JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7294

MACHINE LEARNING TO IDENTIFY ELIGIBILITY OF STUDENTS RECEIVING SINGLE TUITION RELIEF

M. Ghofar Rohman[1*]; Zubaile Abdullah[2]; Shahreen Kasim[2]; M. Ulul Albab[3]

Informatics Engineering, Fakulty of Science and Technology[1]
Mathematics Education, Faculty of Teaching Training and Education[3]
Universitas Islam Lamongan, Lamongan, Indonesia[1,3]
www.unisla.ac.id[1,3]
m.ghofarrohman@unisla.ac.id*, mululalbab@unisla.ac.id

Faculty of Computer Sciences and Information Technology[2]
University Tun Hussein Onn Malaysia, Johor, Malaysia[2]
www.uthm.edu.my[2]
zubaile@uthm.edu.my, shahreen@uthm.edu.my

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *The cost of higher education in Indonesia varies greatly and often becomes a financial burden for students. Socio-economic factors such as parental income, occupation, number of dependents, vehicle ownership, and place of residence influence the determination of single tuition as regulated by the Ministry of Education Regulation No. 55 of 2013. This study aims to classify freshmen eligibility for single tuition relief using five machine learning models: RF, LR, KNN, SVM, and NB. The dataset contains 2000 rows of data with six socio-economic attributes divided into two classes: eligible and ineligible. The data were split into 80% training and 20% testing, and model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Results show that without SMOTE, all models suffer from severe majority-class bias, yielding critically low recall for the minority class SVM = 0.014; NB = 0.004. SMOTE significantly improves minority-class detection, with RF and SVM achieving the highest performance F1-scores of 0.820 and 0.801, and ROC-AUC of 0.966 and 0.990, respectively. SHAP analysis identifies Number of Dependents of Parents as the most influential predictor across all models, highlighting its central role in financial need assessment. These findings demonstrate that combining SMOTE with ensemble or margin-based models enhances classifiying fairness and sensitivity in educational support systems. The future work recommend expanding features to include behavioral, academic, and regional indicators, using multi-institutional data, and exploring deep learning or advanced resampling methods to enhance generalizability and robustness.*

Keywords: *Classification, Higher Education, Machine Learning, SHAP, Single Tuition*

Intisari— *Biaya pendidikan tinggi di Indonesia sangat bervariasi dan sering kali menjadi beban finansial bagi mahasiswa. Faktor sosial ekonomi seperti pendapatan orang tua, pekerjaan, jumlah tanggungan, kepemilikan kendaraan, dan tempat tinggal memengaruhi penentuan besaran uang kuliah tunggal sebagaimana diatur dalam Peraturan Menteri Pendidikan Nomor 55 Tahun 2013. Penelitian ini bertujuan untuk mengklasifikasikan kelayakan mahasiswa baru dalam penerimaan keringanan uang kuliah tunggal menggunakan lima model pembelajaran mesin: RF, LR, KNN, SVM, dan NB. Dataset terdiri dari 2000 baris data dengan enam atribut sosial ekonomi yang dibagi menjadi dua kelas: layak dan tidak layak. Data terbagi menjadi 80% pelatihan dan 20% pengujian, dan performa model dievaluasi menggunakan metrik akurasi, presisi, recall, F1-score, dan ROC-AUC. Hasil penelitian menunjukkan bahwa tanpa SMOTE, semua model mengalami bias berat terhadap kelas mayoritas, menghasilkan recall yang sangat rendah untuk kelas minoritas SVM = 0,014; NB = 0,004. Penerapan SMOTE secara signifikan meningkatkan deteksi kelas* *minoritas, dengan RF dan SVM meraih kinerja tertinggi pada F1-score masing-masing 0,820 dan 0,801, serta ROC-AUC sebesar 0,966 dan 0,990. Analisis SHAP mengidentifikasi Jumlah Tanggungan Orang Tua sebagai prediktor paling berpengaruh di semua model, menegaskan perannya yang sentral dalam penilaian kebutuhan finansial. Temuan ini menunjukkan bahwa kombinasi SMOTE dengan model berbasis ensemble atau margin seperti RF dan SVM meningkatkan keadilan dan sensitivitas klasifikasi dalam sistem dukungan pendidikan. Penelitian mendatang disarankan untuk memperluas fitur dengan menambahkan fitur perilaku, akademik, dan regional, menggunakan data lintas institusi, serta mengeksplorasi pendekatan deep learning atau metode resampling lanjutan guna meningkatkan generalisasi dan ketangguhan model.*

***Kata Kunci**: Klasifikasi, Pendidikan Tinggi, Pembelajaran Mesin, SHAP, Uang Kuliah Tunggal*

INTRODUCTION

Higher education is an educational institution that provides higher-level learning services which are the final stage of choice in formal education. These higher education institutions are generally in the form of universities, academies, institutes or high schools. Types of higher education include vocational, academic, and professional. Based on the level, higher education provides diploma, bachelor's, master's, specialist, and doctoral programs. Single Tuition is a student education fee that is set for one semester, and has received a reduction in fees through government subsidies so that there are no more fees outside the cost of Single Tuition. According to Permendikbud Number 55 of 2013, the determination of the cost of Single Tuition is adjusted to the economic status of each student. One of the most common ways for campuses to find out the economic conditions of students is to conduct interviews within candidate students[1].

The implementation of Single Tuition regulations raises symptoms of problems that are generally part of students' critical arguments, for example, the inaccuracy of the economic resources of students' families regarding the nominal Single Tuition group obtained from the institution where they study. Student criticism of the direct decision at the beginning for the Single Tuition group of students who take part in independent selection before students officially pass is perceived as not considering the economic resources of students, parents, or families who finance them. There is the word "single" in the Single Tuition phrase, but all tuitions are not covered by Single Tuition such as study costs, books, hospitality, real work lectures. The last criticism of students is about the lack of transparency of everything covered in Single Tuition[2]. The solution to overcome these problems requires a decision support system based on a classification algorithm that is able to map students' economic conditions objectively and measurably. The output of the classification algorithm can provide fairer and more transparent Single Tuition relief recommendations. This approach also allows institutions to identify students who really need help more accurately and consistently. Classification[3] is an organized grouping process, referring to the technique of arranging data or grouping entities according to predetermined rules. In each context, classification involves features, including class features, that provide groups to the entities[4]. Its application requires finding a model that explains the class features that act as input features. The main goal of classification is to develop a model or algorithm that can predict the class or label of data based on the features it has[5].

The Naïve Bayes (NB) algorithm[6] is one of the most effective and efficient inductive learning algorithms for machine learning and data mining. The NB performance is competitive in the classification process even though it uses the assumption of feature independence (no relationship between features). The NB is used for data classification techniques using probability and statistical methods that predict future opportunities based on previous experience, so it is known as Bayes' Theorem. This theorem is combined with Naïve where it is assumed that the conditions between one feature and another are independent. The NB classification assumes that the presence or absence of certain characteristics in a class has nothing to do with the characteristics of other classes[7].

One of the classification methods is the Support Vector Machine (SVM)[8]. The SVM is a learning system that uses a hypothesis space in the form of linear functions in features that have high dimensions and are trained using a learning algorithm based on optimization theory[6]. The SVM algorithm model is one of the algorithms of the classification method, which works by finding a line (hyperplane) to separate two groups of data[9][10][11][12]. The hyperplane with the best separator can be found by measuring the margin of the hyperplane and finding the maximum point. The kernel must be used to achieve the success of many classification algorithms for linear surfaces. It can

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

be seen that the type of kernel can affect the classification results performed. The hyperplane is the best dividing line between two classes. To find a hyperplane can be done by finding the margin of the hyperplane and finding the maximum point. The margin is the distance between the closest data between two different classes, which is called the support vector[13].

