Jurnal Natural
Vol. 23, (2) 2023
DOI 10.24815/jn.v23i2.29738
pISSN 1411-8513
eISSN 2541-4062
ORIGINAL RESEARCH

The classification of "Program Sembako" recipients in Payobasung West Sumatra Based on the K-nearest neighbors classifier

HAZMIRA YOZZA[1*], NINDI MAULA AZIZAH[1], LYRA YULIANTI[1], IZZATI RAHMI HG[1]

[1]Faculty of Mathematics and Natural Sciences, Andalas University, Padang, Indonesia

Abstract. The "Sembako Program" is a program carried out by the Indonesian government to improve the welfare of low-income communities. The purposes of this study are: (a) to determine the classification of households that deserve to receive basic-food assistance in Koto Panjang Payobasung, West Sumatra, using the KNN classifier and (b) to determine the optimal number of nearest neighbors used in the classification process. The measure of proximity between objects used is the Gower dissimilarity coefficient. This research used primary data consisting of 175 households collected purposively in a survey conducted on all households in Payobasung. The optimal K value is determined by implementing a 5-fold cross-validation procedure. The result showed that the best classification process is when K = 3 nearest neighbors are used since it produces the highest accuracy coefficient and Mattews correlation coefficient (MCC). Therefore, for further work, in deciding the eligibility of a household to receive the Sembako Program in Payobasung, KNN can be used by considering its 3 nearest neighbors.

Keywords: Sembako Program, KNN classifier, nearest neighbor, Gower dissimilarity coefficient

INTRODUCTION

One of the efforts made by the Indonesian government to improve the welfare of low-income people is through the provision of a Non-cash Food Assistance Program, which is transformed into assistance for the basic food program known as the Sembako Program [1]. Payobasung is one of the villages located in Payakumbuh City, West Sumatra, with a population of 2,704 people. According to the information held by the local government of Payobasung, there is a considerable number of households below the National Poverty Line who are proposed to receive the benefit of this assistance program. However, only a small portion of households meet the government's criteria and are declared eligible to receive the Sembako Program.

The COVID-19 pandemic that has hit Indonesia since 2020 has had an impact on the economic sector, including in Payobasung. This increased the number of poor people who are predicted to be eligible to receive the Sembako Program. To meet the government's goal to fulfill the need of poor people for nutritious food, it is necessary to know the eligibility of a household to be categorized as a recipient of the Sembako Program in Payobasung. The first step is to determine the criteria that make a household eligible to be a recipient of the Sembako Program. Based on these criteria, households in Payobasung are classified into the group of households that are eligible to receive food assistance and the group of households that are not eligible. Statistically, it can be done through various classification methods.

Classification is the process of finding a model or rule that describes and distinguishes data classes or concepts. Classification is used to predict categorical class labels and classifies newly or unlabelled available data by assigning them to the class of the most similar characteristics. There are several classification techniques: ID3 algorithm, C4.5 algorithm, support vector machine, Naïve Bayes, K-nearest neighbor, and many more [2], [3]. Generally, the accuracy of each algorithm depends on the specific dataset and problem being addressed [4],[5],[6]. According to [7] K-nearest neighbor (KNN classifier) is easier to use than other methods. ID3 and C4.5 are classification methods that are based on a decision tree algorithm. The ID3 can only work

*Corresponding Author:
hazmirayozza@sci.unand.ac.id

Received: December 2022| Revised: May 2023 | Accepted: May 2023

with categorical data, so it is not suitable for handling numerical data [8],[9]. On the other hand, the KNN classifier works well with all types of data, unlike Naïve Bayes, support vector machine, and logistic regression, which require certain assumptions about underlying data distribution[10]. KNN is a non-parametric method that does not assume data distribution. It makes KNN more flexible and adaptable to different types of data.

The KNN classifier method is a distance-based classification process in determining the proximity between data that will be the nearest neighbors. KNN classifier classifies objects based on learning data of which the closest distance to the object [11]. The KNN classifier algorithm uses information from variable values from the training data set. In the process of classifying new data, the proximity of the new data to all training data is calculated. The class for the new data is determined based on its K nearest neighbors' majority class [12].

