International Journal of Engineering and Computer Science Applications (IJECSA)
Vol. 4, No. 2, September 2025, pp. 91∼100
ISSN: 2828-5611
DOI: doi.org/10.30812/ijecsa.v4i2.5142

Handling Imbalanced Data in K-Nearest Neighbor Algorithm using Synthetic Minority Oversampling Technique-Nominal Continuous

Anjani Tri Pramudita, Memi Nor Hayati, Surya Prangga
Universitas Mulawarman, Samarinda, Indonesia

Article Info

Article history:

Received May 30, 2025
Revised June 3, 2025
Accepted June 14, 2025

Keywords:

Credit
K-fold Cross Validation
K-NN
SMOTE-NC

ABSTRACT

Classification is a part of data mining that aims to predict the class of data using a trained machine learning model. K-Nearest Neighbor (K-NN) is one of the classification methods that uses the concept of distance to the nearest neighbor in creating classification models. However, K-NN has limitations in handling imbalanced class distributions. This core problem can be addressed by applying a class balancing technique. One such technique is the Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTE-NC), which is suitable for datasets containing nominal and continuous variables. **This research aims** to classify Honda motorcycle loan customer data at Company Z using the K-NN combined with SMOTE-NC to address data imbalance. **This experimental research method uses** a 10-fold cross-validation approach to partition training and testing data. The input variables include gender, occupation, length of installment, income, installment amount, motorcycle price, and down payment, while the output variable is payment status (current or non-current). **The results of this research** are: the optimal K value for classification using K-NN with SMOTE-NC is K = 1, with an average Average Probability of Error Rate (APER) of 0.143. The best result is in subset 8 with an APER value of 0.033. In this subset, out of 61 data points, 34 current-status customers are correctly classified as current, and 25 non-current-status customers are correctly classified as non-current, with only one misclassification in each class. **This study concludes** that the combination of SMOTE-NC and K-NN (K=1) provides high classification accuracy for imbalanced data, and can be effectively used to support credit risk assessment in motorcycle financing.

Corresponding Author:

Memi Nor Hayati,
Universitas Mulawarman, Samarinda, Indonesia,
Email: meminorhayati@fmipa.unmul.ac.id

How to Cite: A. T. Pramudita, M. N. Hayati, and S. Prangga, "Handling Imbalanced Data in K-Nearest Neighbor Algorithm using Synthetic Minority Oversampling Technique-Nominal Continuous," *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 4, no. 2, pp. 91-100, Sep 2025. doi: 10.30812/ijecsa.v4i2.5142.

1. INTRODUCTION

Data mining is extracting information and patterns from very large data. Data collection, data extraction, data analysis, and data statistics are part of data mining, which aims to find previously unknown patterns [1]. Based on its function, data mining has been formed into six groups, namely clustering, regression, anomaly detection, association rule learning, summarization, and classification [2]. Classification is a process of obtaining a number of models or functions that can describe, recognize, and distinguish classes of data or concepts [1]. Classification aims to predict the label of an unknown category of objects based on attributes or features [3]. Classification is also defined as a form of data analysis that produces a model to describe important data classes. Classification techniques that are often used in solving cases include C4.5 Algorithm, ID3, Naïve Bayesian Classification, Classification and Regression Tree (CART), K-Nearest Neighbor (K-NN) Algorithm, and others. K-NN is the most basic and simple method to classify objects based on training data that is closest to the object. The training data is projected onto a multi-dimensional space, where each dimension represents a feature of the data. However, K-NN has a weakness, which is that it is weak in classifying data that has unbalanced classes, which will result in misprediction or misclassification of the data. Classification on unbalanced data is a special situation when the class distribution is unbalanced, where the majority class is more than the minority class [4].

