International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-9, Issue-10; Oct, 2022
Journal Home Page Available: https://ijaers.com/
Article DOI: https://dx.doi.org/10.22161/ijaers.910.60

Applying the Algorithm of Fuzzy K-Nearest Neighbor in Every Class to the Diabetes Mellitus Screening Model

Maizairul Ulfanita, Alfian Futuhul Hadi, Mohamat Fatekurohman

Department of Mathematics, Jember University, Indonesia

Received: 14 Jun 2022,

Received in revised form: 08 Oct 2022,

Accepted: 13 Oct 2022,

Available online: 31 Oct 2022

***Keywords*— Diabetes Mellitus, Machine Learning, Fuzzy K-Nearest Neighbor in Every Class, Classification.**

***Abstract*—** *Heart disease is the number one killer in the world. Someone who has the potential to experience heart disease is a person with Diabetes Mellitus (DM). Diabetes that is detected early can reduce the risk of heart disease and various other complications. This study aims to analyze the performance of the model that is expected to be an alternative for screening DM by using a machine learning method, namely the Fuzzy K-Nearest Neighbour in Every Class (FKNNC) algorithm. The input in this study was clinical data from Hospital X which consisted of 7 predictor variables and 1 response variable. We used the confusion matrix and the Area Under Curve (AUC) value of the Receiver Operating Characteristic (ROC) curve to maseure the performance of the FKNNC model, with the help of the Python programming language. The results obtained indicate that the FKNNC model is classified as a "good classification" model. This can be seen from the accuracy of the FKNNC model of 86% and F_1-score of 81,8%.*

I. INTRODUCTION

Heart disease is the number one killer in the world. Someone who has the potential to have heart disease is a person with Diabetes Mellitus (DM) [10]. DM is a chronic metabolic disorder caused by the pancreas not producing enough insulin or the body is inability to effectively use the insulin it produces [6]. Therefore, early detection of DM is very important to reduces the risk of heart disease and various other complications such as stroke, obesity, disorders of the eyes, kidneys and nervers [9].

DM is diagnosed by examining HbA1c accorion has a high level of accuracy in diagnosing DM. However, the HbA1c examination has not been evenly distributed in all regions due to the efficiency and effectiveness of the tool due to the high price and availability of human resources capable of operating it. Therefore, another method is needed besides the HbA1c examination which is expected to be an alternative for early DM detection, namely the machine learning method.

Early detection or screening of DM using machine learning techniques that combine computer science, engineering, and statistics [2]. In its application the machine learning method requires a classification algorithm with a series of commands in the form of the Python programming language, namely Fuzzy K-Nearest Neighbor in Every Class (FKNNC) and requires indicators that are expected to be more practical than the HbA1c examination. The indicators used in this study include gender, glucose levels, blood pressure, insulin levels, body weight, diabetes pedigree function and age [5].

II. RESEARCH METHODS

In this study there are several stages of steps. The steps taken to complete this research can be seen in the following scheme.

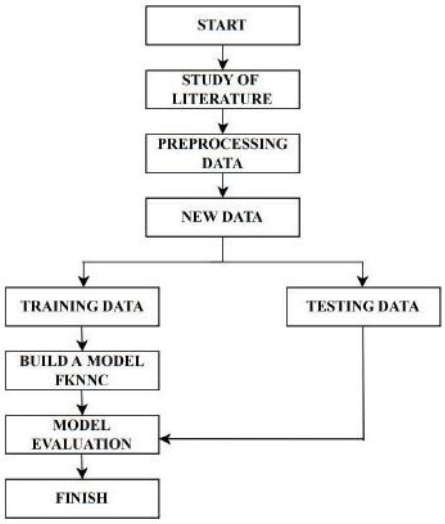

Fig. 1 Schematic of research steps

The first step that will be carried out in this research is a literature study. The author collects several references related to this research. Data preprocessing is an initial technique using Python software to convert the raw data that will be used in this research into new data that is used for further processing.

Several steps in the data preprocessing process are as follows.

- Cleaning Data
  Cleaning data is a step used to overcome missing values and noise.
- Checking the Type Data
  Category data type variables must first be converted to a numeric data type before being processed in Python.
- Data Normalization
  The values in the data need to be normalized so that the learning process is not biased because the number of attributes in the data usually has values in different intervals. The normalization method used is the min-max method [8].

$$x_i' = \frac{x_i - min_A}{max_A - min_A}(newmax_A - newmin_A) + newmin_A \qquad (1)$$

where

x_i : the original value of exploratory variable;
x_i' : the normalized value of x;
min_A : minimum value of attribute A;
max_A : maximum value of attribute A;
$newmin_A$: the restricted minimum value of new values;
$newmax_A$: the restricted maximum value of new values.