The performance of the KNN classifier depends on the dissimilarity measure used to describe the distance between objects. For that reason, choosing the right similarity/dissimilarity is important. It depends on the data type. When all variables are numeric, Euclidean distance is appropriate [13]. When all variables are categorical, the commonly used similarity measure is the simple matching coefficient [14]. For mixed-type data (a mixture of numeric and categorical data), the dissimilarity measure that can be used is the Gower dissimilarity coefficient [15],[16].

In the KNN classifier, the K parameter defines the number of nearest neighbors. It also plays an important role in determining the class prediction of the new data and subsequently affects the accuracy of the whole prediction [12],[17]. The large dataset does not always require a large K value; conversely, small data does not always require a small K. In cases with very unbalanced classes, it is recommended to use a small K value. In some studies, it is recommended to determine the optimal K value by repeatedly running the KNN algorithm with different K values and choosing the one that provides the highest accuracy [17]. Other studies use different numbers of nearest neighbors to different test samples [18].

Several coefficients can be used to measure the accuracy of a classification method. The commonly used measure is the accuracy coefficient, calculated from a confusion matrix [19]. The confusion matrix is a type of contingency table to evaluate or visualize the behavior of models in supervised classification contexts [20]. A confusion matrix of size n x n associated with a classifier shows the predicted and actual classification, where n is the number of different classes [21]. The rows of the confusion matrix represent the actual class of the instances and the columns represent their predicted class [20]. The accuracy coefficient measures the proportion of observations that are correctly classified by a classification method. The higher the accuracy coefficient, the better the classifier results. The other coefficient recommended if positive and negative cases are of equal importance is the Matthews correlation coefficient (MCC). The higher the MCC, the better the classifier results [22],[23].

This study aims to: (a) determine the eligibility classification of households in Payobasung, West Sumatra as recipients of the Sembako Program using the KNN classifier, and (b) determine the optimal K value for the KNN classifier. A similar study has been conducted by [7], but this study used a different distance measure that is more appropriate to be used for mixture variables, i.e. Gower coefficient.

METHODOLOGY

Population and Sample

The population of this study is 838 households in Payobasung at Payakumbuh City, West Sumatra. The sample consists of 175 households selected purposively from the population.

Data Collection

This study used the primary data collected in a survey conducted on all households in the sample. The survey is carried out by interviewing the head of the household.

Variables

The variables involved in this study are:

1. Household dependency (X_1), that is the number of household dependencies. This is a numeric variable.
2. Household income (X_2); household income is ordinal and classified into four categories, i.e:
   - Less than 1.000.000 rupiahs
   - 1.000.000 – 3.000.000 rupiahs
   - 3.000.000 – 5.000.000 rupiahs
   - More than 5.000.000 rupiahs
3. The occupation of the head of the household (X_3); the household head occupation is nominal and classified into 5 categories, namely:
   - Labor,
   - Self-employee,
   - Private employee,
   - Civil servant,
   - Others

4. Home ownership status (X_4)
   This variable is nominal and classified into 2 categories, i.e:
   - Sole ownership
   - Tenancy

Data Analysis
Data analysis is carried out using three main stages.

Stage 1: Classifying the new observations
In this research, households in Koto Panjang Payobasung were grouped into two classes, namely, eligible and non-eligible recipients of the Sembako Program. Here, the KNN classifier was used to identify which category a new observation (household) belongs to. The classification process was carried out by splitting data into a training dataset and a testing dataset, with 80% data in the training dataset and the remaining 20% in the testing dataset.

a. Set the number of nearest neighbors (K value)

b. For every new observation (observation in the testing dataset), do:
   i. Calculate the Gower coefficient between new data (testing data) to each training data
   The Gower coefficient is a dissimilarity used for a mixture of continuous and categorical data [15], [16]. Denote $\boldsymbol{x_c}$ as the-c variable and d_{ijc} is a dissimilarity measure between the i-th and j-th objects by the c-th variable (c = 1, ..., m),
   - If $\boldsymbol{x_c}$ is nominal, the dissimilarity between two objects is expressed as

$$\boldsymbol{d_{ijc}} = \begin{cases} \mathbf{0}, & \boldsymbol{x_{ic} = x_{jc}} \\ \mathbf{1}, & \boldsymbol{x_{ic} \neq x_{jc}} \end{cases} \quad (1)$$

