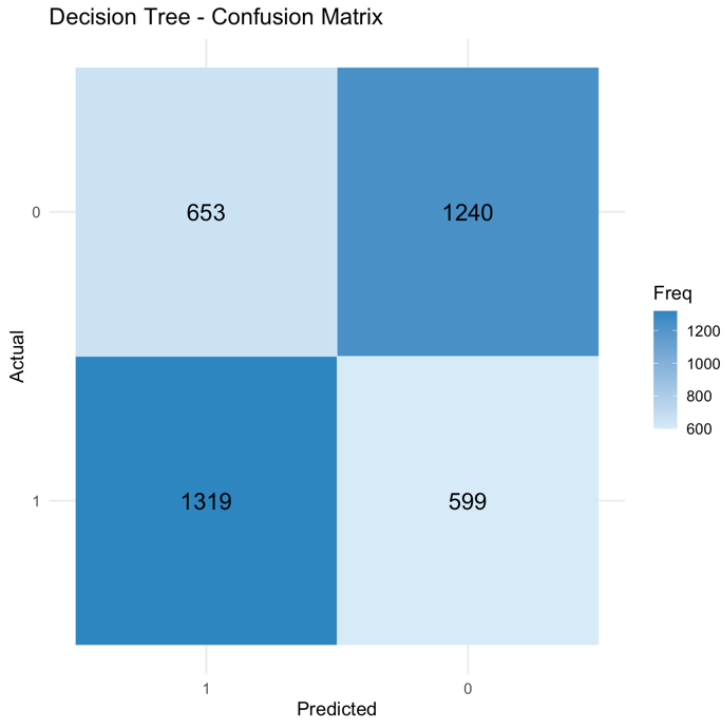

Figure 6. Decision Tree Confusion Matrix

The data presented in **Figure 6** demonstrates the effectiveness of the Decision Tree model in classification tasks. Specifically, the model accurately classified 1,319 instances of class 1, while it misidentified 653 instances of class 0 as class 1. Furthermore, it incorrectly classified 599 instances of class 1 as class 0, but it successfully identified 1,240 instances of class 0 as class 0.

2. Statistics

The analysis indicates that the accuracy of the decision tree model is 67.15%, accompanied by a precision of 66.89%. This suggests that the model effectively predicts Class 1 with a reasonable degree of

correctness. Furthermore, the recall is recorded at 68.77%, reflecting the model's capability to accurately identify all instances of Class 1. The F1-Score of the model stands at 67.81%, highlighting a balanced performance in terms of both precision and recall.

3. Weakness

Decision trees possess certain limitations, particularly their vulnerability to overfitting if the complexity parameters are not properly calibrated. Furthermore, the structural design of decision trees often lacks the robustness exhibited by ensemble models such as Random Forest, which are better suited for handling more intricate datasets. Random Forest is an ensemble algorithm that integrates multiple decision trees to enhance predictive accuracy and mitigate the risks of overfitting. To optimize performance, critical parameters, including the number of trees (n-tree) and the maximum number of features considered at each split (mtry), are fine-tuned using cross-validation techniques.

3.5 Random Forest Model

Researchers utilize the random forest methodology due to its ability to enhance accuracy and mitigate overfitting. This approach employs multiple decision trees, making it particularly effective for handling large datasets, although it does require significant computational resources. The results obtained from random forest techniques demonstrate superior performance when compared to traditional decision tree models.