- Feature Selection
  This stage is useful in reducing data dimensions, eliminating irrelevant data and increasing accuracy results [12].

After performing the data preprocessing stage, it was followed by the stage of forming a model using the FKNNC algorithm to produce a classification model, where the FKNNC algorithm model produces information and becomes an insight in solving this research problem which is also a process in input and output. Furthermore, testing the FKNNC classification model that has been formed to measure the performance of a model. Measuring the performance of the FKNNC model in this study uses a confusion matrix with several measures of evaluation of the model used in this study.

2.1 The FKNNC Classifier

FKNNC using the *k*-nearest neighbor for each class. FKNNC use a number of *k*-nearest neighbors in each class of a test data [7]. The way FKNNC works begins with determining the *k*-nearest neighbor. For each test data, the *k*-nearest neighbours must be found for each class. Near or far neighbours can be calculated based on Euclidean Distance. Euclidean Distance give a straight distance between the two pieces of data with *N* dimensions. The Euclidean Distance formula is as follows.

$$d(x_i, y_j) = \sqrt{\sum_{l=1}^{N}(x_{il} - x_{jl})^2} \qquad (2)$$

where

$d(x_i, x_r)$: distance
N : the number of variables
l : data variables, l=1,2,...,N
x_{il} : the l^{th} variable for subject i from test data
x_{jl} : the l^{th} variable for subject j from train data

the distance of the test data to all *k*-neighbors from each the *k* class is added up. Formula used are as follows.

$$S_{iK} = \sum_{j=1}^{k} d(x_i, x_j)^{\frac{-2}{m-1}} \qquad (3)$$

the value of *d* is the accumulated distance of the test data to *k*-neighbors in the *k* class as much as (many classes) class. The value of *m* here is the weight exponent, with $m > 1$. Next, the accumulated test data distance to each bclass are summed, symbolized *D*. The formula used is as follows.

$$D_i = \sum_{K=1}^{C}(S_{iK}) \qquad (4)$$

where

C : the number of classes
K : class, K=1,2,...,C

the FKNNC framework uses fuzzy logic, where a test data has a membership value with an interval of [0,1] in each class. A data in all classes with the number of membership values equal to 1, as in the following equation.

$$\sum_{K=1}^{C} u_{iK} = 1, \qquad (5)$$

where is the membership value of test data to class *K* to get the test data membership value in each to *K* class (there are *C* classes) can be used the following formula.

$$u_{iK} = \frac{S_{iK}}{D_i} \qquad (6)$$

to determine the class for the predicted test data, the class with the value of is chosen largest membership of the data. The formula used is as follows.

$$y' = \arg\max\{u_{iK}\} \qquad (7)$$

where *y'* is the prediction class.

2.1 Model Evaluation

To evaluate model fit, we used a confusion matrix including accuracy, sensitivity, specificity, precision, and the f1-scores. We also usesd the AUC vakue of ROC curve to determine the performance of the validated model. Model evaluation can be performed using specific measurements and is calculated as follows [4].

- True Positive (TP), observation has a positive actual value being true clssified with a positive predicted value.
- True Negative (TN), observationhas negative actual value and negative predictied value
- False Positive (FP), observationhas has a negative actual value being miss-clssified with a positive predictied value.
- False Negative (FN), observationhas set has a positive actual value and a negative predictied value.

These four terms are called the confusion matrix.

		Predicted class	
		Negative	Positive
Actual class	Negative	TN	FP
	Positive	FN	TP

Fig. 2 Confusion matrix

The following measures the evaluation of the model used in this study [11].

- Accuracy or recognition rate (ACC)

$$\text{ACC} = \frac{\text{TP+TN}}{\text{P+N}} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad (8)$$

- Sensitivity or recall or True Positive Rate (TPR)

$$\text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP+FN}} \qquad (9)$$

- Specificity or True Negative Rate (TNR)

$$\text{TNR} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN+FP}} \qquad (10)$$

- Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (11)$$

- F_1-score or F-score

$$\text{F} - \text{score} = \frac{2\times\text{precision}\times\text{recall}}{\text{precision+recall}} \qquad (12)$$

- False Positive Rate (FPR), proportion of negative samples were misclassified

$$\text{FPR} = 1 - \text{TNR} = \frac{\text{FP}}{\text{FP+TN}} = \frac{\text{FP}}{\text{N}} \qquad (13)$$

- False Negative Rate (FNR), proportion of positive samples were misclassified

$$\text{FNR} = 1 - \text{TPR} = \frac{\text{FN}}{\text{FN+TP}} = \frac{\text{FN}}{\text{P}} \qquad (14)$$