The problem of unbalanced class classification often occurs because the number of observations of one class label is much lower than that of other class labels. Techniques to overcome the problem of unbalanced data include undersampling and oversampling techniques. The oversampling technique has been extended using the Synthetic Minority Oversampling Technique (SMOTE) approach. In its application, the SMOTE approach creates "synthetic" data, replicating the data from the minority class. SMOTE can build synthetic minority class sample data by interpolating between neighboring minority class instances (observations). SMOTE is a good and effective oversampling technique to handle overfitting in the oversampling process and imbalances in the minority class. SMOTE can only be performed if the input variable is numeric, while for data that has a mixed type of nominal and continuous, an extension of SMOTE, namely Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTE-NC) [5, 6]. Previous research on the classification of unbalanced data has been conducted by Utari (2023) entitled "Integration of SVM and SMOTE-NC for Classification of Heart Failure Patients". In the study, the researcher used a dataset consisting of continuous and nominal attributes from heart failure patients. The data exhibited class imbalance between patient statuses (dead or alive), which was handled using the SMOTE-NC technique prior to classification with the Support Vector Machine (SVM) algorithm. The results showed that the integration of SMOTE-NC with SVM improved classification performance, as indicated by higher accuracy and F1 scores compared to SVM without SMOTE-NC [7].

There **are gaps** that have not been resolved by previous research, namely the lack of studies applying the SMOTE-NC method in domains beyond the health sector, particularly in financial or credit-related datasets with mixed-type attributes (nominal and continuous). Moreover, existing studies predominantly integrate SMOTE-NC with advanced classifiers such as SVM, while simpler, interpretable models like K-NN remain underexplored in this context, especially in optimizing the number of neighbors (K) for performance improvement on synthetic balanced datasets. **The difference** between this research and the previous one is that this study applies SMOTE-NC to unbalanced data from a completely different domain, motorcycle credit customers, where the target variable reflects credit payment smoothness. Furthermore, the research investigates the application of the K-NN algorithm, focusing on determining the optimal K and the best subset of features to achieve the most accurate classification results [8].

The objective of this study is to determine the optimal value of K and the best feature subset for classifying Honda motorcycle credit customers at Company Z using the K-NN method after applying SMOTE-NC to balance the dataset. The contribution of this research lies in broadening the applicability of SMOTE-NC to financial datasets and demonstrating its effectiveness in combination with a basic but practical classification algorithm like K-NN. **This contributes** to the development of data mining practices in the financial services industry, particularly in credit risk analysis and customer profiling.

2. RESEARCH METHOD

The data used in this study is secondary data obtained directly from the internal archives of company Z, which is engaged in motorcycle financing in Malinau Regency, North Kalimantan. This data was collected as digital documents in the form of customer credit application reports from March to May 2024. Each data entry represents one customer who applied for a Honda motorcycle loan, with 319 customer data points. The variables in this study consist of 8 variables, which include seven input variables and one output variable. Input variables include demographic data and customer credit history, such as age, occupation, income, gender, marital status, housing status, and number of dependents. The output variable represents the credit repayment status, i.e., current or non-current. Summary statistics of each variable can be seen in Table 1, which includes frequency distributions for categorical data and descriptive statistical measures (such as mean and standard deviation) for numerical data.

Table 1. Research Variables

Symbol	Description	Scale
Y	Status of Payment	Nominal
X_1	Gender	Nominal
X_2	Jobs	Nominal
X_3	Installment Length	Nominal
X_5	Installment Amount (Million Rupiah)	Rasio
X_6	Price of Motorcycle (Million Rupiah)	Rasio
X_7	Down Payment (Million Rupiah)	Rasio

Figure 1 shows the flow of data analysis stages carried out in this study. The process starts from the data collection stage to classification using the K-Nearest Neighbor (K-NN) algorithm. Each step in this flow is designed to handle the problem of unbalanced data by applying the Synthetic Minority Oversampling Technique (SMOTE) and using the Gower distance, which is appropriate for data with mixed types (nominal and continuous).