   - If $\boldsymbol{x_c}$ is ratio/numeric, the dissimilarity between two objects is

$$\boldsymbol{d_{ijc}} = \frac{|\boldsymbol{x_{ic} - x_{jc}}|}{\mathbf{max}(\boldsymbol{x_c}) - \mathbf{min}\,(\boldsymbol{x_c})} \quad (2)$$

   where $\mathbf{max}(\boldsymbol{x_c})$ and $\mathbf{min}(\boldsymbol{x_c})$ are the maximum and minimum values of $\boldsymbol{x_c}$, respectively.
   - If $\boldsymbol{x_c}$ is an ordinal variable, then all the categories are transformed using the formula

$$\boldsymbol{x_{ic}} = \frac{\boldsymbol{r_{ic} - 1}}{\boldsymbol{R_c - 1}} \quad (3)$$

   where r_{ic} : the rank number of the i-th ordinal category (r = 1, ..., Rc)
   Rc: the maximal rank number of x_c.
   After this transformation, the dissimilarity is calculated using the formula for numeric variables (equation 2).

   Finally, the Gower coefficient between two objects is calculated by

$$\boldsymbol{d_G(X_i, X_j)} = \frac{\sum_{c=1}^{C}(\boldsymbol{w_{ijc} \times d_{ijc}})}{\sum_{c=1}^{C} \boldsymbol{w_{ijc}}} \quad (4)$$

   where and w_{ijc} takes the value zero, if either the i-th or the j-th object by the c-th variable is missing; otherwise, it takes the value one [16].
   ii. Determine its K nearest neighbors.
   iii. Identify the class of each nearest neighbors
   iv. Grouped new observation based on the majority class of its nearest neighbors.

Stage 2: Determining the Accuracy of the Classification Result
The accuracy of the classification result was measured as follows:

a. Construct the confusion matrix
   The confusion matrix for this research is shown in Table 1.

Table 1. Confusion matrix

Predicted class	Actual class	
	Eligible (Positive)	***Non-eligible (Negative)***
Eligible (Positive)	TP	FP
Non-eligible (Negative)	FN	TN

b. The accuracy coefficient is defined as

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \quad (5)$$

c. The Matthews' correlation coefficient (MCC) is defined as [20]

$$\boldsymbol{MCC} = \frac{(\boldsymbol{TP.TN}) - (\boldsymbol{FP.FN})}{\sqrt{(\boldsymbol{TP + FP})(\boldsymbol{TP + FN})(\boldsymbol{TP + FP})(\boldsymbol{TP + FN})}} \quad (6)$$

Stage 3: Selecting the Optimal K value
The number of the nearest neighbor (K value) has an important role in the KNN classifier. To select the optimal value of K, a 5-fold cross validation method is implemented by following steps.
a. Randomly split data into 5 sub-datasets.
b. Iteratively, repeat stage 1 by using a sub-dataset as a validation dataset and the remaining data as a training dataset.
c. Construct a confusion matrix for all data and calculate the accuracy coefficient and MCC
d. Repeat stage a-c for K= 1,3, 5,…, 29.
e. The optimal K value is the K value that provides the highest accuracy coefficient and the highest Mattews' correlation coefficient (MCC).

The experiment stages are shown in Figure 1.

RESULTS AND DISCUSSION

In this research, the KNN classifier is used to classify households in Payobasung West Sumatra into two classes according to their eligibility as the recipient of the Sembako

