Available online at http://scik.org

J. Math. Comput. Sci. 2022, 12:173

https://doi.org/10.28919/jmcs/7478

ISSN: 1927-5307

ANALYSIS CLASSIFICATION OF HOUSEHOLDS WHO RECEIVED "RASKIN" IN SEMARANG CITY USING FUZZY K-NEAREST NEIGHBOR (FKNN) AND SUPPORT VECTOR MACHINE (SVM)

DWI ISPRIYANTI[1,*], ALAN PRAHUTAMA[1], MUSTAFID[1], SUGITO[1], RETNO DIAN IKA WATI[2]

[1]Statistics Department, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

[2]BPS-Statistics Agency Center of Semarang city, Semarang, Indonesia

Abstract: Data mining or Big Data is a very important part of going to the industrial revolution era 4. Data mining is inseparable from statistical analysis for classification methods. Data mining is data with a very large size. Two of the methods in data mining is classification using Fuzzy K Nearest Neighbor (FKNN) and Support Vector Machine (SVM). The concept of FKNN is based on fuzzy members, while the SVM method is based on a hyperplane. In this study, the classification of poor rice receipt in the city of Semarang used the FKNN and SVM methods. These methods applied to classification the household wom receipt "raskin" in Semarang city, Indonesia. "raskin" is one of Indonesia government program to assist the households who categorized in the poor households. We used some variables independent such as the characteristic of house and the criteria of head households. The data collected from SUSENAS-social economic survey 2016 in Semarang city with 930 households. From the results of the analysis, it was found that the characteristics of residential houses more influenced the factors of "raskin" revenue compared to the characteristics of the head of the household. The SVM method produces better accuracy than FKNN. The best accuracy value reaches 90% with the radial base function kernel function.

Keywords: Classification method; FKNN; SVM; poor rice receipt.

*Corresponding author

E-mail address: ispriyanti.dwi@gmail.com

Received May 6, 2022

1. INTRODUCTION

One of the goals of Sustainable Development Goals (SDGs) is to reduce poverty, and one of the other goals is not to starve. Therefore, the Indonesian government, in implementing the SDG's value, is contained in the Medium Term Development Plan and Long-Term Development Plan. One of the Medium Term Development Plan programs is reducing poverty with one of the programs being ‘raskin” for groups of households. “Raskin” is the assistance from the government to help the poor households [1]. To implement this program so that it is right on target, it is necessary to study what factors determine the classification of “raskin” receipts, which often refer to the income of poor households, which is classified as income per month below the poverty line. Also, the receipt of “raskin” depends on the quota of the amount of “raskin” received from a village, so the determination of who has the right to get “raskin” is a problem for the government.

Semarang City is the provincial capital of Central Java and is one of the economic centers in Central Java. The strategic location in the northern position of Java makes Semarang a strategic city for development. Although it is strategic at the economic level, it does not guarantee that all residents in the city of Semarang are free from poverty [2].

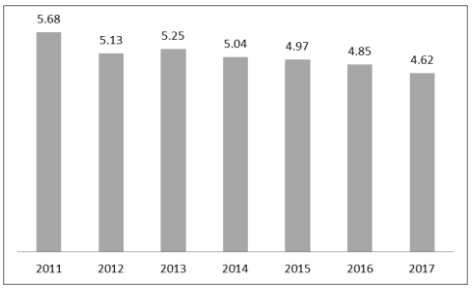

Figure 1. Percentage of poverty in Semarang city in 2011-2017

Figure 1 shows that a percentage graph of poverty in the city of Semarang from 2011-2017. Based on the graph, it can be seen that the percentage of engineering in the city of Semarang has decreased from year to year. However, in 2013 it increased by 0.12% from the previous year so that it can be seen in the graph that the reduction in the percentage of poverty from year to year is quite small.