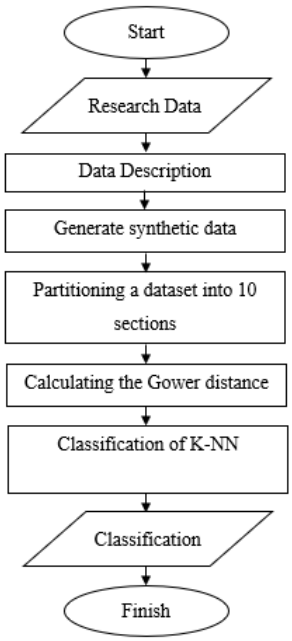

Figure 1. Flowchart of Analysis Stages

Figure 1 shows a flowchart of the stages of analysis carried out in this study with the help of R software. Each step in the diagram is explained as follows.

1. Descriptive Statistics

   The initial stage was calculated using descriptive statistics on Honda motorcycle loan customer data in company Z, Malinau Regency, North Kalimantan. This analysis included creating a pie chart to see the class distribution on the credit payment status variable and to identify data imbalance problems.

2. K-NN Classification with SMOTE-NC

   a) Synthetic data Generation

   The SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous) technique was used to handle the imbalanced data problem. This technique generates synthetic data by considering the nearest neighbor similarity (K) for nominal variables and using the interpolation formula for continuous variables, as in Equation (1) [9]. In the process of generating synthetic data using SMOTE-NC, several notations are used to explain the main components of the formula.

The notation x_syn refers to the synthetic data to be created, while x_knn represents the data with the closest distance to the replicated data. The notation x_r denotes the data to be replicated, and δ is a random value ranging from 0 to 1.

$$x_{syn} = x_r + (x_{knn} - x_r) \times \delta \tag{1}$$

b) Perform data randomization

Divide the dataset partition into 10 parts with a k-fold CV scheme to determine training data and testing data. The determination of the amount of data in each subset uses Equation (2) [10]. In the k-fold cross-validation method, the notation B represents the amount of data in each subset or fold, N is the total amount of data in the dataset, and k denotes the number of folds used in the k-fold scheme.

$$B = \frac{N}{k} \tag{2}$$

c) Classification of K-NN with SMOTE-NC

3. The classification of K-NN with SMOTE-NC is carried out to classify the testing data on data that has been balanced with SMOTE-NC.

a) Determine the neighbor values with K = 1, 3, 5, 7, and 9.

b) Calculate the distance of the Gower data testing to the training data on each variable that corresponds to the measurement scale of the random variable. The calculation of the Gower distance on the nominal using Equation (3), ordinal using Equation (4), interval and ratio data scales using Equation (5) [11].

$$d_q\left(x_{i,q};\ x_{j,q}\right) = \begin{cases} 0, & x_{i,q} = x_{j,q} \\ 1, & x_{i,q} \neq x_{j,q} \end{cases} \tag{3}$$

Where, the distance between observations on the qqq-th variable is denoted as $d_q\left(x_{i,q};\ x_{j,q}\right)$, which represents the distance between the observations $x_{i,q}$ and $x_{j,q}$. Here, $x_{i,q}$refers to the value of the q-th variable for the i th entity, while $x_{j,q}$ refers to the value of the q-th variable for the j-th entity. The indices i and j represent the observation identifiers, where $i, j : 1,\ 2,\ \ldots,\ n$ and q denotes the variable index, where $i, j :\ 1,\ 2,\ \ldots,\ n$

$$d_q\left(x_{i,q}; x_{j,q}\right) = \frac{\left|R\left(x_{i,q}\right) - R_q\left(x_{j,q}\right)\right|}{\text{maks}\ (\text{R}) - \min\left(\text{R}\right)} \tag{4}$$