The Random Forest algorithm was used in this study because of its advantages in increasing the accuracy of graduation predictions[5]. This method is also able to identify important variables that affect student graduation and performance evaluation is carried out through a confusion matrix to assess model performance. To facilitate its use, a web application is created using Streamlit which can be accessed online via a browser. The Logistic Regression (LR) can be used to perform multivariate analysis[14]. The LR is better than linear regression, because in making a LR model there are already decisions that must be taken. The LR can be used to simulate respondents' opinions [15] [13].

The LR[16] is a method used to analyze data that describes one response variable (dependent) or more predictor variables. The LR is used when the predictor variable (y) has a categorical or nominal scale consisting of two or more categories. Therefore, this method was developed with the aim of ensuring that, no matter what the estimate is, it is always in the range between 0 and 1. The LR is a statistical method used to solve binary classification problems by estimating the probability of an observation belonging to one of two classes. The LR model is defined by a linear combination of input features, with each feature assigned a weight and a bias term. This model is commonly used for various tasks such as anomaly detection[17].

K-Nearest Neighbor (KNN) functions to classify new subjects based on training sample data and features. This process involves grouping the results of new test samples according to the majority of categories contained in KNN[16]. The KNN selects the k-most comparable data points from class information and utilizes them to make a prediction. The class label for the query instance is determined by the majority vote or the average for regression of the KNN[18]. The models RF, KNN, SVM, LR, and NB were chosen because of their ability to handle small to medium datasets effectively and produce results that are consistent and simple to understand. A thorough comparative study of data features is made possible by these models, which represent a variety of computational approaches, including tree-based, distance, margin, regression, and probabilistic[19][20].

On the other hand, because models like XGBoost and neural networks need a lot of processing power and can overfit small data sets, they are better suited for large and complicated datasets[21][22]. On basic data, a number of studies have demonstrated that deep learning-based models frequently perform on par with or even more steadily than classical models like RF and SVM[23][24]. As a result, the choice of these five traditional models is thought to be the most pertinent and proportionate to the goal of the study, which is to compare how well algorithms perform when classifying student single tuition categories.

The SHapley Additive exPlanations (SHAP) is a model-interpretability approach grounded in cooperative game theory that quantifies the contribution of each feature to a model’s prediction[25]. This method enables researchers to understand both the magnitude and direction of each feature’s influence, facilitating the identification of key factors driving model decisions and uncovering potential prediction errors. Beyond interpretability, SHAP can also be leveraged to enhance model performance, for instance through feature selection strategies informed by feature contribution scores[26]. SHAP not only promotes transparency and accountability in machine learning models but also supports the development of more accurate and reliable predictive systems.

This study aims to compare several machine learning algorithms that are often used in student graduation classification. These algorithms involve NB, SVM, RF, LR and KNN. By conducting an in-depth comparison, this study intends to evaluate the advantages and disadvantages of each algorithm, as well as identify the most appropriate algorithm for student Single Tuition data classification.

MATERIALS AND METHODS

In this study, the data uses secondary data on study by[27] with additional data. This data has 2000 rows and 6 columns. The data categorization of this data is divided into two classes: eligible and ineligible, with the aim of determining Single Tuition relief. Other categorigal data including place of residence and parental occupation, are nominal data, while parental income, number of dependents of parents, and number of vehicles are ratio (scale) data. The data is categorized as the dependent variable of the study and the other five attributes as the independent variables. Data collection is an important stage before experiencing preproseccing. The process that is in the stages preprocessing involves data cleaning, data labeling, data transformation, and data segmentation. Data is always dirty and requires cleaning, including removing extra spaces. Data cleaning is part of the data preparation process, as is the data labeling process. Data transformation to prepare data for further analysis can be done through normalization and encoded data or dimension reduction[28]. Data normalization transforms the values of different data variables into standardized values to avoid bias in the process of discovering new patterns from a data set. Data normalization uses min-max normalization:

$$x_{norm} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \qquad (1)$$

The variable *Xnorm* represents a data sample obtained from the normalization process. The variable X_i denotes the actual data value at the I^{th} observation. The variable X_{min} refers to the smallest value in the actual dataset, while X_{max} represents the largest value in the actual dataset.

The next stage involves dividing or splitting the data into training and testing data, which are applied to achieve optimal accuracy. This critical stage in machine learning model development involves the training and testing data processes[3]. The percentage distribution used in this study is 80% training data and 20% testing data. Repeated k-Fold Cross-Validation (RKCV) is employed in this study as a robust validation strategy to assess and enhance the reliability of machine learning model performance. Specifically, a Repeated Stratified k-Fold Cross-Validation scheme with 5 folds and 10 repetitions is applied, resulting in 50 independent model evaluations to ensure a more stable and unbiased accuracy estimate. By repeatedly partitioning the dataset into balanced training and testing subsets while preserving class distribution, this approach reduces variance in performance metrics and minimizes risks of overfitting[29].

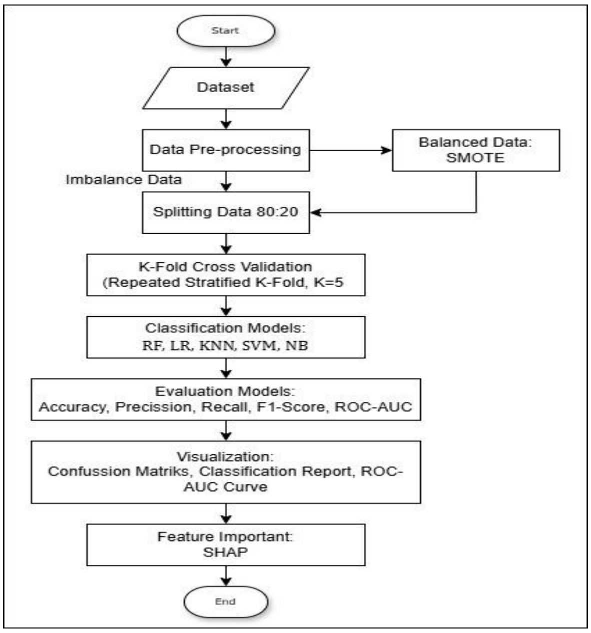
Machine learning method applied after the process of dividing the training data and testing data is completed to build a prediction model. Machine learning methods applied in this study are Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), dan Naïve Bayes (NB). Machine learning methods are suitable for identifying the eligibility of students to receive Single Tuition relief[30][27][31][32][33]. In this study, five classification models were selected to compare their performance in predicting students' eligibility for single tuition relief. The model selection was based on the diversity of learning approaches, encompassing linear, non-linear, probabilistic, and distance-based methods. The RF, LR, SVM, KNN, and NB—were selected based on their complementary learning characteristics and suitability for socio-economic classification tasks such as single tuition eligibility prediction. The rationale for selecting each model is described as follows.

The RF was chosen for its ability to handle heterogeneous data (both categorical and numerical attributes)[34][35] and to model non-linear relationships[36] between socio-economic factors such as income, occupation, and number of dependents. Its ensemble structure[37] also provides interpretability through feature importance[38][39], making it suitable for identifying the most influential socioeconomic indicators.ssss. The LR remains a fundamental tool in binary classification due to its interpretability[40], robustness[41], and effectiveness in small datasets[40][42]. It provides a solid benchmark for evaluating more complex models and continues to be a valuable method in various research and application domains[42][43].