In this era of industrial 4.0 revolution, where the Internet of Think (IoT) became a trigger for technological progress, the role of Big Data or data mining was very supportive in the current era. Data mining or Big Data is data with observations and a large number of variables. One method in data mining that is an advantage is the classification method. In data, mining classification is included in supervised learning, while clustering is included in unsupervised learning. The classification method in data mining based on supervised learning includes K-Nearest Neighbor (KNN), Fuzzy-K Nearest Neighbor (FKNN), Fuzzy-K Nearest Neighbor in every Class (FKNNC), Support Vector Machine (SVM) [3]. FKNN has an advantage compared to KNN. Namely, the selection of class members is based on the calculation of fuzzy concepts so that accuracy is higher compared to KNN. SVM is a classification method based on the hyperplane line. Some of researches regarding classification methods are classification poverty in Eastern Samar province used KNN, decision tree, and naïve Bayes method. It figured out that classification poverty can be captured by those methods [4]. In health field, the comparison of machine learning methods to predict readmission of diabetic patient. The results can be used to the global economic recovery in general and the reduction of medical equipment supply for the care and treatment of diabetics [5]. SVM also FKNN are used to analyze the Synthetic Aperture Radar (SAR) data. SAR data are of high interest for different applications in remote sensing specially land cover classification [6]. Bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor has been done. It can be used as a new candidate of powerful early warning systems for bankruptcy prediction with excellent performance [7]. In poverty, classification poverty of households has been done by [8] used Neural Network method. It resulted that the method can be good classifier to identify the poor households.

In this study, the classification of “raskin” receipts in the city of Semarang will be carried out based on the factors that influence the receipt of “raskin” used in this study, including X_1: Gender of the head of the households; X_2: age of hoseholds head; X_3: the number of households members; X_4: Highest diploma head of households; X_5: Does the head of the household work ?; X_6: business field; X_7: employment status; X_8: home ownership status; X_9: roof building material; X_{10}: Main wall material of the house; $X_{11:}$ Main House Floor Material; X_{12}: Use of

defecation facilities; X_{13}: Stool final disposal site; X_{14}: Source of Drinking Water; X_{15}: The main cooking fuel.

2. MATERIAL AND METHOD

2.1. Fuzzy K-Nearest Neighbor (FKNN)

Fuzzy K-Nearest Neighbor (FK-NN) method was introduced by Keller et al (1985) by developing K-NN which combined with fuzzy theory in conveying the definition of class labeling on predicted test data. As with fuzzy theory, a data has a membership value in each class which means that a data can be owned by a different class with the value of the degree of membership in the interval [0,1] [9].The formula used is:

$$u_{(x,c_i)} = \frac{\sum_{k=1}^{K} u_{(x_k,c_i)} d_{(x,x_k)}^{\frac{-2}{(m-1)}}}{\sum_{k=1}^{K} d_{(x,x_k)}^{\frac{-2}{(m-1)}}} \quad (1)$$

with $u_{(x,c_i)}$ is membership of the data x to ith-class $c_{i;}$ K is the number of nearest neighbor that used; $u_{(x_k,c_i)}$ is memberships value of neighbor data in K neighbor in class of c_i , is 1 if testing data x_k its class of c_i or 0 if not its class of c_i; $d_{(x,x_k)}$ is distance from data of x to data of x_k in K nearest neighbor; m is weight exponent where m > 1

In the FK-NN method, the calculation of the distance between the two data is adjusted to the data type, where each data type has its formula (Prasetyo, 2012). Calculation of distance to be used in this study is distance calculation using Euclidean distance where each variable is nominal. So before entering the Euclidean calculation, the data is calculated first with a formula for nominal data.

$$d_{ij} = \begin{cases} 0, & if\ x_i = x_j \\ 1, & if\ x_i \neq x_j \end{cases} \quad (2)$$

So that the Euclidean formula used is as follows:

$$d_{(x_i,x_j)} = \sqrt{d_{ij}^2} \quad (3)$$

Where d_{ij} is the distance from testing data (x_i) to training data (x_j) with the type of data is nominal, and $d_{(x_i,x_j)}$ is Euclidean distance. Even though FK-NN uses membership values to declare data membership in each class, to provide a final output, FKNN must still give the final output of the predicted results. For this purpose, the FK-NN chose the class with the largest membership value in the data.

2.2. Support Vector Machine

Suppose that given a set $\mathbf{X}=\{\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_n}\}$, with $\mathbf{x}_i \in \Re^p$, it is known that X has a certain pattern that is if $\mathbf{x}_i$ is included in a class then labeled $y_i=+1$, if not labeled $y_i=-1$ for that label each denoted $y_i \in \{+1,-1\}$ so that the data is in the form of pairs $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ where $i=1,2,\ldots,n$ which n is a lot of data. It is assumed that both class -1 and +1 can be completely separated by the dimensionless separator function *p*, which is defined $(\mathbf{w}^{\mathrm{T}}.\mathbf{x})+b=0$ where *w* and *b* are the model parameters [10].

