JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE
VOLUME 3, No 1, March 2024
e-ISSN: 2828-9986
https://ejournal.unib.ac.id/index.php/jsds/index

Classification of Hypertension Patients in Palembang by K-Nearest Neighbor and Local Mean K-Nearest Neighbor

Rosdiana[1*]
[1]Statistics Indonesia of North Sumatera, Indonesia
*Corresponding Author: rosdianabarlian@gmail.com

Article Info

Article History:
Received: January, 11 2024
Accepted: May, 15 2024
Available Online: August, 21 2024

Key Words:
Classification
KNN Method
LMKNN Method

Abstract

Classification is a multivariate technique for separating different data sets from an object and allocating new objects into predefined groups. Several methods that can be used to classify include the *k*-Nearest Neighbor (KNN) and Local Mean *k*-Nearest Neighbor (LMKNN) methods. The KNN method classifies objects based on the majority voting principle, while LMKNN classifies objects based on the local average vector of the k nearest neighbors in each class. In this study, a comparison was made on the results of classifying hypertensive patient data at the Merdeka Health Center in Palembang City with the KNN and LMKNN methods by looking at the accuracy and the smallest APER value produced. The results showed that by using the same proportion of training and testing data and choosing different k values, the results of classifying hypertension patient data at the Merdeka Health Center in Palembang City with the KNN and LMKNN methods resulted in the APER value or the same error rate and accuracy, namely sequentially equal to 0.0303 and 96.97%.

1. INTRODUCTION

Big data or "big data" is a term popularized by Fremont Rider, an American librarian from Westleyan University in 1944. At that time, he speculated that the volume of university collections in America would reach 200 million copies in 2040, so several issues emerged. -issues, such as the addition of large data users, storage capacity, and the need to have data analysis [1]. Currently, the term increase is known as big data. Big data or large chunks of data demand to be explored optimally to obtain new information. These needs can be answered by data mining.

In data mining, large data is processed to produce information, as well as quantitative or qualitative information that can show facts [2]. The data mining function is for classification, clustering, association, sorting and forecasting. Classification as a method in data mining, identifies facts or conclusions based on data filtering to explore data patterns in order to obtain a definition of the characteristics of a group of data as a phenomenon. This data mining function can be applied to examples of cases in the health sector, such as the classification of data on hypertension sufferers.

Hypertension is currently a disease that is no longer foreign to ears. The number of hypertensive sufferers continues to increase from year to year, of course, causing the volume of data on hypertensive patients in various parts of the world to also increase. In Indonesia, the prevalence of hypertension ranges from 6-15%. Hypertension has also been named the main cause of premature death in the world. This disease tends to damage body organs, such as the heart, kidneys, brain, eyes and other organs [3]. The nickname silent killer of hypertension also does not occur without reason, this is because hypertension is a disease that is difficult to detect and manage.

At the initial examination for hypertension, the patient conveys his complaint to the medical officer and all this data is usually recorded in a medical report/record, known as the medical record. Medical record data records the unique variables of each patient with hypertension. If viewed based on class, hypertension is divided into two groups, namely primary (essential) hypertension and secondary hypertension. Hypertension can be detected early by checking the patient's blood pressure, age, weight and cholesterol levels. However, in reality, after anamnesis (initial examination of the patient) is carried out, the existing medical record data is not used to classify the type of hypertension present. In other words, the initial management of hypertension in patients is not completely optimal.

Follow-up is required in the form of classifying hypertension on the initial examination of the patient. This will of course greatly influence the treatment and therapy given to patients as well as an effort to prevent the emergence of other complications/complications in the future. Finally, proper treatment will of course also have an effect on reducing the number of hypertension sufferers. Therefore, in this study, the classification of hypertension sufferers was carried out at the Merdeka Health Center in Palembang City using the k-Nearest Neighbor (KNN) method and the Local Mean k-Nearest Neighbor (LMKNN) method which is expected to determine the classification of hypertension sufferers.

2. METHOD

The method in this research compares the KNN and LMKNN methods.