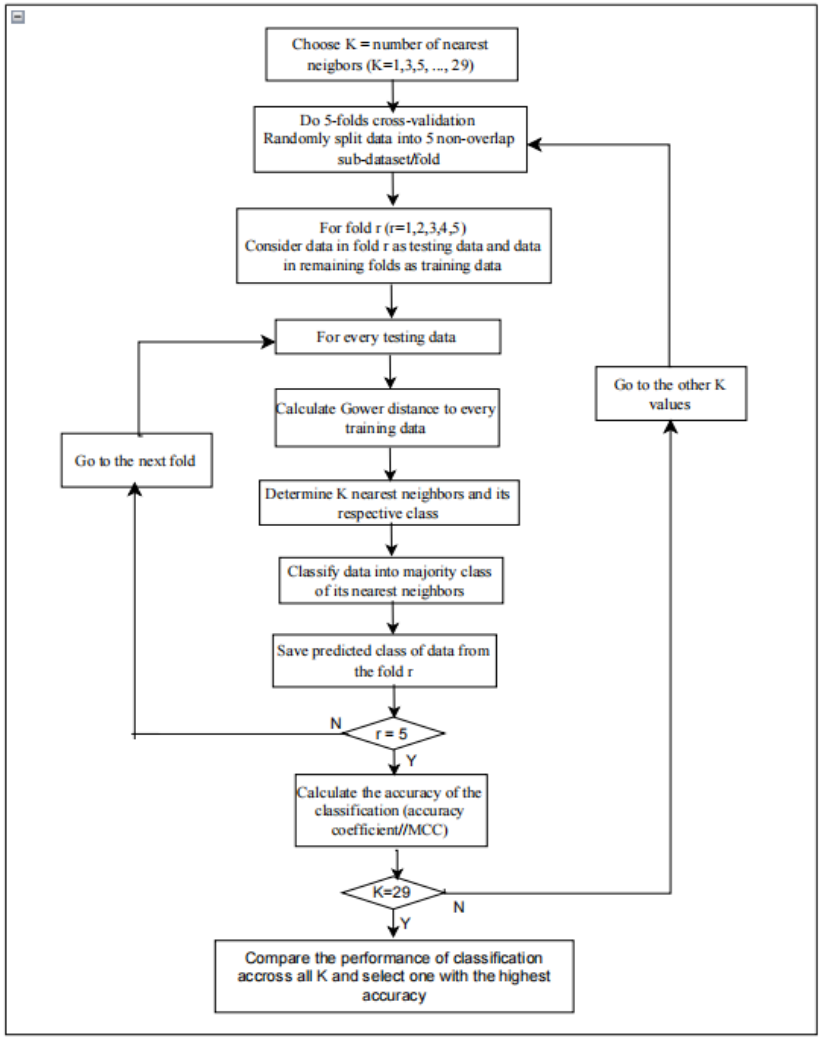

Figure 1. The experiment stages

Program provided by the Indonesian government. The sample consists of 175 households. The data is partitioned into the training and the testing data consisting of 140 data (80% data) and 35 data (20% data) respectively. Variables used are household dependency, household income, occupation of household head, and home ownership status. First, we will illustrate the classifying process of households in Payobasung using the KNN algorithm with K=3 nearest neighbors. The same algorithm is applied with different K values and based on the resulting accuracies we determined the optimal K value which is the K value that provides the highest accuracy.

The KNN classifier for K=3

The first stage to classify new data is to calculate the Gower distances between each new data and all training data. The following illustrates the calculation of the Gower coefficient between new data, denoted by Test-1, and five data in the training datasets, denoted by Train1-Train5. Table 2 shows data for each observation.

The Gower distance between two objects is determined by calculating the dissimilarity between the two objects, separately based on each variable.

- Household dependency
  Household dependency is a ratio variable with maximum and minimum values are 1 and 7, respectively. The dissimilarity between Test1 and Train1 based on this variable is:

$$d_{(New1,Train1)1} = \frac{|x_{i1} - x_{j1}|}{\max(x_1) - \min(x_1)} = \frac{|5-4|}{7-1} = \frac{1}{6}.$$

Table 2. Illustrated data

Data	Obs.	Household Dependency	Household Income	Homeowners hip status	Occupation	Eligible/ non-eligible
New data	Test1	5	<1.000.000	Sole	Labor	-
Training Data	Train1	4	<1.000.000	Sole	Labor	Eligible
	Train2	5	> 5.000.000	Sole	Civil servant	Non- Eligible
	Train3	1	<1.000.000	Sole	Labor	Non-Eligible
	Train4	2	<1.000.000	Sole	Labor	Eligible
	Train5	7	1.000.001-3.000.000	Sole	Self -Employment	Non-Eligible

Table 3. Household income transformation

r	Household income	Transformed value
1	<1.000.000	$x_{12} = \frac{1-1}{4-1} = 0$
2	1.000.001-3.000.000	$x_{12} = \frac{2-1}{4-1} = \frac{1}{3}$
3	3.000.001-5.000.000	$x_{13} = \frac{3-1}{4-1} = \frac{2}{3}$
4	> 5.000.000	$x_{14} = \frac{4-1}{4-1} = 1$

- Household Income
  Since household income is ordinal, all categories were transformed using (3). The result is shown in Table 3.

