INTENSIF, Vol.9 No.2 August 2025
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

Implementation of SMOTE to Improve the Performance of Random Forest Classification in Credit Risk Assessment in Banking

Received:
29 March 2025
Accepted:
8 June 2025
Published:
12 July 2025

[1]Nafa Nur Adifia Nanda, [2*]Yuniar Farida, [3]Wika Dianita Utami
[1,2,3]Mathematics Department, UIN Sunan Ampel Surabaya
E-mail: [1]adifianafa13@gmail.com, [2]yuniar_farida@uinsa.ac.id, [3]wikadianita@uinsa.ac.id
*Corresponding Author

Abstract— **Background:** Credit is essential in banking operations, facilitating investment, corporate expansion, and financial satisfaction. Credit risk may emerge if the borrower defaults on payment commitments. **Objective:** This study aims to evaluate an individual's creditworthiness by classifying and assessing their eligibility for credit. **Methods:** This study uses the Random Forest technique to categorize credit risk evaluation. Random Forest is a decision tree technique recognized for its high accuracy in data classification, utilizing an ensemble method of many decision trees. Before executing the classification process, issues frequently arise when data cannot be directly processed due to class imbalance. This study employs the SMOTE (Synthetic Minority Over-sampling Technique) algorithm to address class imbalance. The SMOTE algorithm is a method that emphasizes oversampling and is designed to augment the data in the minority class by generating synthetic data that aligns with the minority class data. The findings indicated that the ideal ratio for partitioning training and testing data was 80:20, and implementing the SMOTE technique within Random Forest enhanced performance assessment. **Results:** This research contributes to improving the accuracy of credit risk classification using the Random Forest algorithm, which effectively handles complex data and is supported by the implementation of SMOTE to overcome the class imbalance in the data. The classification accuracy value rose from 91.54% to 94.41%. The precision value rose from 90.83% to 97.03%, while the recall value increased from 60.26% to 91.55%. **Conclusion**: This method helps banks identify high-risk debtors more objectively and efficiently and supports appropriate credit decision-making.

Keywords— Classification; Credit Assesment; Credit Risk; Imbalance Data; Random Forest; SMOTE

Corresponding Author:
Yuniar Farida,
Department of Mathematics,
UIN Sunan Ampel Surabaya,
Email: yuniar_farida@uinsa.ac.id
Orchid ID: https://orcid.org/0000-0001-8666-4980

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

I. INTRODUCTION

As the main instrument in banking activities, credit is often given to various parties to encourage investment, business growth, or fulfill other financial needs [1]. However, credit is also risky: the borrower cannot meet the payment obligation according to the agreement. Changes in economic conditions can trigger these problems, the borrower's internal management problems, or the borrower's inability to fulfill the payment contract [2]. Uncontrolled credit risk can result in financial losses for financial institutions, damage market stability and confidence, and create a financial burden for the government. Thus, credit risk management and monitoring borrowers are essential aspects of banking operations [3].

The emergence of credit problems is often caused by a lack of caution in the analysis of credit grants, hasty decisions, and a lack of evaluation of the financial condition of prospective borrowers [4]. This results in the provision of credit to parties who cannot return it, thus adversely affecting financial institutions and resulting in non-performing loans. Therefore, financial institutions reduce credit risk by conducting creditworthiness assessments for creditors with in-depth consideration [5]. Based on past data on loans that pass the selection, classification can be carried out to obtain information on what factors can be considered in creditworthiness. Classification is carried out systematically based on relevant data. Classification using Machine learning is needed in credit risk assessment because it can automatically analyze complex historical data patterns, resulting in more accurate, faster, and more objective risk predictions than traditional methods. With the ability to learn from big data and continuously learn from new data, machine learning helps banks proactively identify high-risk potential borrowers, minimize potential bad debts, and increase the efficiency of credit decision-making [6][7][8].

The classification method is a data analysis technique for identifying patterns or relationships, allowing for the grouping of data into predefined categories [9][10][11]. Random Forest is a technique utilized for classification, characterized as a decision tree method that demonstrates a high degree of accuracy in data classification [12][13]. Random Forest is an ensemble approach that combines multiple decision trees as a classifier [14][15][16]. Combining the voting results from various decision trees allows Random Forest to produce more accurate and consistent predictions [17][18]. Combining the voting results from multiple decision trees allows Random Forest to produce more precise and consistent predictions [19][20].

Research by Jackins et al. [21] on AI-based clinical disease prediction using the Random Forest and Na¨ive Bayes classification methods. The Naïve Bayes classification approach achieved accuracies of 74.46%, 82.35%, and 63.74% for diabetes, coronary heart disease, and cancer data, respectively. The Random Forest model categorization demonstrated 74.03%,

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

83.85%, and 92.40% accuracy. The accuracy of the Random Forest model for the three diseases exceeds that of Naïve Bayes. Research by Depari et al. [22] regarding comparing Naïve Bayes, Decision Tree, and Random Forest models for predicting heart disease classification found that Random Forest provides the best performance in classifying heart disease. The performance evaluation results of the three methods showed that the accuracy value for Decision Tree was 71%, Naïve Bayes was 72%, and Random Forest reached the highest score of 75%. Research by Devika et al. [23] regarding the comparison of Random Forest, Na¨ıve Bayes, and K-Nearest Neighbors methods found that Random Forest was the best method in classifying kidney disease with an accuracy rate of 99.844%.

The research above shows that Random Forest is the best classification method. However, a common problem in classification is the imbalance of data classes, where the proportion between classes is not balanced [24]. This may result in the model tending to be accurate for the majority class while the minority class is ignored [25][26][27]. One solution to overcome this problem is to use the SMOTE (Synthetic Minority Over-sampling Technique) algorithm, which oversamples minor classes by creating synthetic data to achieve a balance between classes [28][29][30].