2.1 KNN Methods

The KNN method performs grouping based on voting for the same votes or based on the number of k nearest neighbors from the existing training data. Choosing the right k value can reduce the error rate. The data analysis stages using the KNN Method [4] are as follows:

1. Calculate the distance of each testing data to all training data
2. Sort the distance values from smallest to largest
3. Determine the value of *k*
4. Determine the class that appears most out of *k* nearest neighbor as a class of testing data

2.2 LMKNN Methods

The LMKNN method is an extension of the KNN method. According to Mitani and Hamamoto [5] LMKNN is simple, effective and is a robust non-parametric classifier. LMKNN has also been proven to be able to improve classification performance and reduce the influence of outliers, especially in conditions of small sample sizes [6]. Gou et al [7] explained that the stages of analysis with LMKNN are:

1. Look for the *k* nearest neighbor value of the set T_i of each class c_i for each value of x. Example $\mathrm{T}_{\mathrm{ik}}^{\mathrm{NN}}(\mathbf{x}) = \left\{\mathbf{x}_{\mathrm{ij}}^{\mathrm{NN}} \in \mathbb{R}^{\mathrm{m}}\right\}_{\mathrm{j}=1}^{\mathrm{k}}$ is a kNN set of x in class c_i using the Euclidean distance matrix, with the value $k \leq N_i$ with the formula

$$\mathrm{d}\left(\mathbf{x}, \mathbf{x}_{\mathrm{ij}}^{\mathrm{NN}}\right) = \sqrt{\left(\mathbf{x} - \mathbf{x}_{\mathrm{ij}}^{\mathrm{NN}}\right)^{\mathrm{T}}\left(\mathbf{x} - \mathbf{x}_{\mathrm{ij}}^{\mathrm{NN}}\right)} \quad (1)$$

2. Calculates the local mean vector $\mathbf{u}_{\mathrm{ik}}^{\mathrm{NN}}$ from class c_i using $\mathrm{T}_{\mathrm{ik}}^{\mathrm{NN}}(\mathbf{x})$ set. With

$$\mathbf{u}_{\mathrm{ik}}^{\mathrm{NN}} = \frac{1}{\mathrm{k}} \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathbf{x}_{\mathrm{ij}}^{\mathrm{NN}} \quad (2)$$

3. Allocate x to class c_i if the Euclidean distance between the local mean vectors for c_i with respect to x is the minimum. Where $\mathrm{c} = \arg\min_{c_i}\left(\mathbf{x} - \mathbf{u}_{\mathrm{ik}}^{\mathrm{NN}}\right)^{\mathrm{T}}\left(\mathbf{x} - \mathbf{u}_{\mathrm{ik}}^{\mathrm{NN}}\right)$ (3)

2.3 Confussion Matrix

According to Prasetyo [8] a system that carries out classification is expected can classify all data sets correctly, but it is undeniable that the performance of a system cannot be 100% correct, so a system classification performance must also be measured. Generally, classification performance measures carried out using a confusion matrix. The confusion matrix is a table recorder of classification work results. To see the accuracy of the classification results from the KNN and LMKNN methods, then need to calculate the confusion matrix. Following is the confusion matrix:

Table 1. Confusion matrix

Predicted	Observed		
		Positive Class	Negative Class
	Positive Class	True Positive (TP)	False Positive (FP)
	Negative Class	False Negative (FN)	True Negative (TN)

2.4 Evaluation of the Accuracy of Classification Results

According to Johnson and Wichern [9], a good classification method will produce few classification errors. To determine the accuracy of the classification, Apparent Error Rate (APER) is used. Apparent Error Rate (APER)

is a value used to see the opportunity for errors in classifying objects. APER is a classification evaluation procedure that does not depend on the shape of the population. The APER value states the proportion of samples that are misclassified. The best method is the method that has the smallest APER value so that the method has the greatest classification accuracy. The APER value is defined as follows:

$$APER = \frac{the\ amount\ of\ incorrectly\ predicted\ data}{number\ of\ predictions\ made} \quad (5)$$

$$Acuration = 1\text{-APER} \quad (6)$$

3. RESULTS AND DISCUSSION

3.1 Description of Data

This research uses secondary data derived from medical record data of 662 hypertension sufferers who underwent examinations at the Merdeka Health Center, Palembang City from January to December 2020. Of the 662 people who suffered from hypertension, 587 or 88.67% suffered from primary hypertension, while the remaining 75 people or 11.33% suffered from secondary hypertension.