  Based on this rule, the transform value of household income for Test1 and Train 1 are 0 and 0, respectively. The dissimilarity between Test1 and Train1 is:

$$d_{(New1,Train1)2} = \frac{|x_{i2} - x_{j2}|}{max(x_2) - \min(x_2)} = \frac{|0-0|}{1-0} = 0.$$

- Household head occupation

  This is a nominal variable. Since Test1 and Train1 have the same household occupation, the dissimilarity between these objects is $d_{(New1,Train1)3} = 0$

- Household head occupation

  The occupation of the head of Test1 and Train1 are the same, then the dissimilarity between these objects is $d_{(New1,Train1)4} = 0$

Finally, The Gower distance between Test1 and Train1 is:

$$\boldsymbol{d_G(New1, Train1)} = \frac{\sum_{c=1}^{4}(\boldsymbol{w_{ijc}} \times \boldsymbol{d_{ijc}})}{\sum_{c=1}^{4} \boldsymbol{w_{ijC}}}$$

$$= \frac{\left(\mathbf{1} \times \frac{\mathbf{1}}{\mathbf{6}}\right) + (\mathbf{1} \times \mathbf{0}) + (\mathbf{1} \times \mathbf{0}) + (\mathbf{1} \times \mathbf{0})}{(\mathbf{1} + \mathbf{1} + \mathbf{1} + \mathbf{1})}$$

$$= \mathbf{0.04167}$$

The value w_{ijc} =1 for all variables because the values of x_{ic} and x_{jc} are non-missing. The Gower coefficient between Test1 and Train1-Train5 were shown in Table 4.

The Gower coefficient between Test1 and the rest of the training data (Train6-Train140) were calculated. It was found that the three closest neighbors for Test1 are Train1, Train3, and Train4. From Table 2, it has already known that Train1 and Train4 belong to the eligible class and Train3 belongs to a non-eligible class. Therefore, Test1 is assigned to the eligible class since the eligible class is the majority class of its nearest neighbors. Furthermore, the same procedure is followed to assign all data in the testing data to eligible or non-eligible classes.

The next stage is to determine the accuracy of classification. In the next step, the accuracy value will be used to determine the optimal K-value. For this purpose, a 5-fold cross-validation procedure was employed and a confusion matrix was constructed according to the actual and predicted class of all data. The confusion matrix is presented in Table 5.

The accuracy coefficient for the KNN classifier with K=3 nearest neighbors is

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{95 + 75}{95 + 75 + 0 + 5}$$

$$= 0.9714$$

The accuracy of classification is 0,9714, which means that 97.14% of the testing data are correctly classified by KNN classifier with K=3 nearest neighbors.

The Matthews Correlation Coefficient for KNN classifier with K=3 nearest neighbors is

$$MCC = \frac{(TP.TN) - (FP.FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$= \frac{(95x75) - (0x5)}{\sqrt{(95 + 0)(95 + 5)(75 + 0)(75 + 5)}}$$

$$= 0.9437$$

Table 4. Dissimilarity measure and Gower coefficient between Test1 and Train1-Train5

Observation	Dissimilarity based on variable				Gower coefficient
	Dependency	Income	Occupation	Homeownership status	
Train1	0.1667	0	0	0	0.0417
Train2	0	1	1	0	0.5000
Train3	0.6667	0	0	0	0.1667
Train4	0.5000	0	0	0	0.1250
Train5	0.3333	0.3333	1	0	0.4167

Table 5. The Confusion matrix for K=3

Predicted class	Actual class	
	Eligible	***Non-eligible***
Eligible	TP=95	FP=0
Non-eligible	FN=5	TN=75

The Optimal K

The optimal K value is obtained by performing the above procedure for K=1,3,5,7,…. 29 and calculating the accuracy for each K. The optimal value of K is the value of K that gives the highest accuracy. The accuracy of a classification and The Mattews correlation coefficients (MCC) for KNN with all K values are shown in Table 6 and graphically presented in Figure 2.

