STATISTICS IN TRANSITION new series, March 2024
Vol. 25, No. 1, pp. 109–124, https://doi.org/10.59170/stattrans-2024-007
Received – 13.02.2023; accepted – 03.10.2023

Implementation of K-Nearest Neighbor using the oversampling technique on mixed data for the classification of household welfare status

Nur Mutmainnah Djafar[1], **Achmad Fauzan**[2]

Abstract

Welfare is closely related to poverty and the socio-economic disparities in a society. Based on data from the Central Bureau of Statistics, Kulon Progo in Indonesia had the highest poverty rate in the province of the Special Region of Yogyakarta; an increasing trend was observed every year from 2019 to 2021; Kulon Progo also had a low poverty line (after Gunung Kidul) compared to other regencies/cities in this province. This study aimed to classify the household welfare status in Kulon Progo in March 2021 using the K-Nearest Neighbor (KNN) method. Since imbalance was found between the poor and non-poor classes, an oversampling technique was employed. Imbalanced data affect classification, particularly when predicting the results of the classification. The following oversampling techniques were employed in this study: Random Oversampling (RO), the Adaptive Synthetic (ADASYN) and the Synthetic Minority Oversampling Technique (SMOTE). It was found that, of the three techniques, RO was the most efficient with $k = 5$, which yielded the best performance in terms of sensitivity, specificity, the G-mean, and accuracy reaching 0.643, 0.805, 0.719, and 78.873%, respectively. Therefore, it can be concluded that the classification model performed well enough to classify household welfare status, especially among the poor (minority class).

Key words: ADASYN, KNN, random oversampling, SMOTE, welfare.

1. Introduction

The development and progress made these days have become one of the challenges for countries to also develop and make progress over time, so these countries must make efforts to improve the welfare for their community. Welfare is a benchmark in social life, in which the society members are prosperous. It can be measured from the

[1] Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Islam Indonesia, Indonesia. E-mail: nur.djafar@students.uii.ac.id.

[2] Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Islam Indonesia, Indonesia. E-mail: achmadfauzan@uii.ac.id. ORCID: https://orcid.org/0000-0002-0533-5518.

society's health, economic conditions, happiness, and quality of life (Suud & Harsono, 2006). Welfare, however, has a close association with poverty and socio-economic disparities in society. Poverty is an economic, material, and physical inability to meet basic needs, including both food and non-food items, as measured by expenditure.

Poverty alleviation is the main responsibility for all central and regional governments throughout the world. Therefore, this is also the responsibility of the government of Indonesia, including the government of Kulon Progo Regency and the Province of the Special Region of Yogyakarta (DIY), Indonesia. Kulon Progo is one of the regencies in DIY which has 12 sub-districts and 88 villages. This regency has an area of 586.28 km2 and a population in the first half of 2021 of 442,838 people. In the last three years, the poverty rate in Kulon Progo had an increasing trend every year from 2019 to 2021, reaching 18.38%, which means that there was a total of 81,140 poor people in this regency. In addition, in 2021 Kulon Progo had the highest poverty rate compared to other regencies/city in DIY (BPS-Statistics of DI Yogyakarta Province, 2021).

In addition to using the monthly per capita consumption, according to research by Suryadarma et al. (2005), poverty can also be classified using other important indicators, namely information on asset ownership such as housing, employment status, family health, livestock ownership, food consumption, etc. In addition, it can also be identified from the Head of the Household (KRT), including age, gender, marital status, education level, the number of dependents, and income of the head of the household.

Based on the above-mentioned description, welfare-related issues, particularly poverty in Kulon Progo, is an interesting topic of discussion. The classification of household welfare status uses many indicators as influencing factors. A total of 17 variables were used in this study. The classification method was K-Nearest Neighbor (KNN), i.e. a non-parametric classification based on the closest neighbor according to the distance-based k value (Haseela H A, 2022). The KNN method was used in this study because KNN has several benefits, including a fast, simple, and effective training process although involving large training data. Such simplicity was used in this study to determine the results of classification using the research data. In addition, KNN can also classify data which contain categorical and numerical data.

The data on the welfare status of Kulon Progo showed that very few households were classified as poor compared to those classified as non-poor. The resulting classification, however, tends to classify the majority class well, but it has a poor performance in predicting the minority class, thus causing (Jian et al., 2016). One of the solutions to such imbalance is oversampling. Oversampling is a method to oversample the minority class data to be close to or equal to the majority class (Chawla, 2005). There are various oversampling methods, some of which were used in this research, including Random Oversampling (RO), Adaptive Synthetic Sampling (ADASYN), and Synthetic

Minority Oversampling Technique (SMOTE). The three oversampling techniques have different procedures and all of them were applied to the imbalanced data to determine their effectiveness. RO duplicates existing data, ADASYN uses weighted distribution, and SMOTE generates replicated data.