Research on the implementation of SMOTE has been carried out a lot, including research by Demir et al. [31] concerning the assessment of the oversampling technique in classifying soil liquefaction datasets utilizing Support Vector Machine, Random Forest, and Naïve Bayes, a comparison of three oversampling methods—Simple Oversampling (OVER), Random Oversampling Examples (ROSE), and Synthetic Minority Oversampling Technique (SMOTE)—revealed that the SMOTE algorithm outperforms the other oversampling methods. Research by Prasetya et al. [32] regarding the comparison of the use of the SMOTE algorithm with K-Nearest Neighbors and Random Forest on unbalanced data shows that SMOTE Random Forest can develop an excellent classification model with an accuracy level of 96.28%, specificity of 93.44%, sensitivity of 99.17%, precision of 93.70%, and AUC of 96.30%.

Based on previous studies, it is known that the Random Forest method and the implementation of the SMOTE algorithm can provide optimal results. Integrating SMOTE into Random Forest to overcome the problem of data imbalance, where the amount of data in the minority class (e.g., defaulters) is much less than the majority class. Without SMOTE, the Random Forest model tends to be biased toward the majority class, so the classification accuracy of the minority class is low. With SMOTE, minority data is synthetically augmented so that the class distribution becomes balanced, allowing Random Forest to learn more fairly and produce more accurate predictions for all classes. Therefore, this study offers novelty by integrating the Random Forest algorithm with the SMOTE technique to address data imbalance in credit risk classification. Although using open data from Kaggle, this study is tailored to the local context by adjusting the variable weights based

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

on input from Indonesian banking practitioners. This approach not only improves the accuracy and reliability of the classification model but also provides an empirical contribution to the effectiveness of data balancing in predicting credit risk more objectively and efficiently.

II. RESEARCH METHOD

This study is executed through several methodical phases to categorize credit risk evaluation utilizing the Random Forest technique. Random Forest is an ensemble method in machine learning that works by building many decision trees during training. This method is designed to improve the accuracy of data classification results and reduce the risk of overfitting. Random Forest generates predictions by combining the results of many decision trees. The following is the research flowchart in Fig. 1 and the Random Forest architecture in Fig. 2.

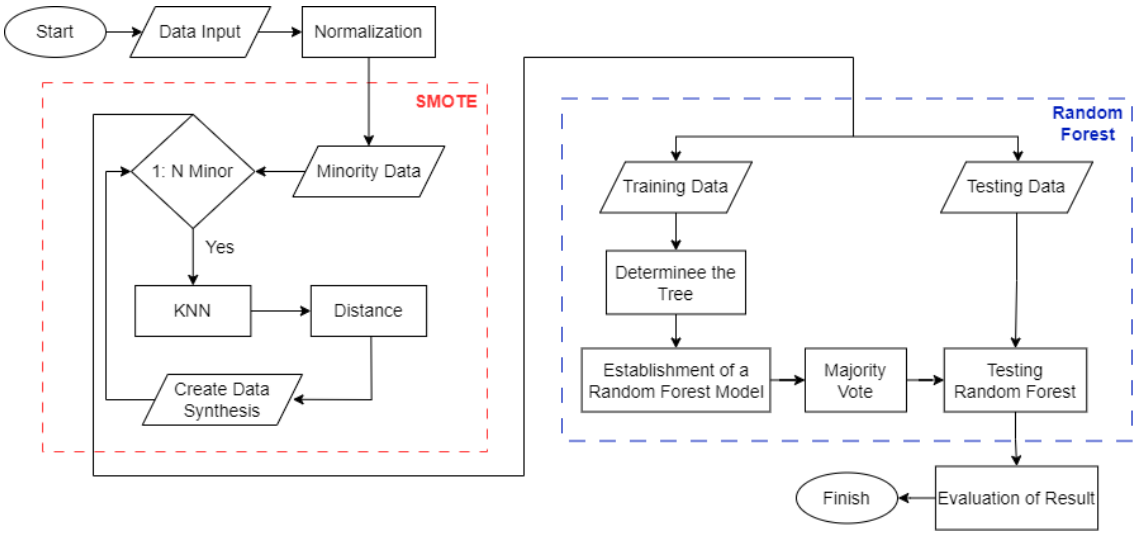

Fig 1. Research Flowchart

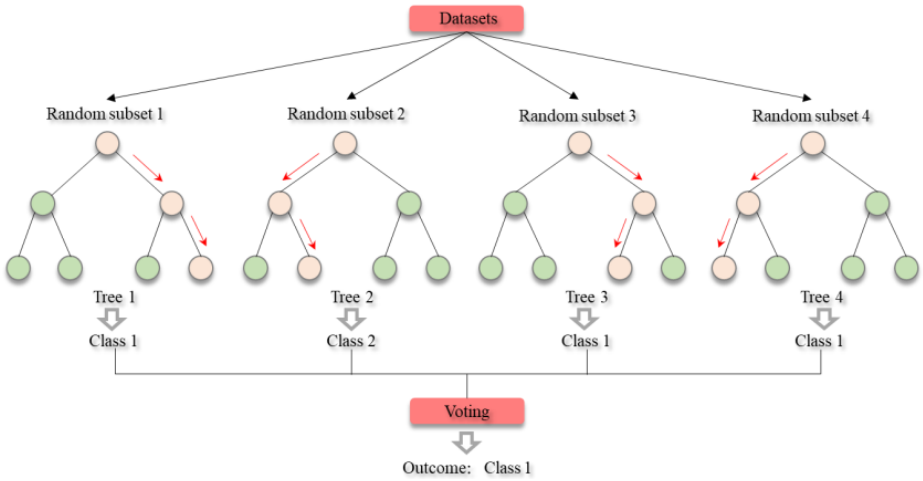

Fig 2. Random Forest Architecture

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

A. Data Input

This study uses secondary data from the Kaggle Dataset [33], which consists of 27.591 data points and 9 variables, where 8 independent variables (Age, Home, Emp Length, Intent, Rate, Percent Income, Default, and Cred Length) and 1 dependent variable (Status). Kaggle is an open-source platform that provides extensive and high-quality datasets, making it suitable for simulation and learning purposes in developing machine learning models. Meanwhile, credit data from financial institutions in Indonesia tends to be difficult to access because it is confidential and has limited distribution, making it impossible to use directly in this study as case study data.