Where, the rank of data for the q-th variable is represented as $R\left(x_{i,q}\right)$ for the i-th observation and $R\left(x_{j,q}\right)$ for the j-th observation. These ranks indicate the relative position of each observation within the distribution of the q-th variable. Furthermore, maks_q and $\min_q$ refer to the maximum and minimum ranks of the q-th variable, respectively. These rank values are used to normalize or compare differences between observations, especially when dealing with ordinal or non-numeric data. The distance between two observations on the q-th variable, denoted as $d_q\left(x_{i,q}; x_{j,q}\right)$, is calculated using the normalized absolute difference:

$$d_q\left(x_{i,q}; x_{j,q}\right) = \frac{\left|x_{i,q} - x_{j,q}\right|}{\text{maks}_q - \min_q} \tag{5}$$

In this formula, maks_q represents the maximum value of the q-th variable, while $\min_q$ denotes the minimum value. This normalization ensures that the distance is scaled between 0 and 1, allowing variables with different ranges to be compared fairly in multivariate analyses or distance-based algorithms such as K-NN.

4. Evaluation of classification model.

The Apparent Error Rate (APER) is a metric used to evaluate the performance of a classification model by measuring the proportion of incorrect predictions among all predictions. It is calculated using the Equation (6). where TP (True Positive) indicates the number of correct positive predictions, and TN (True Negative) refers to the number of correct negative predictions. Meanwhile,

FP (False Positive) represents the number of incorrect positive predictions for instances that are actually negative, and FN (False Negative) represents the number of incorrect negative predictions for instances that are actually positive. A lower APER value indicates better model performance, as it reflects fewer errors in prediction. Where the positive class is a customer who is not fluent in paying installments, while the negative class is a customer who is fluent in paying installments [12].

$$\text{APER} = \frac{\text{FP} + \text{FN}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \tag{6}$$

3. RESULT AND ANALYSIS

This chapter presents the process and results of data analysis carried out based on the previously designed stages. The analysis begins with a description of the initial data, followed by a synthetic data generation process to overcome class imbalance using the SMOTE method. Next, the data was divided into ten parts for cross-validation purposes, and the distance between data was calculated using Gower distance, which can handle both nominal and continuous attributes. The results of this calculation were then used in the classification process with the K-NN algorithm.

3.1. Data Description

The description of the data on the variable of the payment status (YIn company Z, Malinau Regency, with current and non-current categories using a pie chart, is presented in Figure 2. Based on Figure 2, the payment status of 319 Honda motorcycle loan customers at company Z, Malinau Regency, has a class imbalance problem, where there are 305 customers with a current payment status, with a percentage of 96%, and 14 customers with an incurrent payment status, with a percentage of 4%. It can also be concluded that the number of Honda motorcycle loan customers in company Z with current status is 92% more than that of customers who are not current.

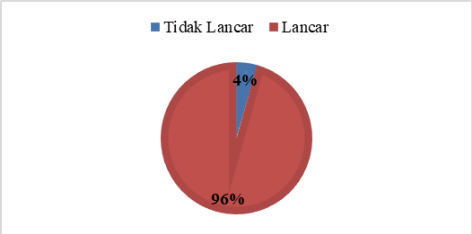

Figure 2. Percentage of Payment Status

3.2. Classification of K-NN with SMOTE-NC

1. Classification of K-NN with SMOTE-NC

The results of synthetic data generation by the SMOTE-NC technique produced 291 synthetic samples. The synthetic data produced is presented in Table 2.

Table 2. Minority-Class Synthetic Data

No	Customer	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	320	1	0	0	1	1	1.347	19.810	2.270
2	321	1	0	0	1	1	1.273	19.810	2.376
3	322	1	0	0	1	1	1.195	20.469	2.425
4	323	1	0	0	1	1	1.188	20.705	2.108
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
291	610	1	0	0	2	1	1.357	23.181	2.614

2. Data randomization

The results of the data randomization that have been combined between the original data and the synthetic data are presented in Table 3.