1. Model Evaluation Result

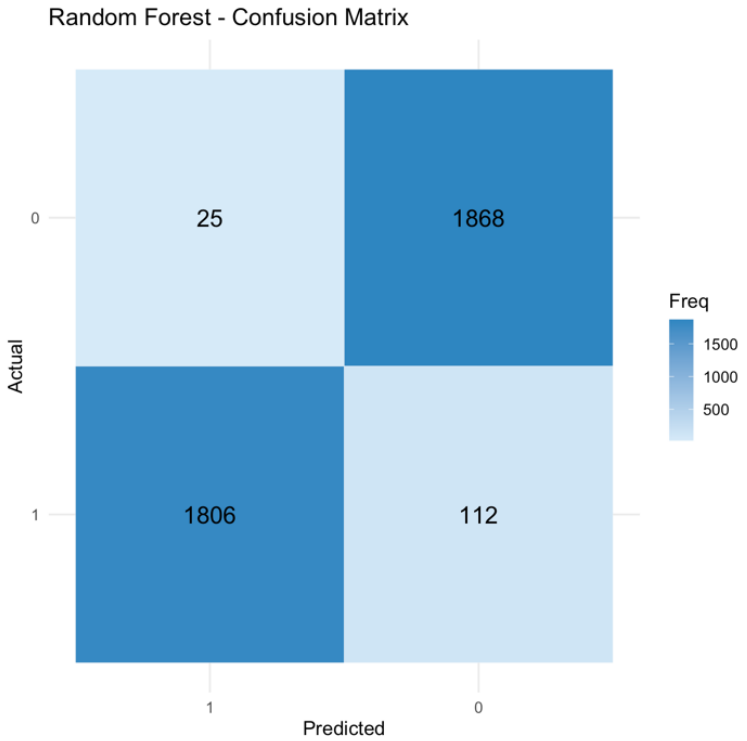

Figure 7. Random Forest Confusion Matrix

The results in **Figure 7** indicate that the Random Forest model correctly classified 1,806 instances of class 1 as class 1. Additionally, it misclassified 25 instances of class 0 as class 1 and 112 instances of class 1 as class 0. The model also accurately identified 1,868 instances of class 0 as class 0.

2. Statistics

The decision tree model demonstrated an accuracy of 96.41%, along with a precision of 98.63%. This indicates a high capability for correctly predicting instances of class 1. The recall rate was recorded at 94.16%,

reflecting the model's effectiveness in capturing all instances of class 1. Additionally, the F1-Score achieved was 96.34%, highlighting the model's overall performance.

3. Advantages and Limitations of the Model

The Random Forest model is known for its robustness against overfitting, as it integrates predictions from multiple nodes. Furthermore, it demonstrates excellent generalization capabilities when applied to complex datasets. However, it is important to note that Random Forest models typically require greater computational resources compared to Decision Trees.