The dataset is partitioned into training and testing subsets in three ratios, 70:30, 80:20, and 90:10, to balance the model training process and performance evaluation. This proportion allows the model to obtain sufficient learning information while providing representative data for accuracy testing. The proportion selection is also adjusted to the size of the dataset. It is a standard practice in machine learning research to maintain the validity and replicability of the classification results [34][35]. The data classifying imbalanced data with the **accepted** class (as many as 22.435) is much larger than the **rejected** class (as many as 5.156). The following description of the variables of this study is presented in Table 1.

Table 1. Description of Data Variables

Variables	Feature Name	Description	Data Type	Value
X1	Age	Age of the loan applicant	Numerical	[20, 84]
X2	Home	Homeownership status	Category	[Rent, Own, Mortgage, Other]
X3	Emp Length	Length of employment in years	Numerical	[0, 41]
X4	Intent	Purpose of loan	Category	[Personal, Medical, Education, Home Improvement, Venture, Debt Consolidation]
X5	Rate	Loan interest rate	Numerical	[5.42, 19.91]
X6	Percent Income	Loan amount as a percentage of revenue	Numerical	[0, 0.83]
X7	Default History	Has the applicant defaulted before?	Category	[Y, N]
X8	Cred Length	Length of applicant's credit history	Numerical	[2, 30]
Y	Status	Loan approval status	Category	[Accepted, Rejected]

Although the data for this study comes from Kaggle, an international open data platform, contextual adjustments are made by assigning weights to each classification variable based on the results of consultations with banking experts in Indonesia. This approach aims to ensure that the

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

classification model built is technically accurate and practically relevant in the context of real application in the national banking sector. Experts determine weights through a structured subjective assessment method (e.g., interviews or closed questionnaires), which reflects the priorities and risk considerations as applicable in the credit assessment system in Indonesia. This is in line with the hybrid model approach in machine learning, where combining global secondary data and local domain knowledge can increase the external validity and generalization of the model to a particular context. The weight of the categorical data is described in Table 2.

Table 2. Category Data Weighting Results

No	Category Data		Weight
1	Age	< 25	0.20
		$25 - 49$	0.40
		$50 - 59$	0.25
		> 59	0.15
2	Home	Own	0.54
		Mortgage	0.15
		Rent	0.23
		Other	0.08
3	Intent	Personal	0.13
		Medical	0.09
		Education	0.17
		Home Improvement	0.22
		Venture	0.35
		Debt Consolidation	0.04
4	Default History	Y	0
		N	1
5	Status	Accepted	0
		Rejected	1

B. Normalization

Data normalization is done to balance the difference in values between features that are too far from each other using the Min-Max Normalization method on the Emp Length, Rate, and Credit Length variables using the following formula.

$$N^* = \frac{N - min\,(n)}{(n) - min\,(n)} \qquad (1)$$

INTENSIF, Vol.9 No.2 August 2025
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

C. SMOTE Algorithm

Addressing imbalanced class data aims to augment the quantity of data in the minority class. The handling employs the SMOTE algorithm, which boosts the amount of minority data to generate synthetic data. This synthesis data will be supplementary information in the training and testing procedures. The procedures of the SMOTE algorithm are as follows:

1. Identifying minority data.
2. The KNN method calculates the distance of each neighbor data by finding the nearest neighbor using the following formula.

$$d_{j,k} = \sqrt{\sum_{i=1}^{n}(x_{ij} - x_{ik})^2}\,; j,k = 1,2,\cdots,P \qquad (2)$$

Description:

x_{ij} : Data on the jth variable

x_{ik} : Data on the kth variable

3. Synthesize the data using the following formula.

$$X_{syn} = x_i + (X_{knn} - x_i) \times \beta \qquad (3)$$

Description:

X_{syn} : Replication result data

x_i : Data to be replicated

X_{knn} : Data with the closest distance from the replicated data

β : Random numbers 0 to 1

D. Random Forest Classification

The output generated by the SMOTE algorithm is further partitioned into training and test datasets according to a predefined evaluation framework. The steps of the Random Forest algorithm are as follows.

1. Train the Random Forest model using the training data.
2. Determine the Mtry value using the following formula.

$$Mtry1 = \frac{1}{2} \times |\sqrt{p}| \qquad (4)$$
$$Mtry2 = |\sqrt{p}| \qquad (5)$$
$$Mtry3 = 2 \times |\sqrt{p}| \qquad (6)$$

Where p is the total variables.

3. Perform Ntree trials, where Ntree is the number of decision trees built using the bagging method, where each tree uses random samples from the original dataset, and the prediction results from all trees are combined with majority voting to determine the final classification.

INTENSIF, Vol.9 No.2 August 2025
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

4. Forming a Random Forest model to determine the Root Node, calculated using the Gini Index.

$$Gini(t) = 1 - \sum[P(j,t)]^2 \qquad (7)$$

P(j,t) is the relative frequency of class j at node t.

5. Calculating the Gini Split to determine the split nodes.

$$Gini(split) = \sum_{i=1}^{k} \frac{n_i}{n} Gini(t) \qquad (8)$$

6. Checking whether the model is good can use the OOB (Out-of-Bag) error, which measures the model's performance based on data not used in training.
7. After all models are built and the OOB evaluation is complete, Random Forest combines the prediction results from each tree. In the context of classification, these results can be used to perform majority voting, referred to as majority voting.
8. After the Random Forest model training process, we continued with testing on the testing data.

E. Evaluation of Results

The next step is evaluating the model results using testing data to calculate the Accuracy, Precision, and Recall values, which will produce a confusion matrix table. The following is a confusion matrix table 3 used to measure model performance.

Table 3. Confusion Matrix

		Prediction	
		Positive	Negative
Actual	Positive	TP	FP
	Negative	FN	TN

Description:

TP (True Positive) : The number of positive data points correctly predicted.
TN (True Negative) : The number of negative data points correctly predicted.
FP (False Positive) : The number of negative data points predicted as positive data.
FN (False Negative) : The number of positive data points predicted as negative data.