To get the best separator function is to find a separating function located in the middle between two class dividing fields and to get the best separator function, it is equal to maximizing the margin or distance between two sets of objects from different classes. Furthermore, it is formulated into the quadratic programming (QP) equation, by minimizing the inverse equation $\frac{1}{2}\|\mathbf{w}\|^2$, where $\|\mathbf{w}\|^2=\mathbf{w}^{\mathbf{T}}\mathbf{w}$ with the condition $y_i[(\mathbf{w}\cdot\mathbf{x}_i)+b]\geq 1, i=1,...,l$ this Optimization requirements can be solved by the Lagrange Multiplier function [3]:

$$L(\mathbf{w},b,\alpha)=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{i=1}^{l}\alpha_i\{[(\mathbf{x}_i\cdot\mathbf{w})+b]y_i-1\} \quad (4)$$

So that the optimal condition of the LaGrange multiplier function is

$$\frac{\partial L}{\partial b}=0 \quad \Rightarrow \quad \sum_{i=1}^{l}\alpha_i y_i=0$$

$$\frac{\partial L}{\partial \mathbf{w}}=0 \quad \Rightarrow \quad \mathbf{w}=\sum_{i=1}^{l}\alpha_i\mathbf{x}_i y_i \,,$$

$$L(\mathbf{w},b,\alpha)=\frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w}-\sum_{i=1}^{n}\alpha_i y_i(\mathbf{w}^T\mathbf{x}_i)-b\sum_{i=1}^{n}\alpha_i y_i+\sum_{i=1}^{n}\alpha_i \quad (5)$$

with $\mathbf{w}^T\mathbf{w} = \sum_{i=1}^{n}\hat{\alpha}_i = \sum_{i=1}^{n}\hat{\alpha}_i = \sum_{i=1}\sum_{j=1}\hat{\alpha}_i\hat{\alpha}_j\left(\mathbf{x}_i{}^T \cdot \mathbf{x}_j\right)y_i y_j$. The constraint as follows:

$0 \le \alpha_i \le C, \quad i = 1,2,...n$ and $\sum_{i=1}^{n}\alpha_i y_i = 0$

From the results of this calculation, α_i most are positive. Data that correlates with α_i positive is called a support vector (Vapnik, 1995). After the solution to the quadratic programming problem is found (value), the class of data to be predicted or testing data can be determined based on the following functions:

$$f\left(\mathbf{x}_t\right) = \sum_{i=1}^{n}\alpha_i y_i x_i x_t + b \tag{6}$$

The SVM method can also be used in non-separable cases by expanding the formulation found in linear cases. The previous optimization problem in both objective functions and constraints was modified by including the Slack variable $\xi > 0$. The slack variable is a measure of misclassification. The formulation is as follows [11].

$$y_i\left[\left(\mathbf{w}^T\mathbf{x}_i\right) + b\right] \ge 1 - \xi_i; \quad i = 1,2,\ldots,n \tag{7}$$

$\Phi\left(\mathbf{w},\xi\right) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i$ so that $\max_{\alpha} L_d = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$

for parameter C functions to control the relationship between slack variables and margins. The greater the value of C, the greater the violation imposed for each classification.

Data whose class distribution is not linear is usually used as a kernel approach to the initial data feature [9]. The mapping process in this phase requires the calculation of dot-product two pieces of data in the new feature space that are denoted as $\Phi\left(\mathbf{x}_i\right)\Phi\left(\mathbf{x}_j\right)$ computational tricks often known as kernel tricks, as follows $K\left(\mathbf{x}_i,\mathbf{x}_j\right) = \Phi\left(\mathbf{x}_i\right)\Phi\left(\mathbf{x}_j\right)$

and predictions on data with the dimensions of features that were newly formulated with

$$f\left(\Phi\left(z\right)\right) = sign\left(w.\Phi\left(z\right) + b\right) = sign\left(\sum_{i=1}^{l}\alpha_i y_i \Phi\left(x_i\right)\Phi\left(z\right) + b\right)$$

with l is the number of data that support vector data, x_i is support vector, and z is testing data that will be predicted. According to [3]the kernel function that used in SVM as follow as:

1. $K(\mathbf{x},\mathbf{y}) = \mathbf{x}.\mathbf{y}$ (Linear);
2. $K(\mathbf{x},\mathbf{y}) = (\mathbf{x}.\mathbf{y} + c)^d$ (Polynomial)
3. $K(\mathbf{x},\mathbf{y}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{y}\|^2\right); \gamma = \frac{1}{2\sigma^2}$ (Radial Basis Function).