3.2 Calculation of Distances

Before carrying out data analysis using the KNN and LMKNN methods, the data pre-processing stage is first carried out, namely determining the proportion of training data and testing data and standardizing the two groups of data. Determining the proportion of training data and testing data greatly influences the classification results and APER values obtained. Data proportions also aim to ensure that each data in each class has the same proportion (proportional) and has an even distribution of data. Meanwhile, data standardization aims to eliminate any negative influences originating from differences in attribute units and different value ranges in research variables. In this study, from 662 data, there were 587 patients with class 1 (primary) hypertension and the remaining 75 patients with class 2 (secondary) hypertension. So that 80% of the training data for classes 1 and 2 are respectively 470 and 60, while 20% of the data testing for classes 1 and 2 were 117 and 15 respectively.

3.2.1 KNN Method

In the KNN method, determining the proportion of training data and test data as well as parameter k is carried out by testing each odd parameter k for each proportion of training data and test data. The results of these trials are summarized in Table 2 below.

Table 2. Summary of APER Values for all data proportions

k	Training Data Proportion: Testing Data Proportion			
	90:10	**80:20**	**70:30**	**60:40**
1	0	0,0758	0,0606	0,0642
3	0	**0,0303**	0,0404	0,0453
5	0	0,0379	0,0505	0,0528
7	0	0,0379	0,0505	0,0491
9	0	0,0379	0,0505	0,0528
11	0	0,0379	0,0505	0,0528
13	0,0303	0,0379	0,0556	0,0528
15	0,0303	0,0455	0,0556	0,0528

The table above shows that the smallest APER value is found in the data proportion of 90:10 with k = 1, but this proportion cannot be used because it produces high fluctuating APER values, so for the KNN method the proportion of training and test data was chosen as 80% and 20%. for each hypertension class with parameter k = 3. Next, standardization is carried out on the training data and test data using the following formula:

$$Z_X = \frac{X - \bar{X}}{s_X} \text{ dan } Z_y = \frac{Y - \bar{Y}}{s_Y} \quad (7)$$

where X is training data and Y is test data. To carry out this standardization, the average, variance and standard deviation for the training data and test data were calculated and the following results were obtained:

Table 3. Summary of Training Data Variables

Variables	Mean ($\overline{X}_i$)	Variance (s_i^2)	Standard Deviation (s_2)
X_1	53,66	89,22	9,45
X_2	57,12	111,52	10,56
X_3	128,98	341,81	18,49
X_4	82,18	59,41	7,71
X_5	197,13	157,43	12,55

The following illustrates the calculation of training data:

$$Z_{X_{11}} = \frac{79-53,66}{9,45} = 2,6828, \ldots,$$

$$\vdots$$

$$Z_{X_{1\,530}} = \frac{43-53,66}{9,45} = -1,1286$$

$$\vdots$$

$$Z_{X_{5\,530}} = \frac{180-197,13}{12,55} = -1,3656$$

With the same formula, standardization is then carried out on the test data for the KNN method. The parameters used to standardize test data are as follows:

Table 4. Summary of Training Data Variables

Variable	Mean ($\overline{X}_i$)	Variance (s_i^2)	Standard Deviation (s_2)
X_1	50,32	50,55	7,11
X_2	56,42	98,21	9,91
X_3	126,29	270,93	16,46
X_4	81,44	32,26	5,68
X_5	193,06	152,03	12,33