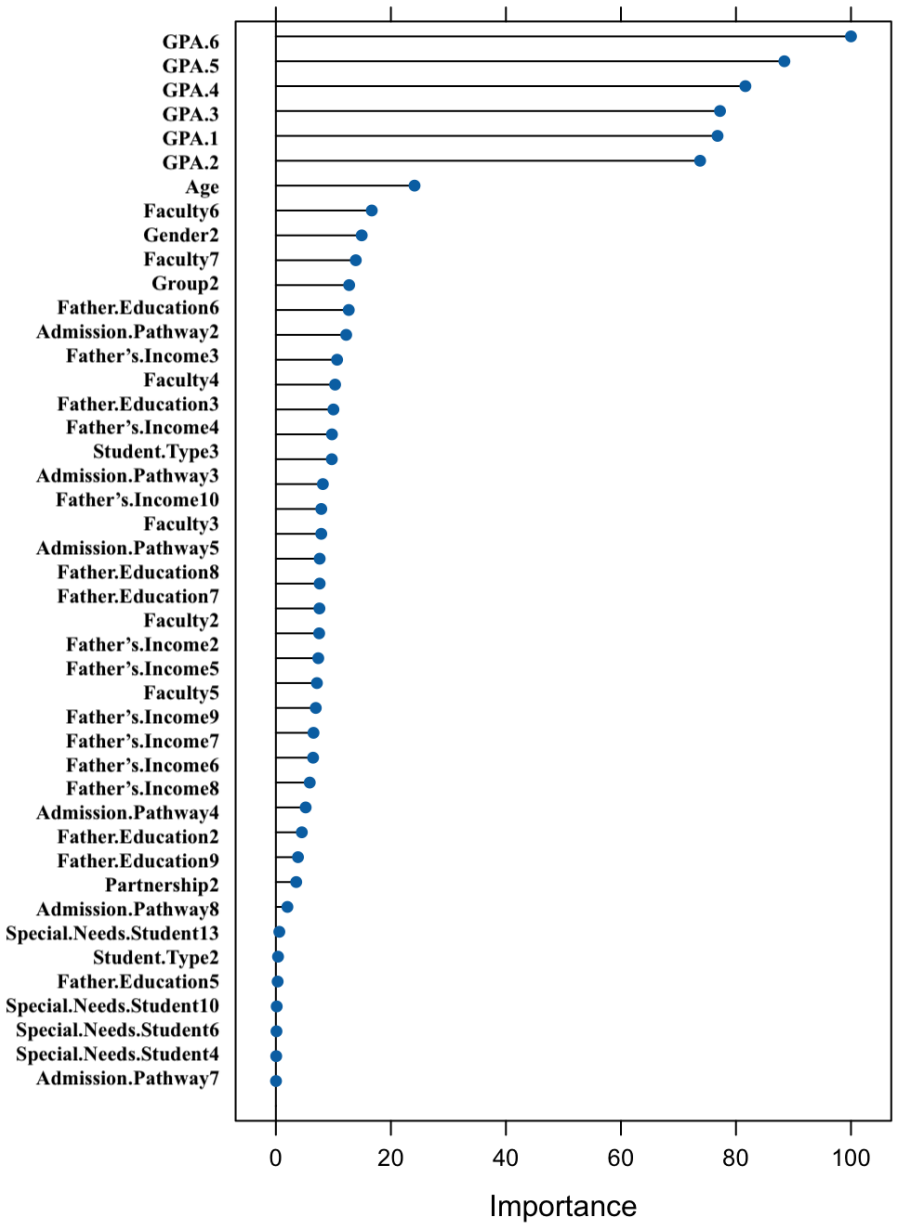

Figure 8. Plot Importance Random Forest

3.6 Comparison Between Two Models

The analysis underscores the significance of selecting appropriate features and effectively managing data to enhance model performance. The Random Forest algorithm has demonstrated superior results due to its capability to capture complex relationships among variables. Nevertheless, Decision Trees continue to serve an important role for those seeking simpler interpretations. This research offers valuable insights into the factors influencing student learning success, specifically highlighting the impact of academic and socio-economic variables. Furthermore, it emphasizes the importance of employing ensemble techniques, such as Random Forest, to achieve more accurate predictions and facilitate deeper interpretations of the data.

3.7 Visualization

In Random Forest analysis, the importance of each feature in making predictions is assessed through Feature Importance metrics. This visualization offers valuable insights into which features have the greatest impact on the predictive model. Based on the conducted analyses, it has been determined that features such as IPK.4, IPK.6, and IPK.7 demonstrate a significant contribution, as indicated by both LASSO feature selection and prior correlation analyses.

1. Partial Dependece Plot (PDP)

The objective of this analysis is to examine the relationship between specific independent variables and their predicted probabilities of study success. As illustrated in **Figure 9**, the GPA.6 (IPK.6) feature indicates that students with higher GPA.6 scores are associated with an increased likelihood of achieving study success. Conversely, the Age (Umur) feature, as represented in the PKP, did not reveal a significant correlation with study success; however, it suggests that Age contributes modestly to the overall model. This is further depicted in the accompanying visual representation.

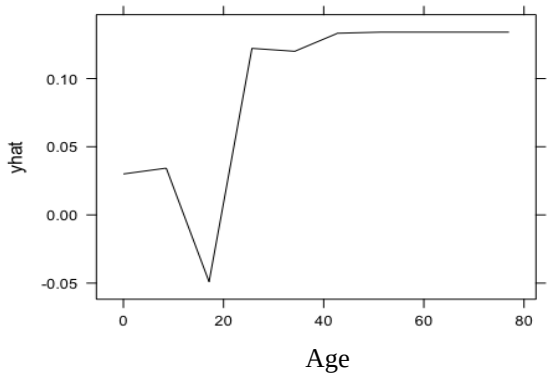

Figure 9. Partical Dependence Plot (PDP)