The study aimed to determine the general description of household welfare status data and to determine the comparison of the results of the KNN classification on data without oversampling and data with different oversampling techniques, namely RO, ADASYN, and SMOTE.

2. Method

2.1. Data

The study used data from the National Socioeconomic Survey (SUSENAS) by the Central Bureau of Statistics in Kulon Progo Regency, Indonesia in March 2021 accessed from http://silastik.bps.go.id. SUSENAS was carried out by direct interviews or self-administered questionnaires.

There were 18 variables used in this study, consisting of 17 independent variables defined as X and 1 dependent variable defined as Y, i.e. household welfare status which was classified into two categories, namely poor (0) and non-poor (1). The variables used are presented in Table 1.

No.	Variables	Type of data
1.	Welfare status	Nominal
2.	Age of head of household	Numerical
3.	Family size	Numerical
4.	Area of house	Numerical
5.	Gender of head of household	Categorical
6.	Marital status of head of household	Categorical
7.	Main source of household income	Categorical
8.	Type of home ownership	Categorical
9.	Latest education of head of household	Categorical
10.	Bank account ownership of head of household	Categorical
11.	Head of household works in the last one week	Categorical
12.	Head of household's health issues in the last one month	Categorical
13.	Main light source	Categorical
14.	Main source of energy used for cooking	Categorical
15.	Main source of drinking water	Categorical
16.	Main source of water for washing	Categorical
17.	Mobile phone ownership of head of household	Categorical
18.	Laptop/notebook ownership of head of household	Categorical

2.2. K-Nearest Neighbors (KNN)

KNN was developed by Evelyn Fix and Joseph Hodges in 1951. KNN is a non-parametric classification method based on the closest neighbor according to a distance-based k value (Haseela H A, 2022). KNN classification requires a distance that is in line with the type of research data (Wu et al., 2008). Several papers on the KNN technique and its development included Alsammak et al. (2020) conducting research to improve the performance of the K-Nearest Neighbor (KNN) classifier to satisfy emerging big data requirements. Kirtania et al. (2020) is working on a new adaptive KNN classifier for addressing imbalances in Magnetic Resonance Imaging (MRI) brain. Awotunde et al. (2022) investigated the feature choice based KNN Model for Rapid Software Defect Prediction. Hoque et al. (2021) created a KNN-DK classifier, which is a modified KNN classifier with dynamic *k* nearest neighbors.

Wilson & Martinez (1997a) explained that one of the distances for numerical and categorical data types is the Heterogeneous Euclidean-Overlap Metric (HEOM). HEOM handles both continuous and nominal attributes with overlap metric for nominal attributes and normalize Euclidean distance for linear attributes (ChitraDevi et al., 2012; Tusyakdiah, 2021). A heterogeneous distance function that uses different attribute distance functions on diverse categories of features has been used to address the issues of applications with continuous and nominal attributes. The unique technique is the overlap metric for combined nominal attributes and normalized Euclidean distance for linear features or numerical data (Dalatu & Midi, 2020). Numerical data are calculated using normalized Euclidean distance (Randall et al., 2000; Wilson & Martinez, 1997), written in equation (1).

$$D(x_{it}, z_{jt}) = \frac{|x_{it} - w_{jt}|}{\max(x_{it}) - \min(x_{it})}, \qquad t = 1,2,3, \dots, m_n \tag{1}$$

Categorical data are calculated using overlap metrics, written in equation (2).

$$D^2(x_{it}, z_{jt}) = \begin{cases} 0 & x_{it} = z_{jt} \\ 1 & x_{it} \neq z_{jt} \end{cases}, \quad t = 1,2,3, \dots, m_c \tag{2}$$

After the distances for both numerical and categorical data have been obtained, the square root of the sum of the two distances was calculated to obtain the HEOM distance in equation (3).

$$D_{HEOM}(x_i, z_j) = \sqrt{\sum_{t=1}^{m_n} D(x_{it}, z_{jt}) + \sum_{t=1}^{m_c} D^2(x_{it}, z_{jt})} \tag{3}$$

with D: distance, $\boldsymbol{x_{it}}$: training data value, $\boldsymbol{w_{jt}}$: data testing value, a: data variable to-$\boldsymbol{i}$, max: the maximum value of each numeric variable, min: the minimum value of each numeric variable, $\boldsymbol{m_n}$: numeric data type, and $\boldsymbol{m_c}$: categorical data type. The use of

HEOM distance will remove the impacts of arbitrary nominal value ordering, but it is an overly simplistic approach to dealing with nominal attributes that fails to make use of extra information offered by nominal attribute values that can aid in generalization (Wilson & Martinez, 1997). The usage of HEOM distance fits this study extremely well, in this case, since the data in this study are mixed data, as seen in Table 1.