The following illustrates the calculation of test data:

$$Z_{Y_{11}} = \frac{48-50,32}{7,11} = -0,3261$$

$$Z_{Y_{1\,132}} = \frac{51-50,32}{7,11} = 0,0959$$

$$\vdots$$

$$Z_{Y_{5\,132}} = \frac{225-193,06}{12,33} = 2,907$$

Next, after obtaining standardized training data and testing data, the Euclidean distance is calculated using the following formula:

$$\textit{Euclidean Distance} = D(\mathbf{x_i}, \mathbf{x_j}) = \sqrt{\sum_{p=1}^{n}(x_{ip} - x_{jp})^2} \quad (8)$$

$$D_{1,1} = \sqrt{(-0,3261-(2,6828))^2 + \cdots + (0,5629-(2,2209))^2} = 3,688$$

$$D_{1,2} = \sqrt{(-0,3261-0,6712)^2 + \cdots + (0,5629-(-0,1701)^2} = 6,105$$

$$D_{1,3} = \sqrt{(-0,3261-1,2005)^2 + \cdots + (0,5629-(-0,1701))^2} = 4,787$$

$$D_{1,4} = \sqrt{(-0,3261-(1,6240))^2 + \cdots + (0,5629-(-0,1701))^2} = 4,313$$

$$\vdots$$

$$D_{1,530} = \sqrt{(-0{,}3261-(-1{,}1286))^2+\cdots+(0{,}5629-(-1{,}3656))^2} = 3{,}112$$

All Euclidean distances that have been calculated for all test data are then sorted from smallest to largest value. The Euclidean distance is then selected by means of trial and error to determine the value of the parameter k until the smallest APER value or error rate is obtained for each value of k. The results of these trials are summarized in Table 4 below.

Table 5. Summary of APER Values for each parameter k

k value	APER value
1	0,0758
3	**0,0303**
5	0,0379
7	0,0379
9	0,0379
11	0,0379
13	0,0379
15	0,0455

In Table 4 above, it can be seen that the k value that produces the smallest APER rate is found when using the value k = 3, so that 3 observations with the smallest Euclidean distance are selected. For the first test data, if it is based on taking k=3, then the first 3 nearest neighbors for the test data have the smallest Euclidean distance, namely (1.2174); (1.4745); and (1.4748). The three smallest distances all have class 1 labels (primary hypertension), so the first test data is predicted to fall into class 1 (primary hypertension). Then the second, third and so on test data are also predicted using the same steps as for the first test data.

3.2.2 LMKNN Method

In the LMKNN method, the selection of training and test data proportions is also carried out based on the smallest APER value which is tested on all training and test data proportions with the help of Matlab 2009a. The test results are summarized in the following table:

Table 5. Summary of APER Values for each k parameter in the KNN Method

k	**Training Data Proportion : Testing Data Proportion**			
	90:10	**80:20**	**70:30**	**60:40**
1	0,3182	0,1894	0,1465	0,1472
3	0	0,0606	0,0758	0,0792
5	0	0,0530	0,0606	0,0604
7	0	0,0379	0,0505	0,0642
9	0	0,0379	0,0455	0,0604
11	0	**0,0303**	0,0556	0,0566
13	0	0,0303	0,0556	0,0604
15	0	0,0379	0,0556	0,0566

From the table above, the proportion of training and test data was chosen as 80:20 with a value of k = 11 because it produces the smallest APER value, namely 0.0303. So before standardizing the data, the data is grouped first based on the class of each data. Then the data is divided into training and test data with respective proportions of 80% and 20% in each class. Then data standardization was carried out using the z-score method.

The parameters used to standardize test data are average, variance and standard deviation. Because the proportions used in the LMKNN method are the same as KNN, the standardization results and Euclidean distance calculation results are the same as the KNN method. Then, the Euclidean distance is grouped based on existing classes and ordered from smallest to largest value.

Next, the k value is selected based on the smallest APER value which has been calculated with the help of Matlab 2009a. The k values are summarized in the following table.