- The ROC curve
  For testing and and visualizing the overall performance of a classification model here used the ROC curve technique [3]. It is a two-dimensional graph with horizontal axis is the FPR, and the TPR on vertical axis. The larger the ROC curve, the better the classification model.
- The value of AUC
  To measure how large the ROC curve we need the value of AUC. Here we used it to determine which classification model better than others. The AUC values ranges from 0 to 1. The better the classification model is the model has AUC closer to 1 [1].
  AUC values can be divided into several groups classification [3] as follow.
  1. Excellent if : $0{,}90 < \text{AUC} \leq 1{,}00$
  2. Good if : $0{,}80 < \text{AUC} \leq 0{,}90$
  3. Fair if : $0{,}70 < \text{AUC} \leq 0{,}80$
  4. Poor if : $0{,}60 < \text{AUC} \leq 0{,}70$
  5. Failure if : $0{,}50 < \text{AUC} \leq 0{,}60$

III. RESULT AND DISCUSSION

Mathematical modeling for DM screening using FKNNC was compiled using medical records of 1000 clinical data at Hospital X. The data consisted of 8 variables, gender, glucose levels, blood pressure, insulin levels, body weight, diabetes pedigree function, age and diagnosis.

3.1 Building The FKNNC Model for DM Cllssifier

Model The FKNNC classification model was built using train data with a sample of 800 clinical data. The stages in FKNNC modeling are broadly divided into two, namely selecting the number of *k*-neighbors and evaluating the model.

- Selection of the number of *k* neighbors
  The selection of the number of *k*-nearest neighbors is part of the optimization of the FKNNC model. The

optimum value of *k* is the value of *k* that produces the greatest accuracy on the training data or produces the lowest error rate. In this study, *k* is sought by determining in advance the value of *k* in the range 1 to 15.

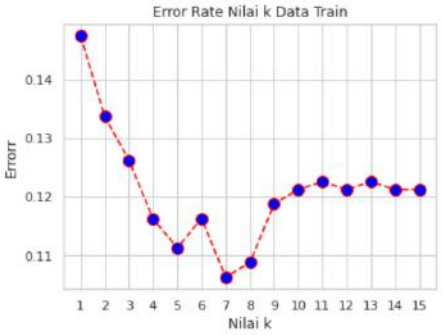

Fig. 3 The k-value

Figure 3. Is a graph of the relationship between the value of k and the resulting error rate. In Figure 3, it is found that the optimum *k* is *k*=7. This indicates that the number of nearest neighbors used is 7. Then it is processed based on the steps of the FKNNC algorithm.

- Evaluation of FKNNC Model

  Before the model is used for DM screening, the model is tested first. The benchmark used is the level of model accuracy in predicting the training data. The prediction results from the FKNNC model can be compared with the actual value, then a confusion matrix and the evaluation value of the model are obtained (Figure 4).

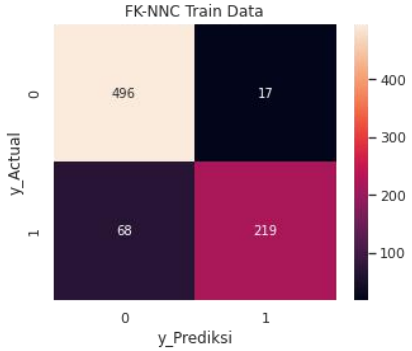

Fig. 4 Confusion matrix

Figure 4 shows the TP, TN, FP and FN values of the FKNNC model, respectively. In each confusion matrix, the first column of the first row is the TN value which indicates the number of negative DM correctly predicted negatively by the model. The second column of the second row is the TP value which indicates the number of DM positives that are correctly predicted to be positive. The first column of the second row is the FN value which indicates the number of positive DMs that are incorrectly predicted to be negative and the second column of the first row is the FP value which indicates the number of negative DM that are incorrectly predicted to be positive.

Table 1. Evaluation of FKNNC classification model on training set

Accuracy	Sensitivity	Specificity	Precision	FPR	FNR	F_1-score
0,894	0,763	0,967	0,928	0,033	0,237	0,837

Table 1 shows that the accuracy on the training set was 89.4%, with details of the positive predictive accuracy of DM at 76.3% and the negative predictive accuracy of DM being 96.7%. This certainly has an impact on the error rate for positive DM and negative DM predictions where the negative DM prediction results have a potential error of 23.7% and positive DM prediction results have a potential error of 3.3%. To conclude how the quality of the resulting model, the benchmark used is the AUC value on the ROC curve.