2.3. Class Imbalance

Class imbalance is one of the problems in data mining. It is a condition where the minority class is very small compared to the majority class (Ren et al., 2017). In a classification with imbalanced data, the accuracy of the minority class tends to be low due to the dominance of the majority class, thus causing biases (Jian et al., 2016). In addition, class imbalance and noise can affect the quality of data in classification performance (Gao et al., 2014).

One of the solutions to class imbalance is resampling. Resampling is a preprocessing technique to balance data distribution to reduce the effect of class imbalance. There are three types of resampling, namely oversampling, undersampling, and hybrid, which combines both over and undersampling (Jian et al., 2016). Oversampling is used because of its benefits, i.e. adding data to the minority class to prevent the loss of data information.

2.4. Oversampling

Oversampling is a method to oversample the minority class data to be close to or equal to the majority class (Chawla, 2005). There are various oversampling methods, some of which were used in this research, including Random Oversampling (RO), Adaptive Synthetic Sampling (ADASYN), and Synthetic Minority Oversampling Technique (SMOTE).

2.4.1. Random Oversampling (RO)

Random Oversampling is a technique to randomly add data from the minority class to the training data, in which the addition process is repeated until the data in the minority class are equal in number to those in the majority class. The difference between the majority and minority classes is first calculated. Then, the repetition is randomly done as many times as the difference resulting from the calculation and added to the dataset.

2.4.2. Adaptive Synthetic Sampling (ADASYN)

ADASYN uses the weighted distribution of the data in the minority class, in which synthetic data are generated from the minority class (Rahayu et al., 2017). The steps for performing ADASYN are as follows (He et al., 2008).

i. Calculating the degree of data imbalance.

$$d_c = \frac{m_s}{m_l} \tag{4}$$

with $d \in [0.1]$. m_s is the number of minority class data and m_l is the number of majority class data.

ii. If $d_c < d_{th}$, where d_{th} is the threshold for the maximum degree of class imbalance, so:

1. Calculating the amount of synthetic data to be generalized for the minority class.

$$G = (m_l - m_s) \times \beta \tag{5}$$

where $\beta \in [0,1]$ is the parameter used to set the level, balance expected after the synthetic data has been generalized. $\beta = 1$ means completely balanced data after the generalization.

2. For each $x_i \in$ minority class, determining the k-nearest neighbor based on *HEOM* distance in equation (3) in n dimensional space and calculating ratio (r_i).

$$r_i = \frac{\Delta_i}{K}, i = 1, \dots, m_s \tag{6}$$

with r_i: ration, Δ_i: the number of samples in KNN but from data that include all classes except the minority and K: the number k in KNN, and interval of $r_i \in [0,1]$,

3. Normalizing r_i

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m} r_i} \tag{7}$$

4. Calculating the amount of synthetic data to be generated for each x_i in the minority class.

$$g_i = \hat{r}_i \times G \tag{8}$$

G is the total amount of synthetic data to be generated for the minority class in equation (5).

5. For each x_i in the minority class, generating synthetic data as many times as g_i by making repetition from 1 to g_i with the steps as follows:

iii. Randomly selecting one of the data in the minority class from x_{knn} in data x_i

iv. Generating synthetic data by equation (9).

$$x_{adasyn} = x_i + (x_{knn} - x_i) \times \lambda \tag{9}$$

where λ is random [0,1]

2.4.3. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is done by adding more data in the minority class by generating synthetic or artificial data. The synthetic data are generated based on the attributes from the k-nearest neighbor. Several studies related to oversampling with the SMOTE technique

include Elreedy & Atiya (2019), Li et al. (2021), Noorhalim et al. (2019), and Srinilta & Kanharattanachai (2021). The steps for performing SMOTE are as follows (Chawla, 2005):

i. Determining the data to be replicated (x_i) from a randomly selected minority class.

ii. Determining the value of k (the number of the nearest neighbors), then calculating the distance from data x_i to the nearest neighbor data (x_{knn}) in the same minority class.

iii. For each x_{knn} that is selected, calculating the difference between x_i and x_{knn}, then multiplying the difference by a random number [0,1], and adding it to the features under study.

$$x_{sintesis} = x_i + (x_{knn} - x_i) \times \delta \tag{10}$$

where δ is random [0,1]

2.5. Validation Technique

In data mining, there are many techniques to measure the performance of an algorithm, one of which is confusion matrix, which is a binary table, where classes are divided into 2 categories (Pramana et al., 2018). Confusion matrix is a table that states the amount of testing data correctly classified, and the amount of testing data misclassified (Indriani, 2014).