2. AUC Model Comparison

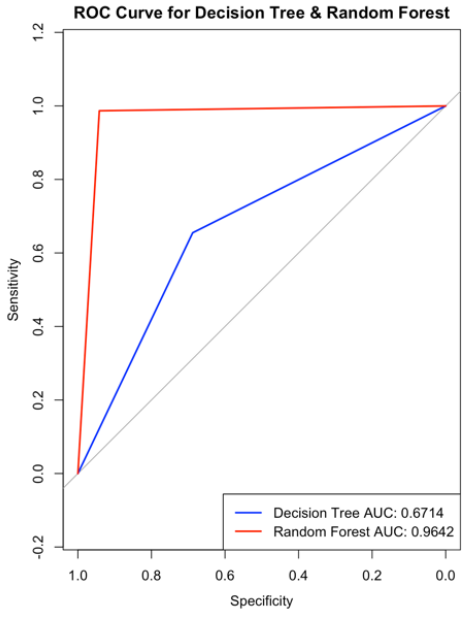

Figure 10. Decision Tree and Random Forest AUC Comparison

The analysis presented in the **Figure 10** highlights notable differences between the two models examined. The Decision Tree model achieved an AUC of 0.6713702, while the Random Forest model demonstrated a significantly higher AUC of 0.9641996. This indicates that the Random Forest model exhibits superior predictive performance compared to the Decision Tree, underscoring its effectiveness in this context.

4. CONCLUSIONS

Previous research has explored the integration of decision trees and random forests across various fields. This study presents a novel approach to enhancing educational outcomes related to student success by combining decision trees and random forests with Lasso regression. Specifically, this research emphasizes the performance of Lasso regression in relation to both decision trees and random forests. The findings reveal significant insights, particularly highlighting the influence of GPA from the first to the final semester (eighth semester) in the classification process. Lasso regression enhances the accuracy of both models by reducing data complexity and focusing on the most relevant variables, resulting in improved precision. The decision tree model achieved an accuracy of 67.15%. However, this indicates some limitations when dealing with high-dimensional data. In contrast, the random forest model exhibited a robust accuracy of 96.41%, demonstrating its resilience in managing high-dimensional and complex datasets. These results are supported by the Area Under Curve (AUC) analysis, which shows that the decision tree's AUC value is 0.671, significantly lower than the random forest's AUC value of 0.964. Furthermore, through SHapley Additive exPlanations (SHAP) and Partial Dependence Plot (PDP) analysis, GPA has been identified as a significant predictor of student study success. In future research, we can consider employing additional methodologies, such as Support Vector Machines (SVM), LightGBM, and other advanced techniques. Utilizing these alternative methods could yield varied outcomes, particularly in relation to the implementation of Lasso regression.

REFERENCES

[1] M. Arridho *et al.*, "THE TECHNOLOGY DEVELOPMENT IN THE EDUCATION FIELD," *INTERNATIONAL OF EDUCATION AND SOSSIAL (AIOES) Journal*, vol. 4, no. 2, pp. 25–29, 2023, doi: 10.55311/aioes.v3i2.199.

[2] I. Oktadiani, H. Fitriawan, M. Nurwahidin, and Herpratiwi, "PENERAPAN MACHINE LEARNING UNTUK PREDIKSI MASA STUDI MAHASISWA DI PERGURUAN TINGGI X," *ELECTRICIAN*, vol. 17, no. 3, Oct. 2023.doi: https://doi.org/10.23960/elc.v17n3.2529

[3] R. Raja and P. C. Nagasubramani, "RECENT TREND OF TEACHING METHODS IN EDUCATION" ORGANISED BY SRI SAI BHARATH COLLEGE OF EDUCATION DINDIGUL," *India Journal of Applied and Advanced Research*, vol. 2018, no. 3, pp. 33–35, 2018, doi: https://doi.org/10.21839/jaar.2018.v3iS1.165.

[4] Dr. Lohans Kumar Kalyani, "THE ROLE OF TECHNOLOGY IN EDUCATION: ENHANCING LEARNING OUTCOMES AND 21ST CENTURY SKILLS," *International Journal of Scientific Research in Modern Science and Technology*, vol. 3, no. 4, pp. 05–10, Apr. 2024, doi: https://doi.org/10.59828/ijsrmst.v3i4.199.

[5] I. B. Suleiman, O. A. Okunade, E. G. Dada, and U. C. Ezeanya, "KEY FACTORS INFLUENCING STUDENTS' ACADEMIC PERFORMANCE," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, p. 41, Sep. 2024, doi: https://doi.org/10.1186/s43067-024-00166-w.