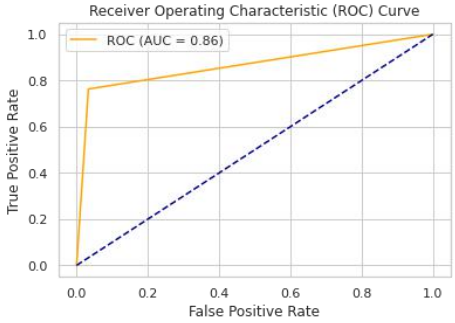

Fig. 5 The ROC curve

Figure 5 is an ROC curve with an AUC value of 0.86 which indicates that the resulting FKNNC model is classified as "good classification" or a classification model with good capabilities.

3.2 Diabetes Mellitus (DM) Screening Using Fuzzy K-Nearest Neighbor in Every Class (FKNNC) Model

The ability of a model to predict a response can be tested with test data. The function of the test data is as a comparison with the initial assumption that the response value is unknown. After the prediction results are obtained, they are then compared with the response to the testing data.

Table 2. The Results screening

		Screening results	
		Negative	Positive
Actual class	Negative	107	18
	Positive	10	65

Table 2 shows the TP, TN, FP and FN values from the predicted results of the test data using the FKNNC model. Table 2 shows the confusion matrix for the FKNNC model with values for TN 107, TP 65, FP 18 and FN 10. The TP, TN, FP and FN values taken from the FKNNC model to calculate performance measures using model evaluation measures.

Table 3. Evaluation of FKNNC classification model on testing set

Accuracy	Sensitivity	Specificity	Precision	FPR	FNR	F_1-score
0,860	0,867	0,856	0,783	0,144	0,133	0,818

Based on table 3, it shows that the accuracy of the FKNNC model on testing set was 86%, with details of the positive predictive accuracy of DM at 86.7% and the negative predictive accuracy of DM being 85.6%. This certainly has an impact on the error rate of positive and negative DM predictions where the negative DM prediction results have a potential error of 13.3% and the positive DM prediction results have a potential error of 14.4%.

IV. CONCLUSION

Based on the results of the research and discussion, it is concluded DM using factors of gender, glucose levels, blood pressure, insulin levels, body weight, diabetes pedigree function and age is classified as a "good classification" model. The accuracy rate of the FKNNC model is 86% and F_1-score is 81,8%.

REFERENCES

[1] Bradley, A. P. (1997). The Use of The Area Under The Curve in The Evaluation of Machine Learning Algorithms. Pattern Recognition: Department of Electrical and Computer Engineering, The University of Queensland, Australia. 30(7): 1145-1159.

[2] Dangeti, P. (2017). Statistics for Machine Learning. 1st ed. Birmingham: Packt, Publishing.

[3] Gorunescu, F. (2011). Data Mining Concepts Model and Techniques. Intelligent Systems Reference Library. 12: Springer-Verlag Berlin Heidelberg.

[4] Han, J., M. Kamber dan J. Pei (2012). Data Mining: Concepts and Techniques. San Fransisco, CA, itd: Morgan Kaufmann (Third). Waltham USA: Elsevier.

[5] Kumiawaty, E and B. Yanita (2016). Faktor-faktor yang Berhubungan dengan Kejadian Diabetes Mellitus Tipe II. Biokimia Fakultas Kedokteran Universitas Lampung. 5(2): Majority.

[6] Kemenkes, RI. (2014). Situasi dan Analisis Diabetes. Jakarta: Kementerian Kesehatan RI.

[7] Prasetyo, E. (2012). Fuzzy K-Nearest Neighbor in Every Class untuk Klasifikasi Data. Seminar Nasional Teknik Informatika (SANTIKA 2012). Teknik Informatika-Fakultas Teknologi Industri Universitas Pembangunan Nasional "Veteran" Jawa Timur: pp 57-60.

[8] Suyanto (2018). Machine Learning Tingkat Dasar dan Lanjut: Informatika Bandung.

[9] V, V. Vijan and A. Ravikumar (2014). Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus. International Journal of Computer Applications. 95(17): 0975-8887.

[10] Widiastuti, N.A., S. Santoso and C. Supriyanto (2014). Algoritma Klasifikasi Data Mining Naive Bayes Berbasis Particle Swarm Optimization untuk Deteksi Penyakit Jantung. Jurnal Pseudocode. 1(1): ISSN 2355-5920.

[11] Wei, W.W.S. (2006). Time Series Analysis Univariate and Multivariate Method. Canada: Addision Welsey.

[12] Yu, L and H. Liu (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). Washington DC: AZ85287-5406.