Table 2: Confusion Matrix

Real Classes	Predicted Classes	
	Poor	Non-Poor
Poor	TP	FN
Non-Poor	FP	TN

True Positive (TP) is the number of poor classes predicted to be poor; False Positive (FP) is the number of non-poor classes predicted to be poor; False Negative (FN) is the number of poor classes predicted to be non-poor; True Negative (TN) is the number of non-poor classes predicted to be poor.

Based on the confusion matrix, several evaluation metrics including accuracy, sensitivity, specificity, and G-mean can be derived.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \times 100\% \tag{10}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

$$G - mean = \sqrt{Sensitivity \times Specificity} \tag{13}$$

Accuracy is a comparison value between correctly classified data and real data. This measurement is used to measure the correctness of the classification (Hamel, 2009). The higher the accuracy value, the better the resulting classification (Widayati et al., 2021). The value of classification accuracy needs to be increased as a measure of standard criteria, especially in cases of class imbalance. Because it will produce good accuracy only for the majority class, while the resulting predictions will be dire for the minority class (Pangastuti, 2018), evaluation of the performance of the method as a whole can be done using the geometric mean (Kubát & Matwin, 1997).

Sensitivity is a value that shows the results of the actual data classification, which is a positive class, and the predicted value is also a positive class (Hamel, 2009). The higher the sensitivity value, the less likely the results of the positive class classification are wrong (Zhu et al., 2010). Specificity is a value that shows the results of the actual data classification, which is labeled negative, and the predicted value is also labeled negative (Jahangiri et al., 2020). The higher the specificity value, the better the classification performance for making predictions because it has low false positives (Maxim et al., 2014). A good classification is a classification that has high sensitivity and specificity values (Zhu et al., 2010). The G-mean is the geometric average value used to measure overall performance. Poor classification results will produce a small G-mean value (Bekkar et al., 2013).

3. Results and Discussion

Based on BPS-Statistics of DI Yogyakarta Province data, there are 71 households in the poor category and 638 households in the non-poor category, or it can be said that there is an imbalance in the data. KNN classification was performed by dividing 80% for the training data and 20% for the testing data. Oversampling was performed using 567 training data.

3.1. RO, ADASYN, and SMOTE Oversampling

RO was done by making 447 repetitions by randomly taking data from the minority class. The ADASYN was performed using $k = 11$ and $\beta = 1$ (with the expected balance of 100%) using the HEOM distance to determine the nearest neighbor. Synthesized data were generated by taking from the poor class, resulting in an addition of 450 data. SMOTE was performed on the data from the minority class using $k = 11$ and generated 8 times from the total observations of the poor class (8×57) using the HEOM distance to determine the nearest neighbor. The replicated data were equivalent to the non-poor class, namely 399 data. Illustration of data results after oversampling and without oversampling is presented in Figure 1.

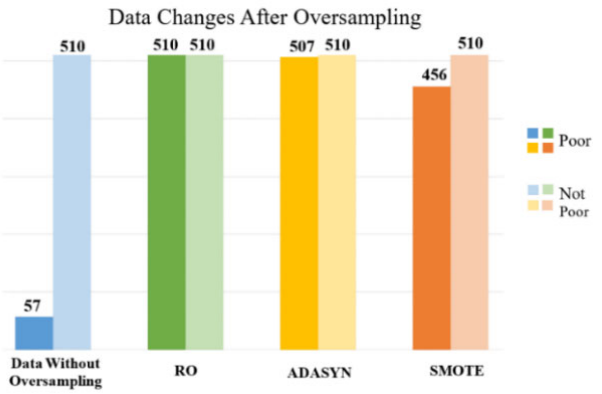

Figure 1: Illustration of data after oversampling.

3.2. KNN with HEOM Distance

The classification using *KNN* with *HEOM* distance was simulated using the data as shown in Table 3.

Table 3: Simulation Data

Data	x_{i1}	x_{i2}	x_{i3}	x_{i4}	…	x_{i15}	x_{i16}	x_{i17}	Y
Training data 1 (x_{1t})	65	1	56	0	…	2	0	1	0
Training data 2 (x_{2t})	57	2	96	0	…	4	0	1	1
Training data 3 (x_{3t})	35	1	87	0	…	3	0	1	1
Training data 4 (x_{4t})	62	0	54	0	…	4	0	1	0
Testing data 1 (z_{1t})	45	1	184	1	…	4	0	1	? (predicted)

Based on Table 3, the household welfare status in testing data 1 (z_{1t}) was predicted, whether it was classified as either poor or non-poor class. The distances for the numerical data, namely X_1, X_2, and X_3 were firstly calculated. The distance calculation used the normalized Euclidean distance according to equation (1).