[6] Arwildayanto, A. Suking, Arifin, and Nellitawati, *MANAJEMEN DAYA SAING PERGURUAN TINGGI*, 1st ed., vol. 1. Bandung: Cendekia Press, 2020.

[7] M. Syukri *et al.*, "KUALITAS PENDIDIKAN DAN KEUNGGULAN KOMPETITIF," *Journal on Education*, vol. 06, no. 02, pp. 11738–11747, Feb. 2024, Accessed: Dec. 06, 2024. [Online]. Available: http://jonedu.org/index.php/joe

[8] R. Purnama, "METODE PENURUNAN MULTIPEL (MULTIPLE DECREMENT METHOD) PADA DATA LAMA STUDI MAHASISWA," *Serambi Saintia*, vol. V, no. 2, Oct. 2017.

[9] R. Setiawan, W. Wagiran, and Y. Alsamiri, "CONSTRUCTION OF AN INSTRUMENT FOR EVALUATING THE TEACHING PROCESS IN HIGHER EDUCATION: CONTENT AND CONSTRUCT VALIDITY," *REID (Research and Evaluation in Education)*, vol. 10, no. 1, pp. 50–63, Jun. 2024, doi: https://doi.org/10.21831/reid.v10i1.63483.

[10] J. Han, M. Kamber, and J. Pei, "DATA MINING. CONCEPTS AND TECHNIQUES, 3RD EDITION (THE MORGAN KAUFMANN SERIES IN DATA MANAGEMENT SYSTEMS)," 2011.

[11] I. H. Sarker, "MACHINE LEARNING: ALGORITHMS, REAL-WORLD APPLICATIONS AND RESEARCH DIRECTIONS," May 01, 2021, *Springer*. doi: https://doi.org/10.20944/preprints202103.0216.v1.

[12] A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön, "SUPERVISED MACHINE LEARNING LECTURE NOTES FOR THE STATISTICAL MACHINE LEARNING COURSE," Mar. 2019.

[13] A. F. A. H. Alnuaimi and T. H. K. Albaldawi, "AN OVERVIEW OF MACHINE LEARNING CLASSIFICATION TECHNIQUES," *BIO Web Conf*, vol. 97, Apr. 2024, doi: https://doi.org/10.1051/bioconf/20249700133.

[14] M. Yağcı, "EDUCATIONAL DATA MINING: PREDICTION OF STUDENTS' ACADEMIC PERFORMANCE USING MACHINE LEARNING ALGORITHMS," *Smart Learning Environments*, vol. 9, no. 1, Dec. 2022, doi: https://doi.org/10.1186/s40561-022-00192-z.

[15] P. Utami, N. Fajaryati, P. Sudira, M. E. Ismail, N. Maneetien, and F. Felestin, "THE ROLE OF TEACHER IN INDUSTRY 5.0, TECHNOLOGY, AND SOCIAL CAPITAL IN FOR VOCATIONAL HIGH SCHOOL GRADUATES IN SCHOOL-TO-WORK TRANSITIONS," *Elinvo (Electronics, Informatics, and Vocational Education)*, vol. 9, no. 1, pp. 113–133, May 2024, doi: https://doi.org/10.21831/elinvo.v9i1.72485.

[16] E. Dewi Sri Mulyani, Y. Purnama Putra, E. Badar Sambani, S. Siti Sundari, T. Mufizar, and M. Satrio Nugraha, "STUDENT COMPETENCY ASSOCIATION ANALYSIS FOR LEARNING EVALUATION USING APRIORI ALGORITHM,"

Elinvo (Electronics, Informatics, and Vocational Education), vol. 6, no. 2, pp. 120–130, Dec. 2021, doi: https://doi.org/10.21831/elinvo.v6i2.42264.

[17] M. A. S. Pawitra, H.-C. Hung, and H. Jati, "A MACHINE LEARNING APPROACH TO PREDICTING ON-TIME GRADUATION IN INDONESIAN HIGHER EDUCATION," *Elinvo (Electronics, Informatics, and Vocational Education)*, vol. 9, no. 2, pp. 294–308, Dec. 2024, doi: https://doi.org/10.21831/elinvo.v9i2.77052.