Support Vector Machine (SVM)[7][44] is a supervised machine learning model used for classification or regression. This model works by identifying the optimal hyperplane that can separate data from different classes with the largest margin. The main goal of SVM is to find a hyperplane that clearly classifies data points in an N-dimensional space. The optimal hyperplane maximizes the margin, defined as the distance between the hyperplane and the nearest data point from one of the classes, known as the support vector. A wider margin tends to improve generalization ability. Maximizing the margin ensures that the model achieves the best separation between classes, thereby reducing classification errors. The SVM was selected because it performs well on small to medium-sized datasets[45] with potentially overlapping classes, as is common in socio-economic data where class boundaries are not always distinct. Its kernel functions enable the model to separate classes that are not linearly separable[46][47].

The KNN predicts a class based on the majority of labels from the k nearest neighbors[48]. The KNN is instance-based and no explicit training process. The KNN works by determining the distance from the training or testing sample to the sample data until each post-selection category is scored and a new category is assigned based on certain rules[49]. The implementation of KNN is an effective, intuitive, and simple model that has been widely studied in data and pattern classification[50][51]. The NB is a simple yet powerful classifier[52] that performs well in social classification problems despite its strong independence assumption [53][54]. Its efficiency and effectiveness make it a valuable tool in various applications, particularly when dealing with small datasets and categorical features[55].

The output results of the implemented machine learning method enter the evaluation process based on several performance metrics[4][7] Accuracy, Precision, Recall, F1-Score, and ROC-AUC. The results of the evaluation process will then be displayed in visualizations in the form of bar charts, pie charts, confusion matrices, and classification reports. The classification report visualization is divided into target classes 0 and 1 or binary. The Confusion Matrix visualization is a metric used to assess model performance in more detail in terms of accuracy, precision, recall, F1-score, and ROC-AUC This is possible because the presentation of the values of these metrics can be used to identify the location of prediction errors made by the model. Data visualization is the process of transforming data into visual or graphical displays, such as diagrams and graphs, to facilitate the communication of information to users. Data visualization offers the advantage of facilitating data understanding and interpretation, attracting attention, and avoiding misinterpretation. This understanding refers to how values within the data are distributed. Interpretation of data visualization results can include identifying outliers[4][7].


Start
Dataset
Data Pre-processing
Balanced Data: SMOTE
Imbalance Data
Splitting Data 80:20
K-Fold Cross Validation (Repeated Stratified K-Fold, K=5
Classification Models: RF, LR, KNN, SVM, NB
Evaluation Models: Accuracy, Precission, Recall, F1-Score, ROC-AUC
Visualization: Confussion Matriks, Classification Report, ROC-AUC Curve
Feature Important: SHAP
End

Source: (Research Results, 2025)

Figure 1 Proposed Methodology

SHAP is utilized in this study as an interpretability technique to assess the contribution of each input feature to the model's predictions. Through the computation of feature contribution values, SHAP quantifies how much each variable influences the deviation of a prediction from a baseline or average reference value[55]. SHAP enables detailed analysis at the individual sample level through tools such as force plots, which visually illustrate the positive or negative impact of each feature on a specific prediction[56]. This approach provides a comprehensive understanding of feature importance and model behavior, supporting more transparent and explainable machine learning analysis within the Material and Methods framework. In general, the proposed methodology of this study can be seen in Figure 1.

RESULTS AND DISCUSSION

The application of machine learning methods to single tuition data produces several diagrams and provides a variety of different discussions of each machine learning method. The training and testing results discuss the comparison between each machine learning method in terms of classification. The results obtained are validated based on accuracy. precision, recall, F1- score, ROC-AUC, confusion matrix, dan classification report.

Dataset

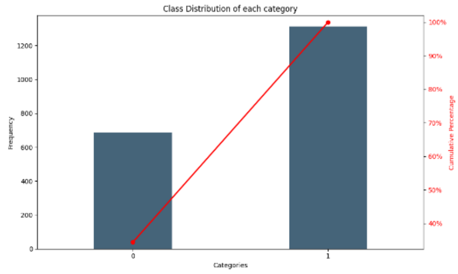
The dataset of Table 1 presents six main attributes which are used in this study: place of residence, parental occupation, parental income, number of dependents of parents, number of vehicles, and eligibility for single tuition relief as the target class. The residence attribute indicates whether a student lives with their parents or lives independently (e.g., in a boarding house or rented accommodation). The occupation attribute represents the employment type of the parents, including categories such as Civil Servants (PNS), Indonesian National Armed Forces (TNI) or Indonesian National Police (POLRI), farmers, fishermen, laborers, housewives, entrepreneurs, and teachers. The income attribute reflects the monthly income of the parents, ranging from IDR 700,000 to IDR 10,000,000. The number of dependents indicates how many family members are financially supported by the parents, ranging from one to five dependents. The number of vehicle attribute captures the number of vehicles owned by the parents, ranging from zero to two.

Table 1. Study Dataset

N o	Place of Residence*)	Parental Occupation	Parental Income (Indonesian Rupiah)	Number of Dependents of Parents	Number of Vehicle	Eligibility for Single Tuition Relief
1	0	PNS	10000000	3	1	0
2	0	TNI/POLRI	8000000	2	2	1
...	...	...	...	...	...	...
1999	0	Farmer	5610587	3	2	0
2000	0	Fisherman	9694991	3	2	1

Source: (Research Results, 2025)

Note: *) The place of residence feature contains a value of 0 or 1, with 0 meaning living with parents, and 1 meaning living in a boarding house or rented accommodation or residence other than with parents.



Source: (Research Results, 2025)

Figure 2. Class Distribution before SMOTE

Preprocessing

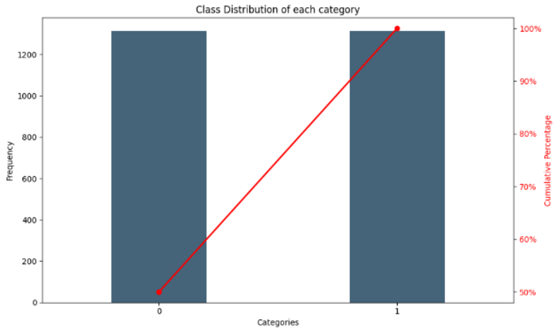
The collected data is the starting point for entering the next stage of preparing data that is free from noise, incompleteness, and inconsistency so that data cleaning is necessary. The collected data must comply with the data mining method and the format of the tool or software which is used by labeling data. The collected data needs to be transformed into values that are within a certain interval which will then be processed for training and testing after the data has been determined as a percentage through data splitting. The data transformation process uses data normalization.

The study data did not contain many errors or biases. The results of the study data identification revealed one data error in row ninety-six, column six. The data error was in the form of entry duplicate data. Handling taken data cleaning extra spaces by taking the first character if there is more than 1 number because entry duplicate data of type string or object. The data used for labeling is taken from the Single Tuition relief eligibility feature. Data labeling uses binary labels 0 and 1. Label 0 indicates eligibility for the Single Tuition relief, while label 1 indicates ineligibility for the Single Tuition relief. The data labeling type for this feature is changed from the integer become a type float. Binary label obtained type float in the form of 0.0 and 1.0.

The next results were obtained from data normalization using min-max normalization after data labeling was completed. The normalized features were parental income, number of dependents of parents, and number of vehicles. Meanwhile, the features parental occupation, place of residence, and eligibility for Single Tuition relief were encoded data. Place of residence and eligibility for Single Tuition relief features encoded data in the form of 0 or 1. Labeling the feature parental occupation results encoded data in lexicographic order, for example Farmer = 0, Fisherman = 1, Housewife = 2, Laborer = 3, PNS = 4, Self-Employed = 5, TNI/POLRI = 6, and Teachers = 7.