The distance between testing data (z_{1t}) and training data 1 (x_{1t})

$$\sum_{t=1}^{3} D^2(x_{1t}, z_{1t}) = \left(\frac{|65-45|}{97-21}\right)^2 + \left(\frac{|1-1|}{5-0}\right)^2 + \left(\frac{|56-184|}{240-9}\right)^2 = 0.57$$

The distance between testing data (z_{1t}) and training data 2 (x_{2t})

$$\sum_{t=1}^{3} D^2(x_{2t}, z_{1t}) = \left(\frac{|57-45|}{97-21}\right)^2 + \left(\frac{|2-1|}{5-0}\right)^2 + \left(\frac{|96-184|}{240-9}\right)^2 = 0.21$$

The distance between testing data (z_{1t}) and training data 3 (x_{3t})

$$\sum_{t=1}^{3} D^2(x_{3t}, z_{1t}) = \left(\frac{|35-45|}{97-21}\right)^2 + \left(\frac{|1-1|}{5-0}\right)^2 + \left(\frac{|87-184|}{240-9}\right)^2 = 0.193$$

The distance between testing data (z_{1t}) and training data 4 (x_{4t})

$$\sum_{t=1}^{3} D^2(x_{4t}, z_{1t}) = \left(\frac{|62-45|}{97-21}\right)^2 + \left(\frac{|0-1|}{5-0}\right)^2 + \left(\frac{|54-184|}{240-9}\right)^2 = 0.407$$

Once the distance for the numerical data had been obtained, the distance for the categorical data on variables X_4 to X_{17} had to be calculated using the overlap metric method by observing the incompatibility of the two vectors, i.e. if vector x_i was different from vector z_j, then the value = 1.

The *HEOM* distance was calculated by calculating distances (z_{1t}) and (x_{it}) in each categorical variable based on equation (2).

The distance between testing data (z_{1t}) and training data 1 (x_{1t})

$$\sum_{t=4}^{17} D^2\,(x_{1t}, z_{1t}) = [D^2(x_{14}, z_{14}) + D^2(x_{15}, z_{15}) + \cdots + D^2(x_{1,17}, z_{1,17})$$
$$= [D^2(0,1) + D^2(1,3) + D^2(0,0) + \cdots + D^2(1,0)] = [1^2 + 1^2 + 0^2 + \cdots + 1^2] = 9$$

The distance between testing data (z_{1t}) and training data 2 (x_{2t})

$$\sum_{t=4}^{17} D^2(x_{2t}, z_{1t}) = [D^2(x_{24}, z_{14}) + D^2(x_{25}, z_{15}) + D^2(x_{26}, z_{16}) + \cdots + D^2(x_{2,17}, z_{1,17})$$
$$= [D^2(0,1) + D^2(3,3) + D^2(0,0) + \cdots + D^2(1,0)] = [1^2 + 0^2 + 0^2 + \cdots + 1^2] = 6$$

The distance between testing data (z_{1t}) and training data 3 (x_{3t})

$$\sum_{t=4}^{17} D^2(x_{3t}, z_{1t}) = [D^2(x_{34}, z_{14}) + D^2(x_{35}, z_{15}) + D^2(x_{36}, z_{16}) + \cdots + D^2(x_{3,17}, z_{1,17})$$
$$= [D^2(0,1) + D^2(1,3) + D^2(0,0) + \cdots + D^2(1,0)] = [1^2 + 1^2 + 0^2 + \cdots + 1^2] = 7$$

The distance between testing data (z_{1t}) and training data 4 (x_{4t})

$$\sum_{t=4}^{17} D^2(x_{4t}, z_{1t}) = [D^2(x_{44}, z_{14}) + D^2(x_{45}, z_{15}) + D^2(x_{46}, z_{16}) + \cdots + D^2(x_{4,17}, z_{1,17})$$
$$= [D^2(0,1) + D^2(0,3) + D^2(0,0) + \cdots + D^2(1,0)] = [1^2 + 1^2 + 0^2 + \cdots + 1^2] = 7$$

After the distance for the numerical data had been obtained using the normalized difference and for the categorical data using the overlap metric, the two distances were combined to obtain the overall distance (x_i, z_j) according to equation (3).

$$\text{HEOM Dist } (x_1, z_1) = \sqrt{0.57 + 9} = 3.094$$
$$\text{HEOM Dist } (x_2, z_1) = \sqrt{0.21 + 6} = 2.492$$
$$\text{HEOM Dist } (x_3, z_1) = \sqrt{0.193 + 7} = 2.682$$
$$\text{HEOM Dist } (x_4, z_1) = \sqrt{0.407 + 7} = 2.722$$

After the distance of each object had been obtained, the classification using *KNN* was done with the specified k. The following is an illustration for $k = 1$ and $k = 3$ using R Programming.