Table 6. Summary of APER Values for each k parameter in the LMKNN Method

k value	APER value
1	0,1894
3	0,0606
5	0,0530
7	0,0379
9	0,0379
11	**0,0303**
13	0,0303
15	0,0379

Based on Table 6, the k value that produces the smallest APER rate is k =11. Then 11 observations with the closest Euclidean distance in each class were selected which will later be used to calculate the local average vector.

There are 11 observations for the first test data based on the smallest Euclidean distance for each class in Tables 7 and 8 below.

Class 1 Observations (Primary Hypertension)

Table 7. Observations in Class 1 (Primary Hypertension)

Observation to-	X_1	X_2	X_3	X_4	X_5	Euclidean Distances
1	0,0280	0,0145	0,4155	0,9122	0,1119	1,2174
2	0,7946	1,1397	0,0108	0,1172	0,1119	1,4745
3	0,7946	0,3529	0,0034	0,9122	0,1119	1,4748
4	0,0656	1,1397	0,0108	0,9122	0,1119	1,4967
5	0,1311	0,6829	0,4155	0,9122	0,5372	1,6367
6	1,2548	0,3529	0,1911	0,9122	0,1119	1,6801
7	0,4619	1,1397	0,4155	0,9122	0,1119	1,7439
8	0,0038	1,1397	0,9567	0,9122	0,1119	1,7675
9	0,4854	0,1245	0,0108	0,9122	2,3407	1,9681
10	0,3490	1,1397	1,4053	0,9122	0,1119	1,9794
11	0,1437	0,3529	0,1911	0,9122	2,3407	1,9851

The local average vector calculation is carried out based on each existing class. The following illustrates the calculation of the local average vector for k = 11 with the following testing data.

$X_{1,1} = -0{,}3261$; $X_{2,1} = 0{,}8664$; $X_{3,1} = 2{,}6561$; $X_{4,1} = 3{,}2677$ dan $X_{5,1} = 0{,}5629$

$$\overline{X}_1 = \frac{X_{(1,1)}+X_{(1,2)}+X_{(1,3)}+\cdots+X_{(1,k)}}{k} = \frac{X_{(1,1)}+X_{(1,2)}+X_{(1,3)}+\cdots+X_{(1,11)}}{11} = \frac{(0{,}0280)+(0{,}7946)+(0{,}7946)+\cdots+(0{,}1437)}{11} = 0{,}4102$$

$$\vdots$$

$$\overline{X}_5 = \frac{X_{(5,1)}+X_{(5,2)}+X_{(5,3)}+\cdots+X_{(5,k)}}{k} = \frac{X_{(5,1)}+X_{(5,2)}+X_{(5,3)}+\cdots+X_{(5,11)}}{11} = \frac{(0{,}1119)+(0{,}1119)+(0{,}1119)+\cdots+(2{,}3407)}{11} = 0{,}5558$$

Class 1 Local Mean Vector (Primary Hypertension):

$$\mathbf{u}_1 = \begin{bmatrix} 0{,}4102 \\ 0{,}6890 \\ 0{,}3660 \\ 0{,}8399 \\ 0{,}5558 \end{bmatrix}$$

Class 2 Observations (Secondary Hypertension)

11 observations are presented with the closest distance to each variable as follows:

Table 8. Observations in Class 2 (Secondary Hypertension)

Observation to-	X_1	X_2	X_3	X_4	X_5	Euclidean Distance
1	0,0280	0,0145	0,1911	0,9122	3,0197	2,041
2	0,0656	0,0145	0,9567	5,0737	0,5372	2,578
3	2,3306	0,1245	2,3073	0,9122	2,7491	2,902
4	4,2271	0,4744	0,9567	0,9122	2,7491	3,053
5	0,9946	0,0961	0,4155	5,0737	2,7491	3,054
6	0,6441	0,1638	4,2431	0,9122	3,7188	3,112
7	1,7289	0,0961	0,1911	5,0737	2,7491	3,137
8	4,2271	0,3529	2,3073	0,9122	2,7491	3,248
9	3,8030	3,6857	0,1911	5,0737	0,5372	3,646
10	4,2271	2,3748	0,1911	5,0737	2,7491	3,823
11	3,8030	1,5799	2,3073	5,0737	2,7491	3,939