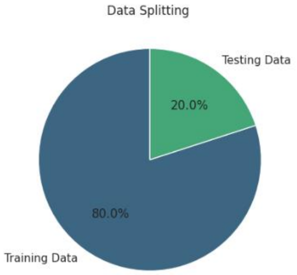
Resampling and Splitting Data

The analysis of data showed that the dataset is imbalanced, and the distribution of each label has that the significant amount of data is label 1. A small distribution is label 0, which can influence this study model biased toward a large enough distribution.



Source: (Research Results, 2025)

Figure 3. Class Distribution after SMOTE



Source: (Research Results, 2025)

Figure 4. Results of Splitting Training Data dan Testing Data

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

To overcome the imbalance data problem, the Synthetic Minority Oversampling Technique (SMOTE) method is applied, as it generates synthetic samples for the minority class to balance the dataset and improve model generalization performance. The class distribution after resampling using SMOTE can be seen in Figure 3. This study separated the data by taking a 4:1 ratio from a total of 100 datasets. A ratio of 4 is equivalent to 80 training data, while a ratio of 1 is equivalent to 20 testing data in Figure 4.

Cross Validation

Table 2 presents 10×5 repeated stratified cross-validation results for five classifiers: RF, LR, SVM, KNN, and NB on an imbalanced dataset, with and without SMOTE. Without SMOTE, all models struggled to detect the minority class, as shown by extremely low recall e.g., SVM: 0.014, LR: 0.021, NB: 0.004, reflecting their inherent bias toward the majority class. High accuracy scores e.g., RF: 0.693, SVM: 0.744, are misleading here, as they mask poor minority-class detection. Metrics like precision, recall, and ROC-AUC offer a clearer picture: low recall signals high false negatives, while ROC-AUC values ranging from 0.525 for RF to 0.778 for LR indicate limited but varying discriminative power still far from optimal.

Table 2. Results of Cross Validation of Classifier Model Performance

Smote Usage	Models	Repeated Stratified 5-Fold Cross-Validation (10x5) – Training Data				
		Accuracy	Precision	Recall	F1-Score	ROC-AUC
None	RF	0.693000 ± 0.020015	0.756651 ± 0.018772	0.785429 ± 0.024806	0.770463 ± 0.015334	0.732525 ± 0.027261
	LR	0.740875 ± 0.019323	0.760334 ± 0.014364	0.884095 ± 0.021415	0.817414 ± 0.013931	0.780541 ± 0.025640
	SVM	0.744062 ± 0.017108	0.762566 ± 0.014442	0.886381 ± 0.019214	0.819663 ± 0.011958	0.778999 ± 0.024515
	KNN	0.712812 ± 0.019195	0.759816 ± 0.016428	0.823143 ± 0.027372	0.789885 ± 0.015035	0.704744 ± 0.025360
	NB	0.734000 ± 0.021053	0.766278 ± 0.015397	0.856000 ± 0.024329	0.808488 ± 0.015855	0.752275 ± 0.031162
With Smote	RF	0.733725 ± 0.021127	0.735637 ± 0.024728	0.731429 ± 0.031616	0.733065 ± 0.022008	0.811912 ± 0.018878
	LR	0.727674 ± 0.021734	0.721650 ± 0.025492	0.743333 ± 0.027409	0.731959 ± 0.020586	0.786314 ± 0.021966
	SVM	0.719051 ± 0.020902	0.707695 ± 0.022127	0.748190 ± 0.033165	0.726933 ± 0.021314	0.784990 ± 0.021996
	KNN	0.727487 ± 0.017097	0.736027 ± 0.020814	0.711143 ± 0.032165	0.722830 ± 0.019210	0.781450 ± 0.017413
	NB	0.676180 ± 0.024709	0.646555 ± 0.022576	0.779714 ± 0.030799	0.706623 ± 0.021895	0.751340 ± 0.025158

Source: (Research Results, 2025)

SMOTE significantly improved minority class detection, notably boosting recall and ROC-AUC. For example, SVM's recall rose from 0.014 to 0.190, and RF's from 0.084 to 0.373 that more than doubling. ROC-AUC also increased across most models, peaking at 0.990 on SVM and 0.966 on RF, reflecting stronger class discrimination. However, this gain often came with reduced precision e.g., RF from 0.756 to 0.707; and SVM from 0.762 to 0.695, indicating more false positives. This trade-off is acceptable in domains like healthcare or education, where missing true positives, false negatives carries higher cost than false alarms.

Among the five models, Random Forest (RF) and Support Vector Machine (SVM) responded most robustly to SMOTE, RF achieved the highest recall gain from 0.084 to 0.373, while SVM attained the highest ROC-AUC with 0.990. RF's strength stems from its ability to handle non-linear patterns and noise via ensemble learning, while SVM benefits from SMOTE's synthetic samples, which enhance margin-based separation in feature space. KNN and LR showed moderate improvement, whereas Naïve Bayes remained largely unchanged, likely due to its restrictive conditional independence assumption, limiting adaptability to SMOTE-induced distributions.

Practically, this suggests that for educational or similar high-stakes domains with rare positive cases, combining SMOTE with ensemble or margin-based models like RF and SVM offers the most effective strategy for improving minority-class detection without compromising model stability. The consistently low standard deviations across metrics further confirm the reliability and reproducibility of these improvements across validation folds.

Classification Model

Implementation machine learning classifier applying the model random forest (RF), logistic regression (LR), k-nearest neighbors (KNN), support vector machine (SVM), and naïve bayes (NB) use the GaussianNB library type. The classification output of the models systematically evaluates and compares its performance explicitly using accuracy, precision, recall, F1- score and confusion matrix as well as classification report. The summary of classification performance of all validated models using metrices evaluation accuracy is as compiled in Table 3.

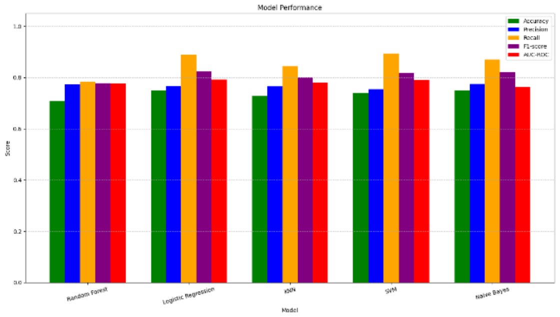
Table 3. Comparison of Classifier Model Performance

Smote Usage	Model	Evaluation Metrics				
		Accu-racy	Preci-sion	Recall	F1-Score	ROC-AUC
None	RF	0.7075	0.7736	0.7824	0.7780	0.7771
	LR	0.7500	0.7664	0.8893	0.8233	0.7922
	KNN	0.7275	0.7647	0.8435	0.8022	0.7812
	SVM	0.7400	0.7548	0.8931	0.8182	0.7912
	NB	0.7500	0.7755	0.8702	0.8201	0.7617
With Smote	RF	0.7295	0.7419	0.7023	0.7216	0.8254
	LR	0.7352	0.7356	0.7328	0.7342	0.8077
	KNN	0.7238	0.7312	0.7061	0.7184	0.7848
	SVM	0.7200	0.7138	0.7328	0.7232	0.8045
	NB	0.7124	0.6844	0.7863	0.7318	0.7710

Source: (Research Results, 2025)

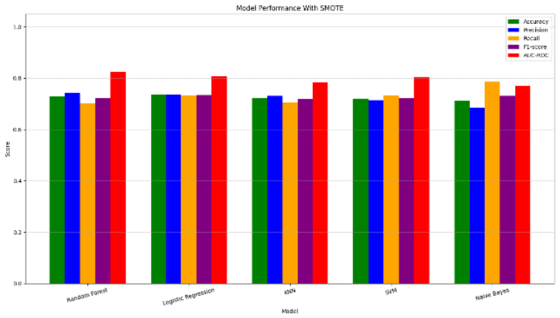
Table 3 evaluates five classifier models under imbalanced conditions and after SMOTE, using Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Without SMOTE, models exhibit deceptively high accuracy (0.725–0.750) but critically low recall, revealing poor minority-class detection, a key limitation in high-stakes domains like education or healthcare. RF outperforms others without SMOTE, achieving the highest F1-Score = 0.7780 and ROC-AUC =0.7771, reflecting its robustness to imbalance via ensemble learning. Upon applying SMOTE, all models show improved recall and F1-Score, with RF again leading F1 score = 0.8201; Recall gain +0.0878, while SVM achieves the highest ROC-AUC = 0.8045, indicating superior class separability. Although precision slightly declines e.g., SVM from 0.7548 to 0.7382, the trade-off favors higher sensitivity, an acceptable compromise where false negatives are costlier than false positives.