[18] B. Mahesh, "MACHINE LEARNING ALGORITHMS - A REVIEW," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, Jan. 2020, doi: https://doi.org/10.21275/ART20203995.

[19] R. K. Dinata and N. Hasdyna, *MACHINE LEARNING*, 1st ed., vol. 1. Lhokseumawe: UNIMAL, 2020.

[20] D. L. Olson and D. Delen, *ADVANCED DATA MINING TECHNIQUES*. Heidelberg: Springer, 2008. doi: https://doi.org/10.1007/978-3-540-76917-0_2.

[21] L. Li and M. Iskander, "USE OF MACHINE LEARNING FOR CLASSIFICATION OF SAND PARTICLES," *Acta Geotech*, vol. 17, no. 10, pp. 4739–4759, Oct. 2022, doi: https://doi.org/10.1007/s11440-021-01443-y.

[22] Suci Amaliah, M. Nusrang, and A. Aswi, "PENERAPAN METODE RANDOM FOREST UNTUK KLASIFIKASI VARIAN MINUMAN KOPI DI KEDAI KOPI KONIJIWA BANTAENG," *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, vol. 4, no. 3, pp. 121–127, Dec. 2022, doi: https://doi.org/10.35580/variansiunm31.

[23] M. L. Ruiz-Rodriguez, J. Andres Sandoval-Bringas, and M. A. Carreno-Leon, "CLASSIFICATION OF STUDENT SUCCESS USING RANDOM FOREST AND NEURAL NETWORKS," in *Proceedings - 2020 3rd International Conference of Inclusive Technology and Education, CONTIE 2020*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 98–103. doi: https://doi.org/10.1109/CONTIE51334.2020.00027.

[24] D. Alita and A. Rahman, "PENDETEKSIAN SARKASME PADA PROSES ANALISIS SENTIMEN MENGGUNAKAN RANDOM FOREST CLASSIFIER," *Jurnal Komputasi*, vol. 8, no. 2, 2020.doi: https://doi.org/10.23960/komputasi.v8i2.2615

[25] S. A. Cushman, K. Kilshaw, R. D. Campbell, Z. Kaszta, M. Gaywood, and D. W. Macdonald, "COMPARING THE PERFORMANCE OF GLOBAL, GEOGRAPHICALLY WEIGHTED AND ECOLOGICALLY WEIGHTED SPECIES DISTRIBUTION MODELS FOR SCOTTISH WILDCATS USING GLM AND RANDOM FOREST PREDICTIVE MODELING," *Ecol Modell*, vol. 492, Jun. 2024, doi: https://doi.org/10.1016/j.ecolmodel.2024.110691.

[26] H. Pramoedyo, D. Ariyanto, and N. N. Aini, "COMPARISON OF RANDOM FOREST AND NAÏVE BAYES METHODS FOR CLASSIFYING ANF FORECASTING SOIL TEXTURE IN THE AREA AROUND DAS KALIKONTO, EAST JAVA," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 16, no. 4, pp. 1411–1422, Dec. 2022, doi: https://doi.org/10.30598/barekengvol16iss4pp1411-1422.

[27] C. Sammut and G. I. Webb, *ENCYCLOPEDIA OF MACHINE LEARNING*. Victoria, 2011. doi: https://doi.org/10.1007/978-0-387-30164-8.

[28] D. K. Dalimunthe and R. B. F. Hakim, "APPLICATION OF RANDOM FOREST ALGORITHM ON WATCH PRICE PREDICTION SYSTEM USING FRAMEWORK FLASK," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 1, pp. 0171–0184, Apr. 2023, doi: https://doi.org/10.30598/barekengvol17iss1pp0171-0184.

[29] Guozheng. Li, *Proceedings : 2013 IEEE International Conference on Bioinformatics and Biomedicine : 18-21 December 2013, Shanghai, China*. IEEE, 2013.

[30] I. A. M. S. Widiastuti, "ASSESSMENT AND FEEDBACK PRACTICES IN THE EFL CLASSROOM," *REID (Research and Evaluation in Education)*, vol. 7, no. 1, pp. 13–22, Jun. 2021, doi: https://doi.org/10.21831/reid.v7i1.37741.