With the same test data used to calculate the local average vector in class 1, the test data is also used to calculate the local average vector in class 2. The following illustrates an example of calculating the local average vector in class 2:

$$\overline{X}_1 = \frac{X_{(1,1)}+X_{(1,2)}+X_{(1,3)}+\cdots+X_{(1,k)}}{k} = \frac{X_{(1,1)}+X_{(1,2)}+X_{(1,3)}+\cdots+X_{(1,11)}}{11} = \frac{(0{,}0280)+(0{,}0656)+(2{,}3306)+\cdots+(3{,}8030)}{11} = 2{,}3708$$

$$\vdots$$

$$\overline{X}_5 = \frac{X_{(5,1)}+X_{(5,2)}+X_{(5,3)}+\cdots+X_{(5,k)}}{k} = \frac{X_{(5,1)}+X_{(5,2)}+X_{(5,3)}+\cdots+X_{(5,11)}}{11} = \frac{(3{,}0197)+(0{,}5372)+(2{,}7491)+\cdots+(2{,}7491)}{11} = 2{,}4597$$

So, Class 2 Local Mean Vector (Secondary Hypertension):

$$\mathbf{u_2} = \begin{bmatrix} 2{,}3708 \\ 0{,}8161 \\ 1{,}2962 \\ 3{,}1821 \\ 2{,}4597 \end{bmatrix}$$

After obtaining the local average vector value for the first test data, the Euclidean distance is then calculated based on the average vector as follows:

$$\mathbf{u}_1 = \begin{bmatrix} 0{,}4102 \\ 0{,}6890 \\ 0{,}3660 \\ 0{,}8399 \\ 0{,}5558 \end{bmatrix}; \mathbf{u_2} = \begin{bmatrix} 2{,}3708 \\ 0{,}8161 \\ 1{,}2962 \\ 3{,}1821 \\ 2{,}4597 \end{bmatrix};$$

Euclidean Distance Calculation

Euclidean Distances = $D(\mathbf{x}, \mathbf{u}) = \sqrt{\sum_{i=1}^{n}(x_i - u_i)^2}$ where x is the test data and u is the local average vector.

a. Euclidean distance in class 1 (Primary Hypertension)

$D(\mathbf{x}, \mathbf{u_1}) = \sqrt{((-0.3261) - 0{,}4102)^2 + \cdots + (0.5629 - 0{,}5558)^2}$ = 1,3384

b. Euclidean distance in class 2 (Secondary Hypertension)

$D(\mathbf{x}, \mathbf{u_2}) = \sqrt{((0.3261) - 2.3708)^2 + \cdots + (0.5629 - 2.4597)^2}$ = 4.9135

From the prediction results using Euclid distance calculations in each of the classes above, the first test data produced the smallest Euclid value, namely 1.3384 in the Euclid distance calculation in class 1 (Primary Hypertension) so that the first test data was categorized into class 1 (Primary Hypertension). Prediction of other test data is also carried out using the steps above.

3.3 Confussion Matrix

To see the accuracy of the classification results from the KNN and LMKNN methods, it is necessary to calculate the confusion matrix. The confusion matrix contains a summary of the prediction results of test data for hypertension classes. From the classification results using both methods, both KNN and LMKNN, the respective confusion matrices are obtained as follows:

3.3.1 k-Nearest Neighbor (KNN) Method

Table 9. Confusion Matrix of KNN Method

		Observed	
Predicted		Positive Class (Class 1)	Negative Class (Class 2)
	Positive Class (Class 1)	116	3
	Negative Class (Class 2)	1	12

3.3.2 LMKNN Method

Table 10. Confusion Matrix of LMKNN Method

		Observed	
Predicted		Positive Class (Class 1)	Negative Class (Class 2)
	Positive Class (Class 1)	117	0
	Negative Class (Class 2)	4	11

According to Johnson and Wichern [9], a good classification method will produce few classification errors. To determine the accuracy of classification Apparent Error Rate (APER) is used. Apparent Error Rate (APER) is a

value used to see the opportunity for errors in classification object. APER is a form-independent classification evaluation procedure population. The APER value states the proportion of samples that are misclassified. The best method is the method that has the smallest APER value so that the method has the greatest classification accuracy.