The differential response to SMOTE underscores algorithmic suitability: RF and SVM benefit most due to their capacity to leverage expanded decision boundaries and handle non-linear patterns, while NB's performance remains stagnant, likely constrained by its conditional independence assumption. KNN and LR show modest gains, suggesting limited adaptability to synthetic data. These findings advocate for combining SMOTE with ensemble or margin-based learners in applications demanding high minority-class sensitivity. Furthermore, the visualization of model performance comparison and class balance improvement after applying SMOTE can be seen in Figure 5 and Figure 6, providing a clearer understanding of how resampling affects classifier behavior.



Source: (Research Results, 2025)

Figure 5. Model Performance before SMOTE



Source: (Research Results, 2025)

Figure 6. Model Performance after SMOTE

Table 4 presents per-class performance metrics by Precision, Recall, F1-Score for binary classification models under imbalanced conditions, a critical approach since aggregate metrics like accuracy can be misleading. By evaluating classes 0 and 1 separately, the table reveals each model's sensitivity to minority instances, aligning with research priorities in fairness and risk-sensitive applications. Precision measures prediction reliability, Recall captures detection capability, and F1-Score balances both collectively offering a nuanced view of model behavior beyond overall accuracy.

Without SMOTE, model performance varies significantly across classes and algorithms. RF exhibits high recall for class 1 with 0.78 but low for class 0 with 0.57, reflecting majority-class bias. LR shows moderate class balance but suboptimal F1-Scores from 0.57 to 0.82. SVM and KNN display

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7294

instability SVM has low recall for class 0 = 0.45, while KNN shows weak precision for class 0 = 0.63. NB performs well on class 0 with recall = 0.84 but poorly on class 1 with recall = 0.52, suggesting its conditional independence assumption limits adaptability to skewed data distributions.

SMOTE significantly improves cross-class consistency, particularly in recall. For RF, recall for class 0 rises from 0.57 to 0.76, while class 1 slightly drops from 0.78 to 0.76, indicating balanced sensitivity. Similar trends occur in LR and SVM, minority-class recall improves at the cost of modest precision loss, a justifiable trade-off in practice. KNN and NB also show improved recall for class 1, though NB remains biased toward class 0. Overall F1-Score gains e.g., RF class 0 from 0.57 to 0.74 confirm SMOTE enhances both detection and precision-recall balance.

These results support that SMOTE effectively reshapes the feature space, enabling non-parametric and ensemble models, particularly RF and LR to learn fairer decision boundaries. Their strong response aligns with literature highlighting their flexibility and generalization capacity in leveraging synthetic samples. In contrast, NB rigid independence assumption limits its adaptability. Practically in education, prioritizing SMOTE with ensemble or linear models ensures predictive fairness and minimizes false negatives.

Table 4. Performance Models of Classification Reports

Smote Usage	Models	Target Class	Evaluation Metrics		
			Precision	Recall	F1-Score
None	RF	0	0.58	0.57	0.57
		1	0.77	0.78	0.78
	LR	0	0.70	0.49	0.57
		1	0.77	0.89	0.82
	SVM	0	0.69	0.45	0.54
		1	0.75	0.89	0-82
	KNN	0	0.63	0.51	0.56
		1	0.76	0.84	0.80
	NB	0	0.68	0.52	0.59
		1	0.78	0.87	0.82
With Smote	RF	0	0.72	0.76	0.74
		1	0.74	0.70	0.72
	LR	0	0.73	0.74	0.74
		1	0.74	0.73	0.73
	SVM	0	0.73	0.71	0.72
		1	0.71	0.73	0.72
	KNN	0	0.72	0.74	0.73
		1	0.73	0.71	0.72
	NB	0	0.75	0.64	0.69
		1	0.68	0.79	0.73

Source: (Research Results, 2025)

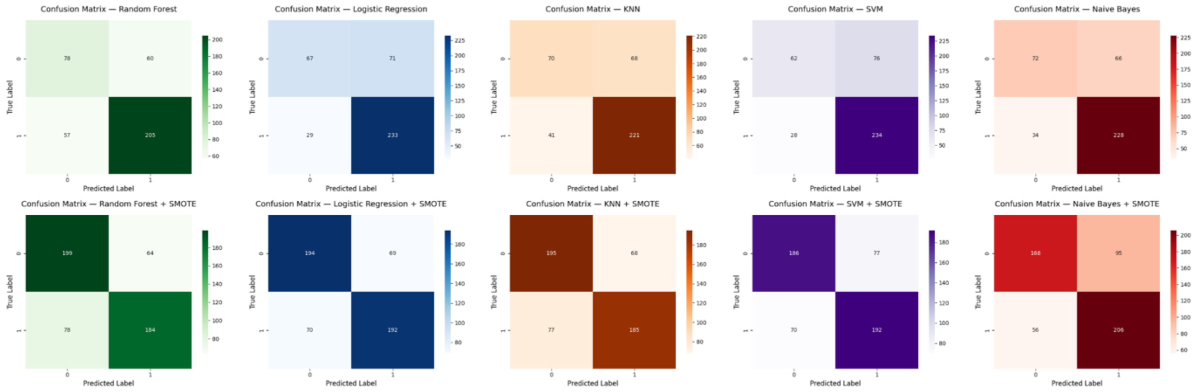
Figure 7 presents the visualizations display confusion matrices for five machine learning classifiers both before and after applying the SMOTE oversampling technique. Each matrix illustrates the distribution of true positives, true negatives, false positives, and false negatives across two classes labeled with 0 and 1. These matrices serve as fundamental diagnostic tools in classification tasks, enabling granular assessment of model behavior beyond aggregate metrics like accuracy. The color intensity reflects prediction frequency, with darker shades indicating higher counts, allowing immediate visual identification of dominant misclassification patterns.

Without SMOTE, all models exhibit a strong bias toward predicting class 1 as the majority class, as evidenced by high values along the diagonal for class 1 e.g., RF = 205 true positives; SVM = 234; NB = 228 but significantly lower true positives for class 0 e.g., RF = 78; SVM = 62; NB = 72. This imbalance is particularly severe for LR and KNN, which misclassify many instances of class 0 as class 1 e.g., LR = 71 false positives; KNN = 68. Such behavior confirms the well-documented challenge of class imbalance in supervised learning, where algorithms inherently favor the majority class unless explicitly corrected, resulting in high false-negative rates for minority-class instances.

SMOTE substantially improves model calibration, particularly in reducing false negatives for class 0. For instance, RF true positives for class 0 jump from 78 to 199, while its false negatives drop from 57 to 78, a clear shift toward balanced detection. Similarly, SVM class 0 true positives increase from 62 to 186, and NB rise from 72 to 168. Although some models like LR and KNN still show residual bias, LR predicts 194 instances of class 0 incorrectly as class 1, the overall trend demonstrates that synthetic oversampling effectively redistributes decision boundaries, enhancing sensitivity to the minority class without completely sacrificing specificity.