The confusion matrix formula evaluates the results by calculating the accuracy, precision, and recall values. The degree of closeness between the predicted and actual values is called the accuracy value. The precision value shows how precise the information the user requests is with the answer given by the system. Meanwhile, the recall value shows how well the system can retrieve information. The following equations can be used to calculate precision, recall, and accuracy:

INTENSIF, Vol.9 No.2 August 2025
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

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

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

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

III. RESULT AND DISCUSSION

A. Normalization

Data normalization is performed to equilibrate the value disparity among excessively divergent characteristics, utilizing the Min-Max Normalization technique. The variables requiring normalization are Emp Length, Rate, and Credit Length—the outcomes of data normalization on the initial dataset utilizing Equation 1. Additionally, the outcomes of data normalization, which yielded a scale ranging from 0 to 1, are presented in Table 4.

Table 4. Data After Normalization

Data	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	0.20	0.23	0.07	0.13	0.73	0.59	0	0.04	1
2	0.20	0.54	0.12	0.17	0.39	0.1	1	0	0
3	0.40	0.15	0.02	0.09	0.51	0.57	1	0.04	1
				⋮					
27.590	0.25	0.15	0.12	0.13	0.42	0.1	1	0.86	0
27.591	0.15	0.23	0.05	0.09	0.32	0.15	1	1	0

B. SMOTE

Overcoming data imbalance uses the Synthetic Minority Over-Sampling Technique, called SMOTE. Data imbalance occurs when the number of samples in the dataset is not balanced between different classes or labels. Before SMOTE, the amount of data in the accepted class denoted “0” was 22,435, and the amount in the rejected class denoted “1” was 5,156. Such a significant difference in the amount of data in the accepted class (0) and dismissed class (1) indicates an unbalanced percentage of class distribution. This condition is included in the imbalanced dataset, so it needs to be processed using the SMOTE algorithm. Overcoming data imbalance involves collecting data in the minority class, namely class “1”. Minority class data samples are shown in Table 5.

INTENSIF, Vol.9 No.2 August 2025
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

Table 5. Minority Data

Data	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	0.20	0.23	0.07	0.13	0.73	59	0	0.04	1
2	0.40	0.15	0.02	0.09	0.51	0.57	1	0.4	1
3	0.20	0.23	0.10	0.09	0.68	0.53	1	0	1
				⋮					
5.155	0.25	0.23	0.10	0.13	0.70	0.31	0	0.61	1
5.156	0.15	0.23	0.07	0.22	0.38	0.46	1	0.93	1

Table 5. shows that there are 5,156 minority data points in class 1. The data will be synthesized using the SMOTE algorithm. The initial phase of implementing the SMOTE algorithm involves identifying the neighbors of the data through the KNN approach, which entails computing the distance between each data point and others using Equation 2.

$$d_{j,k} = \sqrt{\sum_{i=1}^{n} \left(x_{ij} - x_{ik}\right)^2} ; j,k = 1,2,\cdots,P$$

$$d_{1,2} = \sqrt{(0.20-0.40)^2 + (0.23-0.15)^2 + \cdots + (0.04-0.4)^2} = 1.047$$

$$\vdots$$

$$d_{1,5156} = \sqrt{(0.20-0.15)^2 + (0.23-0.23)^2 + \cdots + (0.04-0.93)^2} = 1.386$$

Obtained the closest distance of the 1st data to the 164th data with the smallest distance of 0.086 as X_{KNN} data, namely: [0.20, 0.23, 0.05, 0.17, 0.75, 0.53, 0, 0]. After getting the X_{KNN} data, synthesize the 5,157th data to correct the amount of minority data with the 164th X_{KNN} data to equal the amount of majority data using Equation 3.

$$X_{syn} = x_i + (X_{knn} - x_i) \times \beta$$

$$\begin{aligned} X_{syn(5157)} = &([0.20, 0.23, 0.07, 0.13, 0.59, 0, 0.04, 1]) \\ &+ ([0.20, 0.23, 0.05, 0.17, 0.75, 0.53, 0, 0]) \\ &- ([0.20, 0.23, 0.07, 0.13, 0.59, 0, 0.04, 1]) \times 0.2 \\ &= [0.20, 0.23, 0.07, 0.14, 0.74, 0.58, 0, 0.03] \end{aligned}$$

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

The results of $X_{syn(5157)}$ obtained through the data synthesis process are: [0.20, 0.23, 0.07, 0.14, 0.74, 0.58, 0, 0.03]. These values reflect the synthetic representation of the data in the 5,157th data, which was generated using Equation 3.

The synthetic data generation process was based on a sample of the 5,157th synthesized data to reach 435 data points in each class. This was done by processing each minority data sample until the amount of synthetic data reached 17,279 data points, thus achieving a balance between the amount of data in class "1" of 22,435 data and class "0" of 22,435 data, as shown in Table 6.

Table 6. Data After SMOTE

Data	X1	X2	X3	X4	X5	X6	X7	X8	Y
1	0.20	0.23	0.07	0.13	0.73	59	0	0.04	1
2	0.20	0.54	0.12	0.17	0.39	0.10	1	0	0
3	0.40	0.15	0.02	0.09	0.51	0.57	1	0.4	1
				⋮					
44.869	0.40	0.23	0.02	0.13	0.43	0.20	1	0.34	1
44.870	0.20	0.23	0.12	0.17	0.43	0.33	1	0.04	1

C. Random Forest Classification

Before analyzing the classification, the data is divided into two parts, namely training data and testing data. Training data is helpful for training algorithms for model building, and testing data is used to measure the accuracy and performance obtained from the training data. The division of training data and testing data is carried out in three different scenarios, namely 70:30, 80:20, and 90:10. Training data and testing data are divided into two groups, namely, without SMOTE and with SMOTE, shown in Table 7.