The classification performed with K=1 nearest neighbor provides an accuracy coefficient of 96.57%, indicating that 96.57% of observations are correctly classified by the KNN classifier with K=1. By using K=3, more superior classifier results are produced with a higher accuracy coefficient, 97.14%. Furthermore, when K is between 5 and 19, the accuracy coefficients fluctuate, but the coefficients gained never surpass the accuracy coefficient achieved when K=3. The accuracy coefficients decline as the K values increase, until at K=29, only 90% of the data are classified correctly. Based on this pattern, it can be expected that the accuracy value will be lower for a higher K value. Thus, it can be concluded that the highest accuracy coefficient value will be obtained if K = 3 is used. The pattern as described above is seen more clearly on the MCC chart. Therefore, in classifying households in Payobasung, we can choose K=3 as the optimal K value. The selection of a small K value is also intended to reduce the noise in the classification process.

Many countries have implemented food assistance programs for their citizens who are food insecure. According to various studies, these programs are closely related to insecurity, poverty, and health. Along with other assistance programs, these programs significantly promote food security and good health. These programs effectively reduce poverty by providing benefits to households for grocery purchases and allowing them to spend more on their other basic needs, such as education, health care, and electricity [24]. According to another research, the fundamental purpose of food assistance programs is to provide lower-income families with access to a healthy meal, which improves the nutritional well-being of the individuals they serve. Furthermore, studies have linked the Supplemental Nutrition Assistance Program in the USA to better health outcomes and lower health-care expenses [25]. Food assistance programs are implemented in various forms in Indonesia, including the Raskin, cash transfer, and Sembako programs.

The main challenge in implementing a food assistance program is accurately identifying the

Table 6. The accuracy coefficient and Mattews correlation coefficients (MCC) for K=1,3,5,…,29

k	*TP*	*FP*	*TN*	*FN*	*Accuracy coefficient*	*MCC*
1	96	2	73	4	0.9657	0.9305
3	95	0	75	5	0.9714	0.9437
5	93	0	75	7	0.9600	0.9223
7	93	2	73	7	0.9486	0.8974
9	93	1	74	7	0.9543	0.9098
11	92	3	72	8	0.9371	0.8742
13	93	1	74	7	0.9543	0.9098
15	93	0	75	7	0.9600	0.9223
17	93	0	75	7	0.9600	0.9223
19	93	0	75	7	0.9600	0.9223
21	93	4	71	7	0.9371	0.8728
23	93	8	67	7	0.9143	0.8248
25	93	9	66	7	0.9086	0.8129
27	93	11	64	7	0.8971	0.7895
29	93	10	65	7	0.9029	0.8012

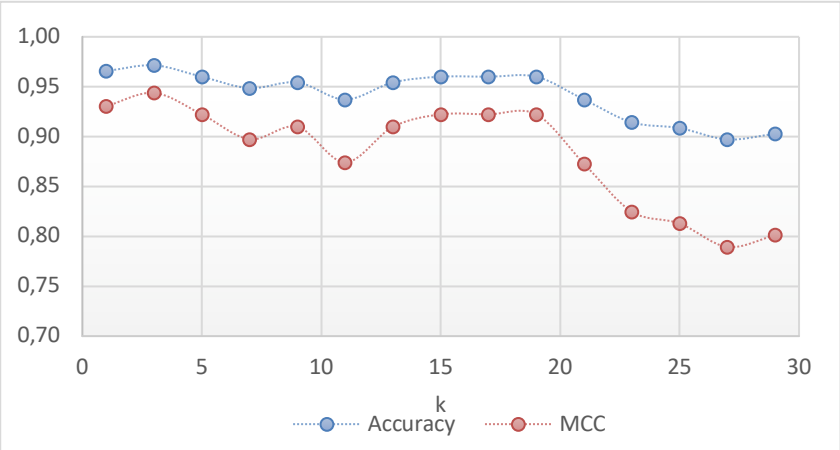

Figure 2. The accuracy coefficient and the Mattews correlation coefficient

eligible recipient of this program. A study conducted by [26] used 5 criteria, i,e. employment, income, dependencies, and housing condition, to select the recipient of the non-cash food assistance program. Similar criteria are used in [7]. As in our research, the latter study attempts to discover a procedure to predict a household's eligibility to participate in the program by adopting the KNN classifier. Unlike our current research, the optimal number of nearest neighbors (K) in this study was derived from the values 15, 30, 45, 60, and 75, and the best accuracy values were found at K = 15 and K = 30. For K = 45, 60, 75, the value of K decreased significantly.