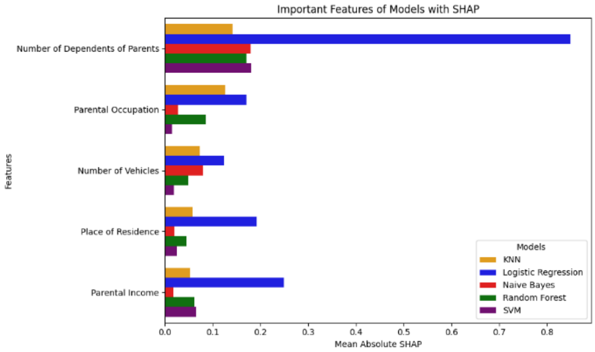
Model responses to SMOTE vary based on algorithmic structure. RF and SVM both capable of handling non-linear boundaries benefit most, showing dramatic increases in minority-class detection with relatively controlled false-positive inflation. NB, despite its naive independence assumption, also improves markedly, suggesting SMOTE's synthetic samples partially compensate for its structural limitations. In contrast, LR and KNN show more modest gains, likely due to their linear or distance-based decision mechanisms, which may struggle to fully exploit the expanded feature space introduced by SMOTE. This differential response underscores the importance of selecting models compatible with resampling strategies for optimal performance under imbalance.

The consistent improvement in detecting class 0 across all models post-SMOTE validates its utility as a preprocessing intervention. From a practical standpoint, Random Forest emerges as the most robust choice, achieving the highest balance between sensitivity and specificity after SMOTE. For applications demanding interpretability, Logistic Regression remains viable if paired with SMOTE, though it requires careful threshold tuning. Ultimately, this visualization reinforces that effective imbalanced classification demands not only appropriate algorithms but also strategic data-level interventions — and that model evaluation must prioritize per-class performance over global accuracy.



Source: (Research Results, 2025)

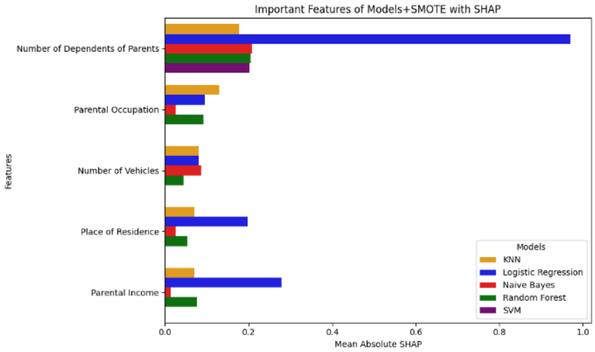
Figure 7. Confusion Matrix each Models before and after SMOTE



Source: (Research Results, 2025)

Figure 8. SHAP result of Models without SMOTE

Figure 8 shows taht the SHAP result reveals that "Number of Dependents of Parents" is the most influential feature across nearly all models, particularly for Random Forest and SVM, which assign it the highest SHAP values >0.7, indicating its dominant role in driving predictions. In contrast, features such as "Parental Income" and "Number of Vehicles" exhibit consistently low importance across models, suggesting limited predictive power. Notably, Logistic Regression and Naïve Bayes show distinct sensitivity to "Place of Residence" and "Parental Occupation," respectively, highlighting how different algorithmic assumptions influence feature attribution. The visual comparison underscores model-specific interpretability: while ensemble and kernel-based models (RF, SVM) prioritize family size-related variables, linear and probabilistic models (LR, NB) respond more strongly to contextual or categorical features offering valuable insights for domain experts seeking to understand and trust model decisions in applications such as educational or social risk prediction.



Source: (Research Results, 2025)

Figure 9. SHAP result of Models with SMOTE

Figure 9 presents that the SHAP result of models with SMOTE confirms "Number of Dependents of Parents" as the top predictor across all five classifiers, especially for RF and SVM with SHAP > 0.8, indicating its robust, imbalance-resistant discriminative power. "Place of Residence" and "Parental Occupation" hold moderate importance, particularly for LR and NB, while "Parental Income" and "Number of Vehicles" consistently show minimal impact, suggesting weak

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

predictive relevance. The stable feature ranking post-SMOTE implies resampling enhances model stability without altering core feature relationships, allowing true structural signals to surface more clearly.

The implementation of SMOTE does not significantly shift the hierarchy of feature importance, reinforcing that "Number of Dependents of Parents" is the most critical and stable predictor across diverse algorithmic architectures. This finding underscores the value of domain-specific knowledge in feature engineering and model interpretation, even when addressing class imbalance, core socio-demographic indicators retain their explanatory power. For practical deployment, especially in educational or social welfare contexts, prioritizing interventions or assessments based on family dependency metrics may yield higher predictive accuracy and actionable insights. Furthermore, the consistent low importance of material assets like "Parental Income" and "Number of Vehicles" suggests that resource-based proxies may be less effective than relational or structural indicators in this setting, guiding future data collection and modeling strategies toward more meaningful, human-centered variables.

CONCLUSSION

This study evaluates the performance of five classification models, RF, LR, KNN, SVM, and NB in predicting student tuition categories based on six socio-economic attributes, with a focus on addressing class imbalance using the SMOTE technique. Results show that without SMOTE, all models exhibit bias toward the majority class, yielding low recall for the minority class, a critical limitation in predictive applications sensitive to rare events. The application of SMOTE significantly improves minority-class detection, particularly through increased recall and F1-Score, with RF and SVM demonstrating the strongest and most stable responses.

Confusion matrix and SHAP analyses confirm that "Number of Dependents of Parents" is the most dominant and consistent feature across all models, even after oversampling, highlighting its high relevance in assessing students' financial capacity. Although global accuracy appears high, per-class metrics, especially recall and F1-Score provide more meaningful insights in imbalanced learning contexts. Overall, combining SMOTE with ensemble or margin-based models such as RF and SVM is recommended as the optimal strategy to ensure fair, sensitive, and reliable predictions.

Future research should expand beyond the current six socio-economic features by incorporating behavioral, psychological, and academic variables, alongside household assets and regional indicators to improve model generalization and accuracy. Leveraging larger, multi-institutional datasets and exploring advanced resampling or deep learning techniques will further enhance robustness, reduce overfitting, and support more externally valid, policy-relevant predictions of student tuition categories.

REFERENCES

[1] N. T. Syam, Irmawati, and Z. Saharuna, "Penerapan Machine Learning untuk Mengatasi Ketimpangan Data dalam Menentukan Klasifikasi Uang Kuliah Tunggal (UKT)," *Journal of Informatics and Computer Engineering Research*, vol. 1, no. 1, pp. 7–14, Jun. 2024, doi: 10.31963/jicer.v1i1.4921.

[2] M. Ardiansyah, T. Suharto, and A. S. Farid, "Upaya Penanganan Uang Kuliah Tunggal (UKT) Bermasalah bagi Mahasiswa yang tidak Mampu pada Perguruan Tinggi," *JIIP - Jurnal Ilmiah Ilmu Pendidikan*, vol. 5, no. 10, pp. 4432–4441, Oct. 2022, doi: 10.54371/jiip.v5i10.1036.

[3] R. A. Sigit, Z. Kurniawan, and R. Rahmaddeni, "Komparasi Algoritma Machine Learning untuk Klasifikasi Kelulusan Mahasiswa," *JSR : Jaringan Sistem Informasi Robotik*, vol. 8, no. 1, pp. 108–113, 2024.