Table 4: Classification Results of Simulation

Data	Y (Welfare Status)	Distance	$k = 1$	$k = 3$
(x_1, z_1)	Poor	3.094		
(x_2, z_1)	Non-poor	2.492	Non-poor	Non-poor
(x_3, z_1)	Non-poor	2.682		Non-poor
(x_4, z_1)	Poor	2.722		Poor

To classify using KNN, the classification results were obtained from the training data with the nearest distance to the testing data. Using $k = 1$, the classification results were non-poor because the nearest distance was training data 2 with category 1 (non-poor). Using $k = 3$, the classification results were still non-poor because the nearest distance was training data 2 and 3 with category 1 (non-poor), while training data 4 with category 0 (poor) only had 1 data. In other words, the dominant class was used as the classification result.

The classification using KNN was carried out four times with four data, namely data without any oversampling treatment (imbalanced data), data with RO, ADASYN, and SMOTE treatment. Validation was carried out with various values of k (3,5,7,9) based on the confusion matrix obtained, by using the R program the results are presented in Table 5.

Table 5: Classification Results

k	Data	Sensitivity	Specificity	G-mean	Accuracy
3	Data without oversampling	0.071	1	0.267	90.845%
	RO	0.500	0.867	0.658	83.099%
	ADASYN	0.500	0.836	0.647	80.281%
	SMOTE	0.429	0.820	0.593	78.170%
5	Data without oversampling	0	1	0	90.141%
	RO	0.643	0.805	0.719	78.873%
	ADASYN	0.500	0.852	0.653	81.691%
	SMOTE	0.500	0.773	0.622	74.648%
7	Data without oversampling	0	1	0	90.141%
	RO	0.714	0.734	0.724	73.240%
	ADASYN	0.500	0.836	0.647	80.282%
	SMOTE	0.500	0.758	0.616	73.240%
9	Data without oversampling	0	1	0	90.141%
	RO	0.786	0.664	0.722	67.606%
	ADASYN	0.500	0.820	0.640	78.873%
	SMOTE	0.643	0.688	0.665	68.310%

It was insufficient to validate the results of the classification through the accuracy due to the imbalanced data between the poor and non-poor classes. Based on Table 5, the data without any oversampling treatment obtained high accuracy and sensitivity, but very low specificity. This means that the model was very bad for the non-poor class classification, yet very good for the poor class classification. Thus, the data without any oversampling treatment were not good in this classification because the results of the classification were dominated by the accuracy of the minority class.

The G-mean is one of the best measurements for evaluating classification, especially in class imbalances in data (Pristyanto et al., 2018). Based on Table 5, all G-mean values in the data that were not oversampled resulted in low values with k values of 3, 5, 7, and 9. Meanwhile, if using data that had been treated RO, ADASYN, and SMOTE with a value of k=5 produces a G-mean of 0.719, 0.653, and 0.622, which means that with balanced data, the resulting classification is good enough for the poor and non-poor classes.

Based on Table 5, it can also be said that with oversampling, the RO oversampling technique with a value of k=5 gives the best results when viewed from the G-mean, accuracy, and ease of forming nearest neighbors. With a G-mean value and accuracy of 0.719 and 78.873%. After calculating the accuracy, sensitivity, specificity, and G-mean, the best classification was generated with the RO-treated data. The model can classify object classes well, especially in the welfare status classification. The findings of this study are also consistent with the findings of several other studies, including Akbar et al. (2019), Hussain et al. (2022), and Xin & Rashid (2021) which specify the performance of k-NN sensitivity values more precisely. Furthermore, research from Islam et al. (2022) and Shi (2020) dealing with imbalance data with oversampling approaches raises the value of precision.

4. Conclusions

This research begins by performing oversampling techniques with three methods, including Random Oversampling (RO), Adaptive Synthetic Sampling (ADASYN), and Synthetic Minority Oversampling Technique (SMOTE), and comparing with data without oversampling techniques. Since the data used are mixed (numeric and categorical), the Heterogeneous Euclidean-Overlap Metric (HEOM) distance is more appropriate for calculating the distance. Then, from the oversampling results, classification is carried out using the k-nearest neighbors (KNN) method with various simulated k values. With the division of 80% training data and 20% testing data, the classification using KNN with k = 5 and the HEOM distance produced the best results on data with a RO treatment. This is evident from the sensitivity, specificity, G-mean, and accuracy i.e. 0.643, 0.805, 0.719, and 78.873% respectively. This means that the classification model was quite good in classifying welfare status, especially in the minority class (poor class).

Acknowledgement

We thank all the parties who have provided support and funding for this research.