By this information, the pattern of accuracy in this study is expected to resemble the pattern obtained in our research. Unfortunately, researchers did not perform a KNN classifier for other K values in the range 1–30; therefore, the pattern of accuracy coefficients for K values in that interval was unrevealed. Considering the fact that the values K = 15 and K = 30 provide the same accuracy coefficient values, the accuracy coefficients within the interval may fluctuate slightly around that value, allowing the researcher to select a much smaller optimal K.

Finally, this research finding can be used for further work in deciding the eligibility of a household to receive the Sembako Program in Payobasung. The eligible recipient of the Sembako Program can be identified using KNN classifier with K = 3 based on the criteria: the number of household dependencies, household income, the occupation of household head and home-ownership status.

CONCLUSION

In this research, the KNN classifier is used to predict the eligibility of households in Payobasung to receive basic food assistance through the Sembako Program provided by the Indonesian government. Variables used as the basis of classifying are household dependency, household income, occupation of household head, and home ownership status. Since K=3 provided the highest accuracy, whether a household in Payobasung is eligible or not eligible to receive Sembako Program can be predicted from the majority class of its 3 nearest neighbors. The nearest neighbors are determined using the Gower distance which is calculated based on household dependency, household income, occupation of household head, and home-ownership status.

REFERENCES

[1] Kemenkeu RI Kanwil DJPb Kalimantan Timur, 2020. *Indeks Manfaat Program Sembako* [in Bahasa] [Online]. Available: http://djpb.kemenkeu.go.id/indeks-manfaat-programsembako. [Accessed 26 September 2021].

[2] Gorade, S.M.; Deo, A.; Purohit, P. 2017. A study of some data mining classification techniques. *International Research J. of Engineering and Technology* **4**(4) 3112-3115.

[3] Nikam, S.S. 2017. A comparative study of classification techniques in data mining algorithms, *Int. J. Mod. Trends Eng. Res.* **4**(7) 58-63. DOI: 10.21884/ijmter.2017.4211.vxayk.

[4] Gunawan, G.; Hanes, H.; Catherine, C. 2021. C4.5, k-nearest neighbor, naive Bayes, and random forest algorithm

comparison to predict students; on Time graduation. *IJAIDM*. **4**(2) 62-71. DOI: 10.24014/ijaidm.v4i2.10833

[5] Sheth, V.; Tripathi, U.; Sharma, A. 2022. A comparative analysis of machine learning algorithm for classification purposes. *Procedia Computer Science* **215** 422-431. DOI: 10.1016/j.procs.2022.12.044

[6] Ashari, A.; Paryudi, I.; Tjoa, A.M. 2013. Performance comparison between naive Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. *International Journal of advanced computer science and Application* **4**(11) 33-39. DOI. 10.14569/IJACSA.2013.041105

[7] Hasanah R.L.; Hasan, M.; Pangesti, W. E.; Wati, F.F.; Gata, G. 2019. Klasifikasi penerima dana bantuan desa menggunakan metode KNN (K-Nearest Neighbor) [in Bahasa]. *J. Techno Nusa Mandiri* **16**(1) 1-6. DOI: 10.33480/techno.v16i1.25.

[8] Patil,S; Agrawal, M; Baviskar, V.R. 2015. Efficient Processing of Decision Tree Using ID3 and Improved C4.5 Algorithm. *International Journal of Computer Science and Information Technology* **6**(2) 1956-1961. DOI: 10.4010/2016.545

[9] Idris, N.F and Ismail, M.A. 2021. Breast cancer disease classification using Fuzzy ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition. *PeerJ Computer Science* **7**:e427 DOI: 10.7717/peerj-cs.427.