2.3. Calculation of Accuracy Value

According to [10] the confusion matrix is a charting table of classification results. For example, confusion matrix elements for classification data with two classes are expressed as fij, then each fij cell states the actual number of records/data entered in class i, but the prediction results classify the data in class j. Table 1 shows an illustration of the confusion matrix.

Table 1. Confusion Matrix

(actual class)	(predicted class)	
	Class 1	Class 2
Class 1	f_{11}	f_{12}
Class 2	f_{21}	f_{22}

The accuracy of classification results that can be calculated with formulas, as follows:

$$The\ accuracy = \frac{f_{11} + f_{22}}{f_{11} + f_{12} + f_{21} + f_{22}}$$

2.4. Methods

In this study consisted of two parts, namely the classification of “raskin” receipt (1: Yes; 2: No) in the city of Semarang using the FKNN and SVM methods. Data were taken from SUSENAS- socio economic survey in 2016 as many as 930 households in Semarang city that conducted by BPS-Statistic agency centers of Semarang city. The independent variables used are as follows X_1: Gender of the households head (1 = Male; 2 = female); X_2: age of households head; X_3: the number of households members; X_4: Highest education of head of households (1: No education; 2: elementary school; 3: junior high school; 4: senior high school: 5: Diploma; 6:

Undergraduate; 7: Master/doctoral degree); X_5: Main wall material of the house (1: Wall; 2: wood; 3: log; 4: Other); X_6: Main House Floor Materials (1: Marble / Granite; 2: Ceramics; 3: Tiles / Tiles; 4; Cement / red brick; 5: Soil; 6: Other); X_7: Use of defecation facilities (1: There are only use of itself; 2: Yes, used together with other households; 3: Yes, in public toilets; 4: None); X_8: Source of Drinking Water (1: Branded bottled water; 2: Refilled water; 3: Led meter; 4: Retail ledge; 5: Drill well / pump; 6: Protected well; 7: Unprotected well; 8: Spring protected: 9: Others); ; X_9: Main cooking fuel (0: no cooking at home; 1: Electricity; 2: gas with capacity 5.5kg (kilograms); 3: gas with capacity 12 kg; 4: gas with capacity 3 kg; 5: kerosene; 6: firewood);

The analysis carried out using the FKNN and SVM methods was divided into 3 scenarios. The scenario I: All independent variables are included in the classification Scenario II: Only variables about the characteristics of the head of the household as independent variables are included in the classification. The characteristics of the head of the household include the sex of the head of the household, the age of the head of the household, the number of household members, and the highest education of the household head. Scenario III: Only variables about the characteristics of residential houses as independent variables are included in the classification. These variables include variables X_5 to X_8

3. RESULTS AND DISCUSSIONS

3.1. The classification of receipt of “raskin” in the city of Semarang uses Fuzzy K-Nearest Neighbor (FKNN).

The first step in classifying FKNN is to determine the K value of the closest neighbor used. After obtaining optimal K, the next step is determining the most influential independent variable in the classification of “raskin” receipts in the city of Semarang. The following table presents accuracy for some K values

Table 2. The accuracy of Scenario I use FKNN Method

The value of K	The accuracy
3	78,76%
5	82,88%
7	80,65%
9	81,59%

Table 2 shows the accuracy value for the scenario I, in the results obtained the optimal number of closest neighbors is for K = 5 reaching 82.88%. Then followed by FKNN analysis using scenario 2, the accuracy value was obtained as follows:

Table 3. The accuracy of Scenario II use FKNN Method

The value of K	The accuracy
3	47,39%
5	68,85%
7	58,65%
9	54,28%

Based on Table 3, the highest accuracy can be seen for K = 5 with an accuracy value of 68.85%. It shows that the characteristic variable of the head of the household only affects the classification results reaching 68.85%. Then followed by FKNN analysis using scenario 3, the accuracy value was obtained as follows:

Table 4. The accuracy of Scenario III use FKNN Method

The value of K	The accuracy
3	76,39%
5	80,85%
7	69,65%
9	74,28%

Table 4 shows the accuracy value for scenario III, based on the table it can be seen that the magnitude of the influence of the characteristics of residential houses in classifying is 80.85%. Based on the results of the analysis using FKNN, variables related to residential characteristics produced better accuracy compared to the characteristics of the head of the household. The optimal number of closest neighbors (K) for each scenario is the same which is equal to K = 5. The next step will be the classification of receipt of poor rice in the city of Semarang using Support Vector Machine (SVM).