Table 7. Proportion of Training Data and Testing Data

Data Split	Total Data Without SMOTE (27,591 Data)		Total Data With SMOTE (44,870 Data)	
	Training Data	Data Testing	Training Data	Data Testing
70:30	19,313	8,278	31,409	13,461
80:20	22,072	5,519	35,896	8,974
90:10	24,831	2,760	40,383	4,487

To achieve an optimal model and minimal Bag (OOB) error, the determination of trees in the random forest algorithm involves consideration of the values of Mtry (predictor variables) and Ntree (number of trees). Three approaches use certain equations to achieve the optimal value of Mtry. The goal is to achieve optimal results with minimal out-of-bag (OOB) error using three ways with Equations 4, 5, and 6.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

$$Mtry1 = \frac{1}{2} \times \left|\sqrt{p}\right| = \frac{1}{2} \times \left|\sqrt{8}\right| \approx 1{,}58 = 2$$
$$Mtry2 = \left|\sqrt{p}\right| = \left|\sqrt{8}\right| \approx 3{,}16 = 3$$
$$Mtry3 = 2 \times \left|\sqrt{p}\right| = 2 \times \left|\sqrt{8}\right| \approx 9{,}48 = 9$$

Based on the results of the above calculations, Mtry values of 2, 3, and 6 were obtained. Furthermore, each Mtry value was tested using the default number of trees (Ntree), namely 25, 50, 100, and 500. The results can be seen in Table 7, which shows the out-of-bag (OOB) error value for each combination of Mtry and Ntree.

Table 8. Tree Determination Result

Mtry	Ntree	Out Of Bag (OOB) Error Without SMOTE	Error Out Of Bag (OOB) With SMOTE
2	25	7.80%	6.37%
	50	7.54%	5.42%
	100	7.30%	4.95%
	500	7.80%	6.37%
3	25	7.37%	6.16%
	50	7.07%	5.30%
	100	7.00%	5.00%
	500	7.37%	6.16%
6	25	6.94%	5.90%
	50	6.93%	5.17%
	100	**6.73%**	**4.88%**
	500	6.94%	5.90%

Table 8 shows that the optimal combination of Mtry and Ntree is Mtry of 6 and Ntree of 100, resulting in the lowest Out of Bag (OOB) error value of 6.73%. With the application of SMOTE, the best combination of Mtry and Ntree is Mtry of 6 and Ntree of 100, with the lowest Out of Bag (OOB) error value of 4.88%. After obtaining the optimal values of Mtry and Ntree, both values are used to make predictions with the Random Forest model on training data and then tested for accuracy on testing data. Next, calculations are performed to form a decision tree by taking random data as much as n = 10, as shown in Table 9.

INTENSIF, Vol.9 No.2 August 2025
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

Table 9. Random Sample Data Tree Formation

Intent	Default	Status
0.04	0	1
0.09	0	1
0.13	0	1
0.13	0	1
0.13	0	1
0.13	1	0
0.13	1	0
0.22	1	0
0.22	1	1
0.35	1	0

Table 9 shows a random sample of data with a credit status of 5 accepted credits and 5 rejected credits. Next, decision tree formation will be done using the random data sample above with the primary branch test on the Intent variable by calculating the relative frequency, Gini index, and Gini split. Calculating the relative frequency of the Intent variable in the random data sample above is used to build the Random Forest model.

$$P(0, \leq 0.04) = \frac{0}{1}$$

$$P(1, \leq 0.04) = \frac{1}{1}$$

Then, determining the root node using the Gini index in the Intent variable helps select the best features when building each tree and measure the separation quality at each node with Equation 7, as follows.

$$I_{Gini}(t) = 1 - \sum [P(j,t)]^2$$

$$I_{Gini}(\leq 0.04) = 1 - [P(0, \leq 0.04)^2 + P(1, \leq 0.04)^2] = 1 - \left[\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right] = 0$$

$$I_{Gini}(> 0.04) = 1 - [P(0, > 0.04)^2 + P(1, > 0.04)^2] = 1 - \left[\left(\frac{5}{9}\right)^2 + \left(\frac{4}{9}\right)^2\right] = 0.493$$

Determining the separation nodes using Gini split on the Intent variable to select the most informative features to separate the data, thus improving the ensemble's ability to make accurate predictions with Equation 8, as follows.

INTENSIF, Vol.9 No.2 August 2025
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

$$Gini(split) = \sum_{i=1}^{k} \frac{n_i}{n} Gini(t)$$

$$Gini(split_{intent:0.04}) = \sum [P(Intent: 0.04) \times I_{Gini}(Intent: 0.04)$$

$$= \left(\frac{1}{10}\right)(0) + \left(\frac{1}{10}\right)(0.493) = 0.44$$

After calculating the relative frequency, Gini index, and Gini split, the Random Forest model decision tree is obtained by taking random data as much as n = 10, depicted in Fig. 3.

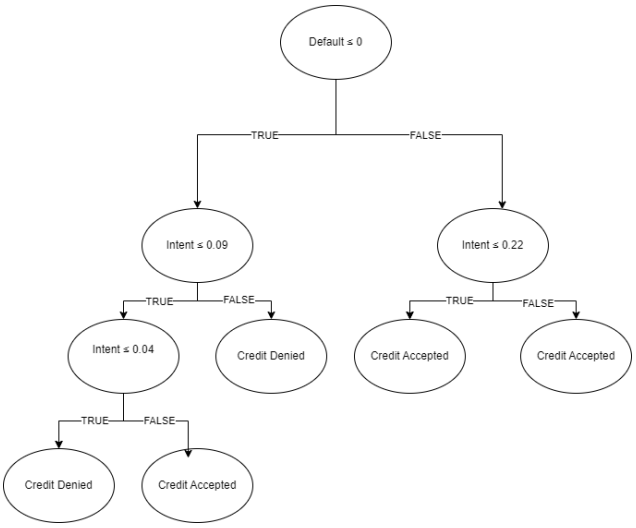

Fig 3. Random Forest Decision Tree

The decision tree in Fig. 3 shows that the Default variable with sample data of 10 is the primary sorting variable and most determines the classification of credit risk assessment. Then, the True branch with variable Intent ≤ 0.09 and the False branch with variable Intent ≤ 0.22 become the second branch. On the Default variable of the True branch with Income ≤ 0.09, the False branch is predicted to be denied credit. On the True branch, the Default variable with Intent ≤ 0.04, the True branch is predicted to be denied credit. The False branch is expected to receive credit on the True branch Default variable with Intent ≤ 0.04. The True branch predicts that credit is accepted on the False branch, the Default variable with Intent ≤ 0.22. Predictive credit is rejected on the Default variable False branch with Intent ≤ 0.22 False branch.