[10] Vaishali H. Kamble, H., Dale, M.P. 2022. Machine learning approach for longitudinal face recognition of children. In Sarangi, P.P (ed). *Machine Learning for Biometrics* 105-128. (London: Elsevier, Inc.)

[11] Lubis, A. R.; Lubis,M.; Khowarizmi, A. 2020. Optimization of distance formula in the k-nearest neighbor method. *Bull. Electr. Eng. Informatics* **9**(1) 326–338. DOI: 10.11591/eei.v9i1.1464.

[12] Indriyanto, J. 2021. *Algoritma K-Nearest Neighbor untuk prediksi nasabah asuransi* [in Bahasa], (Jakarta: Penerbit NEM)

[13] Kataria, A.; Singh, M. 2013. A Review of Data Classification Using K-Nearest Neighbour Algorithm. *Int. J. Emerg. Technol. Adv. Eng.* **3**(6) 354–360.

[14] Chen,L.; Guo, G. 2015. Nearest neighbor classification of categorical data by attributes weighting. *Expert Syst. Appl.* **42**(6) 3142–3149. DOI: 10.1016/j.eswa.2014.12.002.

[15] Tuerhong, G.; Kim, S.B. 2014. Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert Syst. Appl.* **41**(4), 1701–1707. DOI: 10.1016/j.eswa.2013.08.068.

[16] Sulc, Z.; Procházka, J.; Matějkaz, M. 2016. Modifications of the Gower Similarity Coefficient. The 19th Appl. Math. Stat. Econ. ; Slovakia: Matej Bel University [Online]. Available: https://www.researchgate.net/publication/313387106.

[13] Paryudi, I. 2019. What affects K value selection In K-Nearest Neighbor? *Int. J. Sci. Technol. Res.* **8**(7) 86-92.

[14] Zhang, S.; Li, X.; Zong, M.; Zhu, X.; Wang, R. 2018. Efficient kNN classification with different numbers of nearest neighbors. *IEEE Trans. Neural Networks Learn. Syst.* **29**(5) 1774–1785. DOI: 10.1109/TNNLS.2017.2673241.

[19] Kulkarni, A.; Chong, D.; Batarseh, F.A. 2020. *Foundations of data imbalance and solutions for a data democracy*. (Amsterdam: Elsevier Inc.).

[20] Caelen, O. 2017. A Bayesian interpretation of the confusion matrix," *Ann. Math. Artif. Intell*. **81**(3) 429–450. DOI: 10.1007/s10472-017-9564-8.

[21] Visa, S.; Ramsay, B.; Ralescu , A.; Van der Knaap, E. 2011. Confusion matrix-based feature selection. The 22nd Midwest Artificial Intelligence and Cognitive Science Conference 2011; Cincinnati, USA; University of Cincinnati.

[22] Chicco, D.; Tötsch, N.; Jurman, G. 2021. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min.* 14(13) 1-22. DOI: 10.1186/s13040-021-00244-z.

[23] Ballabio, D.; Grisoni, F.; Todeschini, R. 2018. Multivariate comparison of classification performance. *Chemometrics and Intelligent Laboratory Systems*, **174** 33-44. DOI: 10.1016/j.chemolab.2017.12.004

[24] Keith-Jenning, B.; Llobrera, J.; Dean, S. 2019. Links of the Supplemental Nutrition Assistance Program With Food Insecurity, Poverty, and Health: Evidence and Potential. *Am.J.Public Health* **109**(12) 1636-1640. DOI: 10.2105/AJPH.2019.305325

[25] Institute of Medicine and National Research Council. 1999. *Evaluating Food Assistance Programs in an Era of Welfare Reform Summary of a Workshop*. (Washington : National Academies Press).

[26] Siregar, V.M.M.; Irmayanti; Julyanti, E.; Harahap, N.A.; Jannah, M.; Sagala, E.; Siagian, N. F.; Saediman, H.; Achmad, A. D.; Arief, A.S. 2022. Decision support system for selection of food aid recipients using SAW method. AIP Proceeding of 1st International Conference on Technology, Informatics, and Engineering; Indonesia: University of Muhammadiyah Malang; p. 030019. DOI: 10.1063/5.0094385