[4] I. K. N. Ananda, N. P. N. P. Dewi, N. W. Marti, and L. J. E. Dewi, "Klasifikasi Multilabel pada Gaya Belajar Siswa Sekolah Dasar Menggunakan Algoritma Machine Learning," *Journal of Applied Computer Science and Technology*, vol. 5, no. 2, pp. 144–154, Dec. 2024, doi: 10.52158/jacost.v5i2.940.

[5] M. Putra and E. Harahap, "Machine Learning pada Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Random Forest," *Jurnal Riset Matematika*, vol. 4, no. 2, pp. 127–136, Dec. 2024, doi: 10.29313/jrm.v4i2.5102.

[6] J. Jefri and Z. Fatah, "Klasifikasi Data Mining untuk Memprediksi Kelulusan Mahasiswa Menggunakan Metode Naive Bayes," *Jurnal Ilmiah Multidisiplin Ilmu*, vol. 2, no. 1, pp. 29–37, Feb. 2025, doi: 10.69714/mhjq1v85.

[7] M. Fadhilla, R. Wandri, A. Hanafiah, P. R. Setiawan, Y. Arta, and S. Daulay, "Analisis Performa Algoritma Machine Learning untuk Identifikasi Depresi pada Mahasiswa," *Journal of Informatics Management and Information Technology* , vol. 5, no. 1, pp. 40–47, Jan. 2025.

[8] S. S. M. Putri, M. Arhami, and H. Hendrawaty, "Penerapan Metode SVM pada Klasifikasi Kualitas Air," *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 3, no. 2, p. 93, Nov. 2023, doi: 10.30811/jaise.v3i2.4630.

[9] I. M. D. P. Asana and N. P. D. T. Yanti, "Sistem Klasifikasi Pengajuan Kredit dengan Metode Support Vector Machine (SVM)," *Jurnal Sistem Cerdas*, vol. 6, no. 2, pp. 123–133, Aug. 2023.

[10] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, "Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 153–160, Oct. 2023, doi: 10.57152/malcom.v3i2.897.

[11] S. Wahyuni, N. Sutriningsih, and S. Rahayu, "Penerapan Media GeoGebra pada Pembelajaran Matematika," *Cartesian: Jurnal Pendidikan Matematika*, vol. 2, no. 2, pp. 234–240, Apr. 2023, doi: 10.33752/cartesian.v2i2.3508.

[12] M. Mardewi, N. Yarkuran, S. Sofyan, and F. Aziz, "Klasifikasi Kategori Obat Menggunakan Algoritma Support Vector Machine," *Journal Pharmacy and Application of Computer Sciences (JOPACS)*, vol. 1, no. 1, pp. 27–32, Feb. 2023.

[13] O. Daswati, I. Indahwati, E. Erfiani, A. Fitrianto, and M. A. Aliu, "Model Klasifikasi Regresi Logistik Biner untuk Laporan Masyarakat di Ombudsman Republik Indonesia," *Lebesgue: JurnalIlmiahPendidikan Matematika, Matematika dan Statistika*, vol. 5, no. 2, pp. 964–973, Aug. 2024.

[14] H. Achmadi, I. Fatmawati, and S. Samuel, "Karakteristik Siswa Siswi SMA yang Menentukan Pemilihan Perguruan Tinggi Swasta di Indonesia dengan Menggunakan Logistik Regresi," in *6th NCBMA 2023 "Business Analytics and Artificial Intelligence for Supporting Business Sustainability"* , Tangerang: Universitas Pelita Harapan, Indonesia, May 2023, pp. 251–259.

[15] R. Aristawidya, I. Indahwati, E. Erfiani, A. Fitrianto, and M. A. A, "Perbandingan Analisis Regresi Logistik Biner dan Naïve Bayes Classifier untuk Memprediksi Faktor Resiko Diabetes," *Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 5, no. 2, pp. 782–794, Aug. 2024.

[16] D. Nasien, R. Darwin, A. Cia, and et al., "Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes," *Jekin-Jurnal Teknik Informatika*, vol. 4, no. 1, pp. 10–17, Feb. 2024.

[17] V. Vajrobol, B. B. Gupta, and A. Gaurav, "Mutual information based logistic regression for phishing URL detection," *Cyber Security and Applications*, vol. 2, p. 1, Mar. 2024, doi: 10.1016/j.csa.2024.100044.

[18] S. Zhang, "Challenges in KNN Classification," *IEEE Trans Knowl Data Eng*, vol. 34, no. 10, pp. 4663–4675, Oct. 2022, doi: 10.1109/TKDE.2021.3049250.

[19] M. F. Kurniawan and D. A. Megawaty, "Comparison of Logistic Regression, Random Forest, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) Algorithms in Diabetes Prediction," *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2154–2162, Oct. 2025, doi: 10.30871/jaic.v9i5.9815.

[20] M. AlShaikh, Y. Alrajeh, S. Alamri, S. Melhem, and A. Abu-Khadrah, "Supervised Methods of Machine Learning for Email Classification: A Literature Survey," *Systems Science & Control Engineering*, vol. 13, no. 1, pp. 1–15, Dec. 2025, doi: 10.1080/21642583.2025.2474450.

[21] M. Zou, W.-G. Jiang, Q.-H. Qin, Y.-C. Liu, and M.-L. Li, "Optimized XGBoost Model with Small Dataset for Predicting Relative Density of Ti-6Al-4V Parts Manufactured by Selective Laser Melting," *Materials*, vol. 15, no. 15, p. 5298, Aug. 2022, doi: 10.3390/ma15155298.

[22] A. Shmuel, O. Glickman, and T. Lazebnik, "A Comprehensive Benchmark of Machine and Deep Learning Models on Structured Data for Regression and Classification," *Neurocomputing*, vol. 655, pp. 1–14, Nov. 2025, doi: 10.1016/j.neucom.2025.131337.

[23] R. Shwartz-Ziv and A. Armon, "Tabular Data: Deep Learning is not All You Need,"

Information Fusion, vol. 81, pp. 84–90, May 2022, doi: 10.1016/j.inffus.2021.11.011.

[24] R. Guetari, H. Ayari, and H. Sakly, “Computer-Aided Diagnosis Systems: A Comparative Study of Classical Machine Learning Versus Deep Learning-Based Approaches,” *Knowl Inf Syst*, vol. 65, no. 10, pp. 3881–3921, Oct. 2023, doi: 10.1007/s10115-023-01894-7.

[25] Y. Arslan *et al.*, “Towards Refined Classifications Driven by SHAP Explanations,” 2022, pp. 68–81. doi: 10.1007/978-3-031-14463-9_5.

[26] L. Bernal, G. Rastelli, and L. Pinzi, “Improving Machine Learning Classification Predictions through SHAP and Features Analysis Interpretation,” *J Chem Inf Model*, Oct. 2025, doi: 10.1021/acs.jcim.5c02015.

[27] R. Rina, M. H. Puspita, N. Ayu, and R. A. Saputra, “Klasifikasi Keringanan UKT Mahasiswa UHO Menggunakan K-Nearest Neighbor (KNN),” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 6, pp. 11939–11945, Nov. 2024, doi: 10.36040/jati.v8i6.11757.

[28] F. Sulianta, *Basic Data Mining from A to Z*, 1st ed. Bandung, 2023.

[29] D. Abriha, P. K. Srivastava, and S. Szabó, “Smaller is Better? Unduly Nice Accuracy Assessments in Roof Detection Using Remote Sensing Data with Machine Learning and k-Fold Cross-Validation,” *Heliyon*, vol. 9, no. 3, pp. 1–17, 2023, doi: 10.1016/j.heliyon.2023.e14045.