The credit risk assessment classification process results using the Random Forest Algorithm with SMOTE implementation are influenced by different data divisions to explore a more profound understanding of the model's performance. Table 10 shows the Accuracy, Precision, and recall values based on Equations 9, 10, and 11 on the data distribution with a ratio of 70:30, 80:20, and 90:10.

INTENSIF, Vol.9 No.2 August 2025
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

Table 10. Comparison of Model Evaluation Results

Splitting Data	Accuracy	Precision	Recall
70:30	93.81%	96.87%	90.55%
80:20	**94.41%**	97.03%	91.55%
90:10	84.25%	96.77%	91.84%

Based on Table 10. It can be seen that the goodness of the best model of the Random Forest algorithm on the classification of credit risk assessment with the implementation of SMOTE using 80:20 data division for training and testing, with an accuracy value of 94.41%, which shows the ability of the model to make correct predictions.

A comparative evaluation of the results was conducted to determine the superior method between the two approaches used. The review was conducted on both Random Forest algorithms, those that did not apply SMOTE and those that did. An overview of the evaluation results for both methods can be seen in Fig. 4.

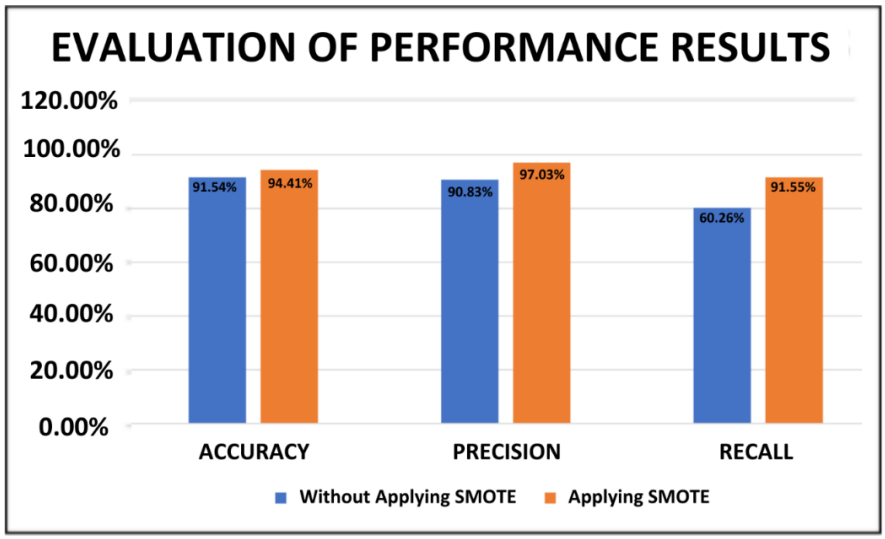

Fig 4. Evaluation Results Comparison Chart

Figure 4 shows the evaluation values, namely accuracy, precision, and recall, of the Random Forest method without the application of SMOTE and with the application of SMOTE. Based on these results, it is concluded that the SMOTE method can improve the performance evaluation of the Random Forest algorithm. This study's results align with the research of Mizwar et al. [36] regarding the classification of heart disease using the SMOTE and Random Forest methods. The study resulted in better accuracy in the application of the SMOTE – Random Forest method compared to the results obtained in research conducted without using SMOTE, which reached

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

92% accuracy; the best result is an increase of 2% from the accuracy results produced by research without SMOTE – Random Forest, which is 90%. Other research results by Polat et al. [37] on a hybrid approach to Parkinson's disease classification using speech signals, using a combination of SMOTE and Random Forest. The research resulted in better accuracy in the application of the SMOTE – Random Forest method compared to the results obtained in research conducted without using SMOTE, which reached an accuracy of 94.89%; this best result was an increase of 7.86% from the accuracy results produced by research without SMOTE – Random Forest, which was 87.03%.

In this study, the accuracy value increased by 2.87% with the application of the SMOTE method from 91.54% to 94.41%. The precision value increased by 6.2% with the application of the SMOTE method from 90.83% to 97.03%. The recall value increased by 31.29% with the application of the SMOTE method from 60.26% to 91.55%. Random Forest with SMOTE and without SMOTE perform differently because SMOTE addresses class imbalance by adding synthetic data to the minority class. Without SMOTE, the model tends to be biased towards the majority class, resulting in poor performance on the minority class (rejected class), especially in the recall metric. With SMOTE, the class distribution is more balanced, allowing Random Forest to learn better from both classes and produce fairer and more accurate performance, especially in detecting minority cases.

IV. CONCLUSION

This study successfully built an optimal credit risk classification model by integrating the Random Forest algorithm and the SMOTE method on Kaggle data that has been adjusted for criteria weights based on input from banking experts. This study's results show that applying the Synthetic Minority Over-sampling Technique (SMOTE) method to the Random Forest algorithm in credit risk assessment classification significantly improves model performance. With a data split of 80:20, where the training data is 35,896, and the testing data is 8,974, the resulting model achieved 94.41% accuracy, 97.03% precision, and 91.55% recall. These results indicate a substantial improvement compared to the model without SMOTE, where accuracy increased by 2.87%, precision by 6.2%, and recall by 31.29%. These improvements indicate that SMOTE successfully addresses the class imbalance in the dataset, improving the model's ability to correctly classify the data, identify positive classes, and capture most of the actual positive instances. This shows that SMOTE can reduce bias towards the majority class and increase the model's sensitivity to previously difficult-to-detect risks.