Table 3. Minority-Class Synthetic Data

Customer	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
527	1	0	0	1	1	1.506	21.951	2.907
64	0	0	0	2	1	1.250	20.570	3.000
217	0	0	0	2	1	1.580	25.960	3.000
257	0	1	1	0	0	1.820	18.090	2.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
485	1	0	1	2	0	1.484	20.89	2.482

3. Determining the value and the amount of data in each subset.

K-fold CV in this study uses, so that the number of k=10. The subsets formed are 10 subsets of data. The calculation of the number of data points in the subset uses data that has been combined from the original data and the synthetic data, with a total of 610 data points. The following is the calculation of the amount of data in a subset, namely, using Equation (2). Based on the calculation numbers of data for each subset that will be used, which is as many as 61 observations. An example of the results of determining the amount of data from 10 subsets is presented in Table 4.

$$B = \frac{610}{10} = 61$$

Table 4. Results of Determining the Amount of Data in a Subset

Subset	Customer	Y	Number of samples
1	527	TL	61
	64	L	
	217	L	
	⋮	⋮	
	348	TL	
2	475	TL	61
	345	TL	
	504	TL	
	⋮	⋮	
	187	L	
3	476	TL	61
	506	TL	
	265	L	
	⋮	⋮	
	54	L	
9	476	TL	61
	506	TL	
	265	L	
	…	…	
	54	L	
10	369	TL	61
	75	L	
	268	L	
	…	…	
	485	TL	

4. Calculating the distance

The calculation of the Gower distance between the testing data and the training data is calculated using Equation (3) for nominal scale data and Equation (5) for continuous data. The results of the calculation of the Gower distance are presented in the form of a matrix, which can be seen in Table 5.

Table 5. Matrix of Gower Distance

Customer	1	2	3	4	5	...	610
1	0	0.1649403	0.1696356	0.6233198	0.3758846	...	0.4424361
2	0.1649403	0	0.0457767	0.6306448	0.2520257	...	0.3080876
3	0.1696356	0.0457767	0	0.6424813	0.206249	...	0.3263574
4	0.6233198	0.6306448	0.6424813	0	0.5614733	...	0.3260094
5	0.3758846	0.2520257	0.206249	0.5614733	0	...	0.5326065
...	...	...	...	...	...	...	...
606	0.3154541	0.4656362	0.4787069	0.3078657	0.6849559	...	0.1610008
607	0.4908761	0.5129593	0.5088694	0.1945373	0.7135757	...	0.2048718
608	0.1991295	0.3609848	0.3152082	0.7901486	0.3946734	...	0.6415656
609	0.0346777	0.1657211	0.2043133	0.6077808	0.4105623	...	0.4493844
610	0.4424361	0.3080876	0.3263574	0.3260094	0.5326065	...	0

5. Classification based on k-value on each subset

After calculating the distance of neighbors using the Gower distance between the training data and the testing data on a subset of data, the Gower distance was sorted from the smallest (closest) to the largest (farthest) in each subset of data. After obtaining the results of sorting the Gower distances, the value of K of the nearest neighbor was determined using odd values of 1, 3, 5, 7, and 9. The voting stage selects categories on the output variable to avoid ties in classification results. This nearest-neighbor classification approach is particularly important in assessing the eligibility of credit recipients, where classification decisions, such as whether a customer is considered creditworthy or not, are based on proximity to similar historical cases [13]. The calculation of the sorted Gower distance can be seen in Table 6.

Table 6. Results of Gower Distance Sequencing 549 Training

Rank	Data Testing		1st Data Testing from Subset 1 (527th Customer)	K-NN Limit	Results of the 1st Data Testing Classification
	Customer	Classification	$d(x_i, x_{527})$		
1	535	TL	0,0033	1	L
2	183	L	0,0053		
3	184	L	0,0053	3	L
4	189	L	0,0056		
5	529	TL	0,0060	5	TL
6	533	TL	0,0068		
...	...	...	...	...	...
549	259	L	0,6950		

Table 6 shows that the 535th customer is the closest training data to the 1st testing data in subset 1, namely the 527th customer. At the limit of 1 nearest neighbor (1-NN), the result of the 1st data testing classification, namely the 1st data in subset 1 (535th customer), is the Current class label (L) on the payment status. This happens because the fluent class label is a classification of the first rank. Meanwhile, at the boundary of the nearest three neighbors (3-NN), the result of the 1st data testing classification is the Non-Current (TL) class label. This happens because the non-fluent class label is the classification that most often appears from the first rank to the third rank, as well as the same thing is done in 5-NN, 7-NN, and 9-NN, so that the classification results of the testing data in the 1st subset have TL results.