3.4 Evaluation of the Accuracy of Classification Results

Based on Table 9, the following APER and accuracy values can be obtained:

$$\boldsymbol{APER}: = \frac{number\ of\ data\ predicted\ incorrectly}{number\ of\ predictions\ made} = \frac{(3+1)}{(116+1+3+12)} = 0{,}0303$$

Accuracy = $(1 - APER)\ x\ 100\% = (1\ -\ 0{,}0303)\text{x}\ 100\% = 96{,}97\%$

Based on Table 10, the following APER and accuracy values can be obtained:

$$\boldsymbol{APER}: = \frac{number\ of\ data\ predicted\ incorrectly}{number\ of\ predictions\ made} = \frac{(0+4)}{(117+0+4+11)} = 0{,}0303$$

Accuracy = $(1 - APER)\ x\ 100\% = (1\ -\ 0{,}0303)\text{x}\ 100\% = 96{,}97\%$

4. CONCLUSION

Based on APER calculations, it shows that classification using the KNN method and the LMKNN method both provide equally good classifications in classifying the diagnosis of patients suffering from hypertension at the Merdeka Health Center, Palembang City. However, if we look at the k value obtained, with the same proportion of training and testing data, the KNN method only needs to take k = 3 to obtain the APER value and the same accuracy as the LMKNN method which takes the value k = 11. For future studies, it is necessary to consider the use of data mining methods with unbalanced data as well as newer methods to obtain better models.

ACKNOWLEDGMENTS

The author would like to thank the entire academic community of the Bengkulu University Masters in Statistics program who have helped the author a lot during the preparation of this journal.

REFERENCES

[1] N.P.Albertus, “Big Data, Data Analyst, and Improving the Competence of Librarian”. Record and Library Journal, e-ISSN 2442-5168, vol 1, no. 2, pp 1, July-December, 2015.
[2] Riduwan, Measurement Scale, Bandung, Alfabeta, 2012.
[3] A.E.Susanto and T.H.Wibowo, “Effectiveness of Giving Deep Relaxation to Reduce Pain in Hypertension Patients in Edelweis Room Down, Kardinah Tegal Hospital”. Jurnal Inovasi Penelitian, vol. 3, no. 4, pp. 1, September, 2022.
[4] A.H..Wahbeh, Q.A.Al-Radaideh, M.N.Al-Kabiand, and E.M.Al-Shawakfa, “A Comparison Study Between Data Mining Tools Over Some Classification Methods”. International Journal of Advanced Computer science and Application Special issues on Artificial Intelligence (IJACSA), vol. 35, pp 18-26, 2011.
[5] Mitani, Y, and Hamamoto, Y. “A Local Mean-Based Nonparametric Classifier”. Pattern Recognition Letter 27. Pp.1151-1159, Elsevier, 2006.
[6] Pan, Zhibbin, Wang, Yidi, and Ku, Weiping. “A New K-Harmonic Nearest Neighbor Classifier Based on The Multi-Local Means Expert Systems with Applications”. School of Electronic and Information Engineering, Xi’an Jiaotong University, 2017.
[7] J. Gou, Z.Ti, L.Du, and T.Xiong, “A Local Mean-Based k-Nearest Centroid Neighbor Classifier”. The Computer Journal, Vol.55, No.9, pp.1058-1071, 2012.
[8] E. Prasetyo, Data Mining: Concepts and Applications Using Matlab, Yogyakarta, Andi, 2012.
[9] R.A.Johnson and D.W.Wichern, Applied Multivariate Statistical Analysis, New Jersey , Pearson Prentice Hall, 2007.