References

Akbar, S., Hayat, M., Kabir, M., and Iqbal, M., (2019). iAFP-gap-SMOTE: An Efficient Feature Extraction Scheme Gapped Dipeptide Composition is Coupled with an Oversampling Technique for Identification of Antifreeze Proteins. *Letters in Organic Chemistry*, *16*(4), pp. 294–302. https://doi.org/10.2174/1570178615666180816101653

Alsammak, I. L. H., Sahib, H. M. A., and Itwee, W. H., (2020). An Enhanced Performance of K-Nearest Neighbor (K-NN) Classifier to Meet New Big Data Necessities. *IOP Conference Series: Materials Science and Engineering*, *928*(3). https://doi.org/10.1088/1757-899X/928/3/032013

Awotunde, J. B., Misra, S., Adeniyi, A. E., Abiodun, M. K., Kaushik, M., and Lawrence, M. O., (2022). A Feature Selection-Based K-NN Model for Fast Software Defect Prediction. In O. Gervasi, B. Murgante, S. Misra, A. M. A. C. Rocha, & C. Garau (Eds.), *Computational Science and Its Applications – ICCSA 2022 Workshops,* pp. 49–61. Springer International Publishing.

Bekkar, M., Djemaa, H. K., and Alitouche, T. A., (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*, *3*, pp. 27–38.

BPS-Statistics of DI Yogyakarta Province, (2021). *Persentase Penduduk Miskin menurut Kabupaten/Kota di Provinsi DI Yogyakarta (Persen), 2009-2021*.

Chawla, N. V., (2005). Data Mining for Imbalanced Datasets: An Overview. In L. Maimon Oded and Rokach (Ed.), *Data Mining and Knowledge Discovery Handbook* (pp. 853–867). Springer US. https://doi.org/10.1007/0-387-25465-X_40

ChitraDevi, N., Palanisamy, V., Baskaran, K., and Prabeela, S., (2012). A Novel Distance for Clustering to Support Mixed Data Attributes and Promote Data Reliability and Network Lifetime in Large Scale Wireless Sensor Networks. *Procedia Engineering*, *30*, pp. 669–677. https://doi.org/10.1016/j.proeng.2012.01.913

Dalatu, P. I., Midi, (2020). Modified Statistical Approach for Data Preprocessing to Improve Heterogeneous Distance Functions. In *Malaysian Journal of Mathematical Sciences* (Vol. 14, Issue 2).

Elreedy, D., Atiya, A. F., (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, pp. 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Gao, K., Khoshgoftaar, T. M., and Wald, R., (2014). Combining Feature Selection and Ensemble Learning for Software Quality Estimation. *The Florida AI Research Society*.

Hamel, L., (2009). Model Assessment with ROC Curves. In *Encyclopedia of Data Warehousing and Mining, Second Edition,* pp. 1316–1323. IGI Global. https://doi.org/10.4018/978-1-60566-010-3.ch204

Haseela H A., (2022). Hybrid Method for Image Classification. *EPRA International Journal of Research and Development (IJRD)*, *7*(2), pp. 59–61. https://doi.org/10.36713/epra2016

He, H., Bai, Y., Garcia, E. A., and Li, S., (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Hoque, N., Bhattacharyya, D. K., and Kalita, J. K., (2021). KNN-DK: A Modified K-NN Classifier with Dynamic k Nearest Neighbors. In J. C. Bansal, L. C. C. Fung, M. Simic, & A. Ghosh (Eds.), *Advances in Applications of Data-Driven Computing,* pp. 21–34. Springer Singapore. https://doi.org/10.1007/978-981-33-6919-1_2

Hussain, L., Lone, K. J., Awan, I. A., Abbasi, A. A., and Pirzada, J.-R., (2022). Detecting congestive heart failure by extracting multimodal features with synthetic minority oversampling technique (SMOTE) for imbalanced data using robust machine learning techniques. *Waves in Random and Complex Media*, *32*(3), pp. 1079–1102. https://doi.org/10.1080/17455030.2020.1810364

Indriani, A., (2014). Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI) Yogyakarta*. www.bluefame.com,

Islam, A., Belhaouari, S. B., Rehman, A. U., and Bensmail, H., (2022). KNNOR: An oversampling technique for imbalanced datasets. *Applied Soft Computing*, *115*, 108288. https://doi.org/10.1016/j.asoc.2021.108288

Jahangiri, M., Jahangiri, M., and Najafgholipour, M., (2020). The sensitivity and specificity analyses of ambient temperature and population size on the transmission rate of the novel coronavirus (COVID-19) in different provinces of Iran. *Science of The Total Environment*, *728*, 138872. https://doi.org/10.1016/j.scitotenv.2020.138872

Jian, C., Gao, J., and Ao, Y., (2016). A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble. *Neurocomput.*, *193*(C), pp. 115–122. https://doi.org/10.1016/j.neucom.2016.02.006

Kirtania, R., Mitra, S., and Shankar, B. U., (2020). A novel adaptive k-NN classifier for handling imbalance: Application to brain MRI. *Intelligent Data Analysis*, *24*, pp. 909–924. https://doi.org/10.3233/IDA-194647

Kubát, M., Matwin, S., (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *International Conference on Machine Learning*.