6. Model Evaluation

Evaluation was carried out on each pair of training and testing subsets. First, a confusion matrix is formed to calculate APER with Equation (6). The APER was obtained for each subset of each K, which became the testing data, and then the average of the overall APER score was calculated. The results of the APER score and the average of the overall APER score can be seen in Table

7. Based on Table 7, it was found that the average APER score of each neighbor K, where the APER value K=1 was 0.1426, the APER value K=3 and K=5 was 0.1590, the APER value K=7 was 0.1508, and the APER value K=9 was 0.1557. **The findings** of this research are that the optimal value of K in classifying Honda motorcycle loan customers at Company Z using the K-NN algorithm with the application of SMOTE-NC is K = 1, with an average APER value of 0.1426. The results of this research are supported by previous research, which found that the use of SMOTE-NC improved classification performance in datasets with a mix of nominal and continuous variables [7]. Although Utari's research focused on SVM, both studies demonstrate that SMOTE-NC is effective in improving classification accuracy on imbalanced datasets, and is supported by the findings of research [14–16], who demonstrated that applying SMOTE-based oversampling significantly improved classification performance on imbalanced datasets that included mixed data types. This supports the relevance of using SMOTE-NC for similar classification problems.

Table 7. APER of each Testing Subset at each K (K-NN + SMOTE-NC)

Subset	K=1	K=3	K=5	K=7	K=9
1	0,1311	0,1475	0,1803	0,1803	0,1803
2	0,1148	0,1148	0,1311	0,0984	0,0656
3	0,1311	0,1148	0,1147	0,1148	0,1148
4	0,4754	0,4754	0,4918	0,5082	0,5410
5	0,0656	0,0656	0,0656	0,0656	0,0656
6	0,1148	0,1475	0,1639	0,1148	0,1639
7	0,0491	0,0984	0,0984	0,0656	0,0492
8	0,0328	0,0984	0,0492	0,0656	0,0820
9	0,2623	0,2459	0,2295	0,2459	0,2623
10	0,0492	0,0819	0,0656	0,0492	0,0328
Average of APER	0,1426	0,1590	0,1590	0,1508	0,1557

4. CONCLUSION

The findings of this research are that the optimal value of K in classifying Honda motorcycle loan customers at Company Z using the K-NN algorithm with the application of SMOTE-NC is K = 1, with an average APER value of 0.1426. At K = 1, subset eight was identified as the best feature subset, producing the lowest APER value of 0.0328. In this subset, out of 61 total customers, 34 customers with current loan status were correctly classified as current, and 25 customers with non-current status were correctly classified as non-current. Misclassifications occurred only in two cases: one current customer was classified as non-current, and one non-current customer was classified as current. The use of SMOTE-NC improved classification performance in datasets with a mix of nominal and continuous variables. Suggestions for future research include testing alternative classification algorithms, such as Random Forest or SVM, to compare their performance with K-NN in a similar context. Additionally, further optimization of SMOTE-NC parameters (such as the number of neighbors used in the oversampling process) could be explored to examine whether it leads to significant improvements in model accuracy and robustness.

REFERENCES

[1] J. Olufemi Ogunleye, "The Concept of Data Mining," in *Artificial Intelligence*, C. Thomas, Ed., vol. 8, IntechOpen, Mar. 2022. DOI: 10.5772/intechopen.99417.