3.2. Classification of "raskin" Receipts in Semarang City using Support Vector Machine (SVM)

In the classification of receipt of "raskin" in the city of Semarang using SVM, the kernel functions used are linear, polynomial and Radial Basis Function (RBF). The results of the accuracy obtained for each scenario are as follows:

Table 5. The accuracy of Scenario I use SVM Method

Kernel Function	The accuracy
Linear	87.36%
Polynomial	87.14%
RBF	90%

Table 5 shows the classification of SVM based on the scenario I with various kernel functions. Based on the table, the greatest accuracy value is in the RBF kernel function with a value of the parameter $C = 100$ and. $\gamma = 0.000283$. The level of accuracy produced reaches 90%.

Table 6. The accuracy of Scenario II use SVM Method

Kernel Function	The accuracy
Linear	68.45%
Polynomial	56.87%
RBF	72.77%

Table 6 shows the classification accuracy value using the SVM method for scenario II. Based on the table the highest accuracy is in the RBF kernel function for parameter $C = 100$ and $\gamma = 0.00465$. The resulting accuracy value is 72.77%.

Table 7. The accuracy of Scenario III use SVM Method

Kernel Function	**The accuracy**
Linear	79.45%
Polynomial	80.83%
RBF	84.71%

Table 7 shows the classification accuracy value using the SVM method for scenario III. Based on the table the highest accuracy is in the RBF kernel function for parameter $C = 100$ and. $\gamma = 0.0052$. The resulting accuracy value is 84.71%.

Based on the results of the classification analysis using the SVM method, the scenario I produce the highest accuracy. Scenario I involves all independent variables. For the classification of using SVM, variables regarding the characteristics of residential houses produce better accuracy compared to the variables regarding the characteristics of the head of the household. It shows that the determination of poor rice revenues is more apparent from the characteristics of residential houses than the household head characteristics. This means that it does not guarantee that the characteristics of household heads do not play a role in the classification of poor rice revenues. The results of classification accuracy increase if the two characteristics of the variable are included. When compared with the FKNN method, the classification of the accuracy of receipt of “raskin” in the city of Semarang with the SVM method produces better accuracy.

4. CONCLUSIONS

The classification of “raskin” receipts involving all variables (X_1 to X_8) results in better acquisitions. Meanwhile, the characteristics of residential houses produce better accuracy compared to the characteristics of the head of the household. It shows that in accepting “raskin”,

it is more seen the characteristics of residential houses compared to the characteristics of the head of the household. The classification of receipt of "raskin" in the city of Semarang uses the SVM method better than FKNN. The best function parameter of the kernel used in SVM is $C = 100$ and $\gamma = 0.000283$.

ACKNOWLEDGMENT

Thanks to Diponegoro University for research funding grants in the scheme research development and implementation (RPP) 2018-2019 with no. of contract 474-26/UN7.P4.3/PP/2019.

CONFLICT OF INTEREST

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] E. Riski Ningtiyas, Counterproductive effects of rice for poor (raskin) program on labor supply, J. Perenc. Pembang. Indones. J. Dev. Plan. 2 (2018), 188–202.

[2] BPS, Profil Kemiskinan kota Semarang 2018, Semarang, 2018.

[3] I. Witten, E. Frank, and M. Hall, Data mining: Practical machine learning tools and techniques. Burlington, USA: Morgan Kauffmann, 2011.

[4] J.H.Q. Celis and A.C. Pagatpatan, Predicting the poverty alleviation in the province of eastern samar using data mining techniques, Int. J. Recent. Technol. Eng. 8 (2019), 7140–7145.

[5] L.D.P. Cuong And D. Wang, A comparison of machine learning methods to predict hospital readmission of diabetic patient, Estud. Econ. Apl. 39 (2021), 1-15.

[6] B. Bigdeli and P. Pahlavani, High resolution multisensor fusion of SAR, optical and LiDAR data based on crisp vs. fuzzy and feature vs. decision ensemble systems, Int. J. Appl. Earth Obs. Geoinf. 52 (2016), 126–136.

[7] H.L. Chen et al., A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method, Knowledge-Based Syst. 24 (2011), 1348–1359.

[8] M.F.V. Ruslau and B.S.S. Ulama, Classifying the poor household using neural network, Proc. IConSSE FSM SWCU, 2015, pp. 66–70.

[9] C. Aggarwal, Data mning: Text book. New York: Springer, 2015.

[10] T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning: Data mining, inference and prediction, 2nd ed. California: Springer, 2008.

[11] E. Reynolds, B. Callaghan and M. Banerjee, SVM–CART for disease classification, J. Appl. Stat. 46 (2019), 2987–3007.