Li, J., Zhu, Q., Wu, Q., and Fan, Z., (2021). A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. *Information Sciences*, *565*, pp. 438–455. https://doi.org/10.1016/j.ins.2021.03.041

Maxim, L. D., Niebo, R., and Utell, M. J., (2014). Screening tests: a review with examples. *Inhalation Toxicology*, *26*(13), pp. 811–828. https://doi.org/10.3109/08958378.2014.955932

Noorhalim, N., Ali, A., and Shamsuddin, S. M., (2019). Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE. In L.-K. Kor, A.-R. Ahmad, Z. Idrus, & K. A. Mansor (Eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017),* pp. 19–30. Springer Singapore.

Pangastuti, S. S., (2018). *Perbandingan Metode Ensemble Random Forest dengan Smote-Boosting dan Smote-Bagging pada Klasifikasi Data Mining untuk Kelas Imbalance (Studi Kasus: Data Beasiswa Bidikmisi Tahun 2017 di Jawa Timur)*. Institut Teknologi Sepuluh Nopember.

Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I., and Nooraeni, R., (2018). *Data Mining dengan R: Konsep Serta Implementasi*. IN MEDIA.

Pristyanto, Y., Pratama, I., and Nugraha, A. F., (2018). Data level approach for imbalanced class handling on educational data mining multiclass classification. *2018 International Conference on Information and Communications Technology (ICOIACT)*, pp. 310–314. https://doi.org/10.1109/ICOIACT.2018.8350792

Rahayu, S., Bharata Adji, T., Akhmad Setiawan, N., and Teknik Elektro dan Teknologi Informasi, D., (2017). Penghitungan k-NN pada Adaptive Synthetic-Nominal (ADASYN-N) dan Adaptive Synthetic-kNN (ADASYN-kNN) untuk Data Nominal-Multi Kategori. *Ktrl.Inst (J.Auto.Ctrl.Inst)*, *9*(2).

Randall, D., And, W., and Martinez, T. R., (2000). An Integrated Instance-Based Learning Algorithm. *Computational Intelligence*, *16*(1).

Ren, F., Cao, P., Li, W., Zhao, D., and Zaiane, O., (2017). Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm. *Computerized Medical Imaging and Graphics*, *55*, pp. 54–67. https://doi.org/https://doi.org/10.1016/j.compmedimag.2016.07.011

Shi, Z., (2020). Improving k-Nearest Neighbors Algorithm for Imbalanced Data Classification. *IOP Conference Series: Materials Science and Engineering*, *719*(1), 012072. https://doi.org/10.1088/1757-899X/719/1/012072

Srinilta, C., Kanharattanachai, S., (2021). Application of Natural Neighbor-based Algorithm on Oversampling SMOTE Algorithms. *2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST)*, pp. 217–220. https://doi.org/10.1109/ICEAST52143.2021.9426310

Suryadarma, D., Akhmadi, Hastuti, and Toyamah, N., (2005). *Objective measures of family welfare for individual targeting: results from pilot project on community based monitoring system in Indonesia*. SMERU Research Institute.

Suud, M., Harsono, (2006). *3 Orientasi Kesejahteraan Sosial*. Prestasi Pustaka.

Tusyakdiah, H., (2021). *Implementasi K Nearest Neighbor (KNN) dalam Klasifikasi Status Kerja Lulusan Sekolah Menengah Kejuruan (SMK) dengan Oversampling Synthetic Minority Oversampling Technique (SMOTE) dan Adaptive Synthetic (ADASYN)*. Universitas Islam Indonesia.

Widayati, Y. T., Prihati, Y., and Widjaja, S., (2021). Analisis dan Komparasi Algoritma Naïve Bayes dan C4.5 untuk Klasifikasi Loyalitas Pelanggan MNC Play Kota Semarang. *TRANSFORMTIKA*, *18*(2), pp. 161–172.

Wilson, D. R., Martinez, T. R., (1997). Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, *6*, pp. 1–34.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D., (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), pp. 1–37. https://doi.org/10.1007/s10115-007-0114-2

Xin, L. K., and Rashid, N. binti A., (2021). Prediction of Depression among Women Using Random Oversampling and Random Forest. *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, pp. 1–5. https://doi.org/10.1109/WiDSTaif52235.2021.9430215

Zhu, W., Zeng, N. F., and Wang, N., (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS. *Northeast SAS Users Group 2010: Health Care and Life Sciences*.