[2] M. Chaudhry *et al.*, "A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective," *Symmetry*, vol. 15, no. 9, p. 1679, Aug. 2023. DOI: 10.3390/sym15091679.

[3] B. Ghasemkhani, K. F. Balbal, and D. Birant, "A New Predictive Method for Classification Tasks in Machine Learning: Multi-Class Multi-Label Logistic Model Tree (MMLMT)," *Mathematics*, vol. 12, no. 18, p. 2825, Sep. 2024. DOI: 10.3390/math12182825.

[4] F. Y. Pamuji, "Pengujian Metode SMOTE untuk Penanganan Data Tidak Seimbang pada Dataset Binary," in *Seminar Nasional Sistem Informasi 2022*, Malang: Universitas Merdeka Malang, Jan. 2023, pp. 3200–3208.

[5] A. Syukron *et al.*, "Penerapan Metode Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung," *Jurnal Teknologi Informasi dan Terapan*, vol. 10, no. 1, pp. 47–50, Jun. 2023. DOI: 10.25047/jtit.v10i1.313.

[6] N. V. Chawla *et al.*, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002. DOI: 10.1613/jair.953.

[7] D. T. Utari, "Integration of SVM and SMOTE-NC for Classification of Heart Failure Patients," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 4, pp. 2263–2272, Dec. 2023. DOI: 10.30598/barekengvol17iss4pp2263-2272.

[8] E. Bu'ulolo *et al.*, "Implementasi Algoritma K-Nearest Neighbor (K-NN) dalam Klasifikasi Kredit Motor," *Bulletin of Information System Research*, vol. 1, no. 1, pp. 18–22, Dec. 2022. DOI: 10.62866/bios.v1i1.34.

[9] A. N. Kasanah, M. Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 2, pp. 196–201, Aug. 2019. DOI: 10.29207/resti.v3i2.945.

[10] J. Allgaier and R. Pryss, "Cross-Validation Visualized: A Narrative Guide to Advanced Methods," *Machine Learning and Knowledge Extraction*, vol. 6, no. 2, pp. 1378–1388, Jun. 2024. DOI: 10.3390/make6020065.

[11] R. Soraya, M. N. Hayati, and R. Goejantoro, "Klasifikasi Status Hipertensi Pasien UPTD Puskesmas Sempaja, Kota Samarinda Menggunakan Metode K-Nearest Neighbor," *EKSPONENSIAL*, vol. 14, no. 2, p. 67, Nov. 2023. DOI: 10.30872/eksponensial.v14i2.1009.

[12] U. M. Martha *et al.*, "Perbandingan Analisis Diskriminan Kuadratik dengan Analisis Diskriminan Kuadratik Robust," *UNP Journal of Statistics and Data Science*, vol. 2, no. 4, pp. 469–474, Nov. 2024. DOI: 10.24036/ujsds/vol2-iss4/315.

[13] T. T. Muryono and I. Irwansyah, "Implementasi Data Mining untuk Menentukan Kelayakan Pemberian Kredit dengan Menggunakan Algoritma K-Nearest Neighbors (K-NN)," *Infotech: Journal of Technology Information*, vol. 6, no. 1, pp. 43–48, Jun. 2020. DOI: 10.37365/jti.v6i1.78.

[14] M. A. Latief *et al.*, "Handling Imbalance Data using Hybrid Sampling SMOTE-ENN in Lung Cancer Classification," *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 3, no. 1, pp. 11–18, Feb. 2024. DOI: 10.30812/ijecsa.v3i1.3758.

[15] H. Hairani and D. Priyanto, "A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023. DOI: 10.14569/IJACSA.2023.0140864.

[16] B. C. Herawati, H. Hairani, and J. X. Guterres, "SMOTE Variants and Random Forest Method: A Comprehensive Approach to Breast Cancer Classification," *International Journal of Engineering Continuity*, vol. 3, no. 1, pp. 12–23, Feb. 2024. DOI: 10.58291/ijec.v3i1.147.

[This page intentionally left blank.]