The credit risk assessment classification results show several characteristics in accepting or rejecting credit for a debtor. Relatively young debtors with a stable income who own their own

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

home, mortgage, or lease are more likely to be approved for credit. In addition, the length of employment and good credit history are also determining factors in credit acceptance. A clear and rational purpose of the loan, such as for business or education, also increases the chances of receiving credit. However, borrowers who have a history of frequent defaults or poor credit history, as well as loan amounts that exceed their repayment capacity or interest rates that are unaffordable, will often be denied credit. In addition, the purpose of the loan to repay debts is usually detected. By considering variables such as age, Income, home ownership, length of employment, loan purpose, loan amount, interest rate, Income in percent, default history, and length of credit history, lending institutions can make more informed decisions in granting credit.

This study successfully built an optimal credit risk classification model by integrating the Random Forest algorithm and the SMOTE method on Kaggle data that has been adjusted for criteria weights based on input from banking experts. For further development, it is recommended that this model be validated using actual data from local financial institutions to improve relevance and external validity. In addition, testing with other ensemble algorithms, such as XGBoost or LightGBM, can be carried out to evaluate the potential for improving model performance. Future research can also consider the application of feature selection and model interpretation techniques, such as SHAP, to identify variables that contribute most to risk prediction. Techniques such as ADASYN or Borderline–SMOTE can be explored as an alternative to handling data imbalance. Finally, the development of a dashboard-based predictive system or recommendation system can be carried out to support real-time credit decision-making in financial institutions.

Author Contributions: *Nafa Nur Adifia*: Conceptualization, Methodology, Writing – Original Draft, Software. *Yuniar Farida*: Methodology, Investigation, Writing – Review & Editing, Supervision. *Wika Dianita Utami*: Investigation, Data Curation.

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The data was derived from the Kaggle Dataset with the link as mentioned in the references.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Nafa Nur Adifia: https://orcid.org/0009-0008-7652-8027
Yuniar Farida: https://orcid.org/0000-0001-8666-4980
Wika Dianita Utami: https://orcid.org/0009-0004-9214-9845

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

REFERENCES

[1] K. Danylkiv, N. Hembarska, and O. Voloshyn, "Efficiency of Using Financial and Credit Instruments To Intensify the Innovative Development of Small Business Structures in Ukraine," J. Lviv Polytech. Natl. Univ. Ser. Econ. Manag. Issues, vol. 4, no. 2, pp. 133–143, 2020, **doi:** 10.23939/semi2020.02.133.

[2] P. E. T. Dewi, "the Legal Obligation of Bank in Implementing Prudential Principles Through Credit Analysis," Int. J. Business, Econ. Law, vol. 15, no. 5, p. 109, 2018, [Online]. Available: **doi:** https://ijbel.com/wp-content/uploads/2018/06/ijbel-243.pdf

[3] E. Gila-Gourgoura and E. Nikolaidou, "Credit Risk Determinants in the Vulnerable Economies of Europe: Evidence from the Spanish Banking System," Int. J. Bus. Econ. Sci. Appl. Res., vol. 10, no. 1, pp. 60–71, 2017, **doi:** 10.25103/ijbesar.101.08.

[4] J. N. Githama and P. Gachanja, "Effects of Credit Appraisal Methods on Non-Performing Loans in Government Owned Financial Institutions, A Case of Kenya Commercial Bank Limited," Int. J. Curr. Asp., vol. 4, no. 2, pp. 1–12, 2020, **doi:** 10.35942/ijcab.v4i2.123.

[5] R. Ranyard, S. McNair, G. Nicolini, and D. Duxbury, "An item response theory approach to constructing and evaluating brief and in-depth financial literacy scales," J. Consum. Aff., vol. 54, no. 3, pp. 1121–1156, 2020, **doi:** 10.1111/joca.12322.

[6] A. Fattahi, J. Sijm, and A. Faaij, "A systemic approach to analyze integrated energy system modeling tools: A review of national models," Renew. Sustain. Energy Rev., vol. 133, no. August, p. 110195, 2020, **doi:** 10.1016/j.rser.2020.110195.

[7] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, "A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier," Knowledge-Based Syst., vol. 205, p. 106270, 2020, **doi:** 10.1016/j.knosys.2020.106270.

[8] T. R. Ramesh, U. K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi, "Predictive Analysis of Heart Diseases With Machine Learning Approaches," Malaysian J. Comput. Sci., vol. 2022, no. Special Issue 1, pp. 132–148, 2022, **doi:** 10.22452/mjcs.sp2022no1.10.

[9] Y. Farida, M. R. Nurfadila, and D. Yuliati, "Identifying Significant Factors Affecting the Human Development Index in East Java Using Ordinal Logistic Regression Model," JTAM (Jurnal Teor. dan Apl. Mat., vol. 6, no. 3, p. 476, 2022, **doi:** 10.31764/jtam.v6i3.8301.

[10] P. Pięta and T. Szmuc, "Applications of rough sets in big data analysis: An overview," Int. J. Appl. Math. Comput. Sci., vol. 31, no. 4, pp. 659–683, 2021, **doi:** 10.34768/amcs-2021-0046.

[11] D. Chicco, V. Starovoitov, and G. Jurman, "The Benefits of the Matthews Correlation Coefficient (MCC) over the Diagnostic Odds Ratio (DOR) in Binary Classification Assessment," IEEE Access, vol. 9, no. Mcc, pp. 47112–47124, 2021, **doi:** 10.1109/ACCESS.2021.3068614.

[12] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," J. Data Anal. Inf. Process., vol. 08, no. 04, pp. 341–357, 2020, **doi:** 10.4236/jdaip.2020.84020.

[13] A. Alfani W.P.R., F. Rozi, and F. Sukmana, "Prediksi Penjualan Produk Unilever Menggunakan Metode K-Nearest Neighbor," JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform., vol. 6, no. 1, pp. 155–160, 2021, **doi:** 10.29100/jipi.v6i1.1910.

[14] Z. Khan et al., "Ensemble of optimal trees, random forest and random projection ensemble classification," Adv. Data Anal. Classif., vol. 14, no. 1, pp. 97–116, 2020, **doi:** 10.1007/s11634-019-00364-9.