[30] E. Novianto, A. Hermawan, and D. Avianto, “Perbandingan Metode K-Nearest Neighbor dan Support Vector Machine untuk Memprediksi Penerima Beasiswa Keringanan UKT,” *Jurnal Media Informatika Budidarma*, vol. 8, no. 1, p. 654, Feb. 2024, doi: 10.30865/mib.v8i1.6913.

[31] A. Khaidar, M. Arhami, and M. Abdi, “Application of the Random Forest Method for UKT Classification at Politeknik Negeri Lhokseumawe,” *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 4, no. 2, p. 94, Nov. 2024, doi: 10.30811/jaise.v4i2.6131.

[32] R. Susetyoko, W. Yuwono, E. Purwantini, and N. Ramadijanti, “Perbandingan Metode Random Forest, Regresi Logistik, Naïve Bayes, dan Multilayer Perceptron Pada Klasifikasi Uang Kuliah Tunggal (UKT),” *Jurnal Infomedia*, vol. 7, no. 1, p. 8, Jun. 2022, doi: 10.30811/jim.v7i1.2916.

[33] R. C. A. Fajardo, F. B. Yara, R. F. Ardeña, M. K. L. Hernandez, and J. C. T. Arroyo, “A Data-Driven Approach in Predicting Scholarship Grants of a Local Government Unit in the Philippines Using Machine Learning,” *International Journal of Engineering Trends and Technology*, vol. 72, no. 6, pp. 74–81, Jun. 2024, doi: 10.14445/22315381/IJETT-V72I6P108.

[34] A. Gunakala and A. H. Shahid, “A Comparative Study on Performance of Basic and Ensemble Classifiers with Various Datasets,” *Applied Computer Science*, vol. 19, no. 1, pp. 107–132, Mar. 2023, doi: 10.35784/acs-2023-08.

[35] J. Hu and S. Szymczak, “A Review on Longitudinal Data Analysis with Random Forest,” *Brief Bioinform*, vol. 24, no. 2, pp. 1–11, Mar. 2023, doi: 10.1093/bib/bbad002.

[36] J. Zhao, C.-D. Lee, G. Chen, and J. Zhang, “Research on the Prediction Application of Multiple Classification Datasets Based on Random Forest Model,” in *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, IEEE, Jul. 2024, pp. 156–161. doi: 10.1109/ICPICS62053.2024.10795875.

[37] G. S. Jamnal, “Instils Trust in Random Forest Predictions,” in *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, Oct. 2023, pp. 1–9. doi: 10.1109/DSAA60987.2023.10302640.

[38] A. Yaqoob *et al.*, “SGA-Driven Feature Selection and Random Forest Classification for Enhanced Breast Cancer Diagnosis: A Comparative Study,” *Sci Rep*, vol. 15, no. 1, Mar. 2025, doi: 10.1038/s41598-025-95786-1.

[39] Y. Miao and Y. Xu, “Random Forest-Based Analysis of Variability in Feature Impacts,” in *2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA)*, IEEE, Jun. 2024, pp. 1130–1135. doi: 10.1109/ICIPCA61593.2024.10708791.

[40] J. K. Harris, “Primer on Binary Logistic Regression,” *Fam Med Community Health*, vol. 9, no. 1, pp. 1–7, Dec. 2021, doi: 10.1136/fmch-2021-001290.

[41] D. Cornilly, L. Tubex, S. Van Aelst, and T. Verdonck, “Robust and Sparse Logistic Regression,” *Adv Data Anal Classif*, vol. 18, no. 3, pp. 663–679, Sep. 2024, doi: 10.1007/s11634-023-00572-4.

[42] S. Naik, P. Kumar, S. Saha, S. Das Bairagya, D. Rawat, and S. K. Baliarsingh, “Predictive Healthcare Analytics: A Multidisease Approach Using Logistic Regression,” in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10725194.

[43] R. Suryawanshi, V. Kulkarni, P. Ghule, K. Patil, H. Patil, and Y. Manala, “Brain Stroke Prediction Using Logistic Regression with Logarithmic Transform,” in *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, IEEE, Oct. 2024, pp. 873–877. doi: 10.1109/ICSES63445.2024.10763034.

[44] A. Putri *et al.*, “Komparasi Algoritma K-NN Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 1, pp. 20–26, May 2023, doi: 10.57152/malcom.v3i1.610.

[45] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, “An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review,” *Information*, vol. 15, no. 4, pp. 1–36, Apr. 2024, doi: 10.3390/info15040235.

[46] Y. Chen, A. Zhu, and Q. Zhao, “Rolling Bearing Fault Diagnosis Based On Flock Optimization Support Vector Machine,” in *2023 IEEE 7th Information Technology and Mechatronics Engineering Conference (ITOEC)*, IEEE, Sep. 2023, pp. 1700–1703. doi: 10.1109/ITOEC57671.2023.10292080.

[47] Z. Jun, “The Development and Application of Support Vector Machine,” *J Phys Conf Ser*, vol. 1748, no. 5, pp. 1–6, Jan. 2021, doi: 10.1088/1742-6596/1748/5/052006.

[48] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, “Comparative Performance Analysis of K-Nearest Neighbour (KNN) Algorithm and Its Different Variants for Disease Prediction,” *Sci Rep*, vol. 12, no. 6256, pp. 1–11, Apr. 2022, doi: 10.1038/s41598-022-10358-x.

[49] H. Vega-Huerta *et al.*, “K-Nearest Neighbors Model to Optimize Data Classification According to the Water Quality Index of the Upper Basin of the City of Huarmey,” *Applied Sciences*, vol. 15, no. 18, pp. 1–19, Sep. 2025, doi: 10.3390/app151810202.

[50] A. A. Amer, S. D. Ravana, and R. A. A. Habeeb, “Effective k-Nearest Neighbor Models for Data Classification Enhancement,” *J Big Data*, vol. 12, no. 86, pp. 1–41, Apr. 2025, doi: 10.1186/s40537-025-01137-2.

[51] R. K. Halder, M. N. Uddin, Md. A. Uddin, S. Aryal, and A. Khraisat, “Enhancing K-Nearest Neighbor Algorithm: A Comprehensive Review and Performance Analysis of Modifications,” *J Big Data*, vol. 11, no. 113, pp. 1–55, Aug. 2024, doi: 10.1186/s40537-024-00973-y.

[52] F. Ramadhani, A.-K. Al-Khowarizmi, and I. P. Sari, “Improving the Performance of Naïve Bayes Algorithm by Reducing the Attributes of Dataset Using Gain Ratio and Adaboost,” in *2021 International Conference on Computer Science and Engineering (IC2SE)*, IEEE, Nov. 2021, pp. 1–5. doi: 10.1109/IC2SE52832.2021.9792027.

[53] D. Prabha, J. Aswini, B. Maheswari, R. S. Subramanian, R. Nithyanandhan, and P. Girija, “A Survey on Alleviating the Naive Bayes Conditional Independence Assumption,” in *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, IEEE, Nov. 2022, pp. 654–657. doi: 10.1109/ICAISS55157.2022.10011103.

[54] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, “Variable Selection for Naïve Bayes Classification,” *Comput Oper Res*, vol. 135, pp. 1–11, Nov. 2021, doi: 10.1016/j.cor.2021.105456.

[55] B. Phatcharathada and P. Srisuradetchai, “Randomized Feature and Bootstrapped Naive Bayes Classification,” *Applied System Innovation*, vol. 8, no. 4, pp. 1–20, Jul. 2025, doi: 10.3390/asi8040094.

[56] A. S. Antonini *et al.*, “Machine Learning Model Interpretability Using SHAP Values: Application to Igneous Rock Classification Task,” *Applied Computing and Geosciences*, vol. 23, p. 100178, Sep. 2024, doi: 10.1016/j.acags.2024.100178.