[15] S. H. Hasanah and E. Julianti, "Analysis of CART and Random Forest on Statistics

Student Status at Universitas Terbuka," INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf., vol. 6, no. 1, pp. 56–65, 2022, **doi:** 10.29407/intensif.v6i1.16156.

[16] N. Arora and P. D. Kaur, "A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment," Appl. Soft Comput. J., vol. 86, p. 105936, 2020, **doi:** 10.1016/j.asoc.2019.105936.

[17] S. Han, B. D. Williamson, and Y. Fong, "Improving random forest predictions in small datasets from two-phase sampling designs," BMC Med. Inform. Decis. Mak., vol. 21, no. 1, pp. 1–9, 2021, **doi:** 10.1186/s12911-021-01688-3.

[18] Z. Sajjadnia, R. Khayami, and M. R. Moosavi, "Preprocessing Breast Cancer Data to Improve the Data Quality, Diagnosis Procedure, and Medical Care Services," Cancer Inform., vol. 19, pp. 7–12, 2020, **doi:** 10.1177/1176935120917955.

[19] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," IEEE Access, vol. 8, pp. 90847–90861, 2020, **doi:** 10.1109/ACCESS.2020.2994222.

[20] S. Ben Atitallah, M. Driss, and I. Almomani, "A Novel Detection and Multi-Classification Approach for IoT-Malware Using Random Forest Voting of Fine-Tuning Convolutional Neural Networks," Sensors, vol. 22, no. 11, 2022, **doi:** 10.3390/s22114302.

[21] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naïve Bayes," J. Supercomput., vol. 77, no. 5, pp. 5198–5219, 2021, **doi:** 10.1007/s11227-020-03481-x.

[22] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, "Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," Inform. J. Ilmu Komput., vol. 18, no. 3, p. 239, 2022, **doi:** 10.52958/iftk.v18i3.4694.

[23] R. Devika, S. V. Avilala, and V. Subramaniyaswamy, "Comparative study of classifier for chronic kidney disease prediction using naïve Bayes, KNN and random forest," Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019, no. Iccmc, pp. 679–684, 2019, **doi:** 10.1109/ICCMC.2019.8819654.

[24] K. M. Hasib et al., "A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem," J. Comput. Sci., vol. 16, no. 11, pp. 1546–1557, 2020, **doi:** 10.3844/JCSSP.2020.1546.1557.

[25] D. C. R. Novitasari et al., "Whirlwind Classification with Imbalanced Upper Air Data Handling using SMOTE Algorithm and SVM Classifier," J. Phys. Conf. Ser., vol. 1501, no. 1, 2020, **doi:** 10.1088/1742-6596/1501/1/012010.

[26] S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi, "The Majority Can Help the Minority: Context-rich Minority Oversampling for Long-tailed Classification," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2022-June, pp. 6877–6886, 2022, **doi:** 10.1109/CVPR52688.2022.00676.

[27] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," IEEE Access, vol. 9, pp. 64606–64628, 2021, **doi:** 10.1109/ACCESS.2021.3074243.

[28] D. A. Nurdeni, "Extracting Information From Twitter Data To Identify Types of Assistance for Victims of Natural Disasters: an Indonesian Case Study," J. Manag. Inf. Decis. Sci., vol. 25, no. S1, pp. 1–14, 2022, [Online]. Available: **doi:** https://www.researchgate.net/profile/Ariana-Yunita-2/publication/369795342_Special_Issue_1_2022_1_Journal_of_Management_Information_and_Decision_Sciences/links/642d4abaad9b6d17dc393e2f/Special-Issue-1-2022-1-Journal-of-Management-Information-and-Deci

[29] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning from Imbalanced Data," IEEE Access, vol. 9, pp. 74763–74777, 2021, **doi:**

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i2.23930

10.1109/ACCESS.2021.3080316.

[30] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," Mach. Learn., vol. 113, no. 7, pp. 4903–4923, 2024, **doi:** 10.1007/s10994-022-06296-4.

[31] S. DEMİR and E. K. ŞAHİN, "Evaluation of Oversampling Methods (OVER, SMOTE, and ROSE) in Classifying Soil Liquefaction Dataset based on SVM, RF, and Naïve Bayes," Eur. J. Sci. Technol., no. 34, pp. 142–147, 2022, **doi:** 10.31590/ejosat.1077867.

[32] J. Prasetya and A. Abdurakhman, "Comparison of Smote Random Forest and Smote K-Nearest Neighbors Classification Analysis on Imbalanced Data," Media Stat., vol. 15, no. 2, pp. 198–208, 2023, **doi:** 10.14710/medstat.15.2.198-208.

[33] Kaggle, "Credit Risk Analysis." Accessed: Sep. 12, 2023. [Online]. Available: https://www.kaggle.com/datasets/nanditapore/credit-risk-analysis

[34] S. Lasniari, J. Jasril, S. Sanjaya, F. Yanto, and M. Affandes, "Klasifikasi Citra Daging Babi dan Daging Sapi Menggunakan Deep Learning Arsitektur ResNet-50 dengan Augmentasi Citra," J. Sist. Komput. dan Inform., vol. 3, no. 4, p. 450, 2022, **doi:** 10.30865/json.v3i4.4167.

[35] R. M. Candra and A. Nanda Rozana, "Klasifikasi Komentar Bullying pada Instagram Menggunakan Metode K-Nearest Neighbor," IT J. Res. Dev., vol. 5, no. 1, pp. 45–52, 2020, **doi:** 10.25299/itjrd.2020.vol5(1).4962.

[36] A. M. A. Rahim, I. Y. R. P. Pratiwi, and M. A. Fikri, "Indonesian Journal of Computer Science," Indones. J. Comput. Sci., vol. 12, no. 2, pp. 284–301, 2023, **doi:** https://doi.org/10.33022/ijcs.v12i1.3135.

[37] K. Polat, "A hybrid approach to Parkinson disease classification using speech signal: The combination of SMOTE and random forests," 2019 Sci. Meet. Electr. Biomed. Eng. Comput. Sci. EBBT 2019, pp. 1–3, 2019, **doi:** 10.1109/EBBT.2019.8741725.