INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND ANALYSIS

ISSN(print): 2643-9840, ISSN(online): 2643-9875

Volume 04 Issue 01 January 2021

DOI: 10.47191/ijmra/v4-i1-03, Impact Factor: 5.522

Page No.- 18-23

Performa Comparison of the K-Means Method for Classification in Diabetes Patients Using Two Normalization Methods

Dwianti Westari[1], Dr. Abdul Halim, M. Eng[2]

[1]Dwianti Westari, Universitas Indonesia

[2]Dr. Abdul Halim, M. Eng, Universitas Indonesia

ABSTRACT: The diabetes classification system is very useful in the health sector. This paper discusses the classification system for diabetes using the K-Means algorithm. The Pima Indian Diabetes (PID) dataset is used to train and evaluate this algorithm. The unbalanced value range in the attributes affects the quality of the classification result, so it is necessary to pre-process the data which is expected to improve the accuracy of the PID dataset classification result. Two types of pre-processing methods are used that are min-max normalization and z-score normalization. These two normalization methods are used and the classification accuracies are compared. Before the data classification process is carried out, the data is divided into training data and test data. The result of the classification test using the K-Means algorithm has shown that the best accuracy lies in the PID dataset which has been normalized using the min-max normalization method, which 79% compared to z-score normalization.

KEYWORDS- diabetes, k-means, min-max normalization, z-score normalization, Pima Indian Diabetes (PID)

I. INTRODUCTION

Diabetes mellitus (DM), also known as diabetes, is a disease caused by pancreatic cells that do not produce enough insulin so that blood sugar rises [1]. Symptoms that often result from high blood sugar include frequent urination (*polyura*), increased thirst (*polydipsia*), and increased hunger (*polyphagia*). Diabetes can also lead to serious long-term complications such as heart disease, stroke, kidney failure, leg ulcers, and eye damage [2].

Apart from long term complications, diabetes also causes premature death. Therefore, it is necessary to do a diabetes diagnosis in the form of tests and analysis of diabetes disease factors. However, this is constrained by time and limited medical personnel who are experienced in their field. Today's machine learning algorithms are used to classify and diagnose a disease thereby eliminating the problems and reducing costs required by producing meaningful and accurate decisions [3].

The Pima Indian Diabetes (PID) dataset from the University of California, Irvine (UCI) Repository of Machine Learning database is used to train and evaluate machine learning algorithms. The irrelevant data (noise) will influence the decision of the algorithm to be used [4]. One of the algorithms used is the K-means algorithm which is used to categorize and classify patients into healthy (non-diabetic) and diabetic categories.

II. METHOD

A. Research Dataset

The dataset used is the Pima Indian Diabetes (PID) dataset from UCI Machine Learning. The dataset is 768 data. The data consists of 8 attributes and 1 output attribute in the form of a class with a range of 0-1. Range 0 means negative (non-diabetic / healthy) and 1 means positive (diabetes). The diabetes attribute data is shown in Table 2.1 [5].

Table 2.1.List of Diabetes Data Attributes

Attribute	Abbreviation	Description	Unit	Data Type
Pregnant	Pregnant	Number of Pregnancy	-	Continuous
Plasma Glucose Concentration	OGTT	Glucose levels 2 hours after meals	mg/dl	Continuous
Diastolic Blood Pressure	Diastolic	Blood Pressure	mmHg	Continuous

Triceps Skin Fold Thickness	TSFT	Skin thickness	Mm	Continuous
2-Hour Serum Insulin	INS	Insulin	Mu U/ml	Continuous
Body mass index	IMB	Bodyweight	Kg/m^2	Continuous
Diabetes Pedigree Function	DPF	Family history of diabetes	-	Continuous
Age	Age	Patient's age	Year	Continuous

B. Pre-processing Data

The preprocessing stage is carried out as an initial stage and an important stage in research in determining the quality and quantity of data for diagnosis and prediction accuracy in each classification model. This results in low data quality so that no quality results can be found. In making data, it is important to make adjustments for data analysis in terms of time, cost, and quality [6].

1. . Data reduction

Reduced data provides a reduced representation of a data set that is much smaller in size but still produces the same analytical output. It includes dimensional and number reduction strategies [6]. In this methodology, the number reduction strategy is used by using a new algorithm proposed by reducing the number of samples from 768 to 333.

2. Data Transformation

Data transformation was performed by refining and normalizing data [7]. Normalization is the process of scaling the attribute values of data so that they can be placed in a certain range (0 and 1) [8]. The normalization stage used in this study is the Min-Max Normalization by carrying out a linear transformation of the original data so as to produce a balance comparison of values between the data before and after the process [9] and the Z-Score Normalization based on the mean (average value) and standard deviation data. (standard deviation) by knowing the actual minimum and maximum value of the data [10]. Equations 3 and 4 are used in this method.

$$\textit{Normalized}\ (X') = \frac{X - Xmin}{Xmax - Xmin} \qquad (1)$$

$$\text{New value}\ (X') = \frac{actual\ value - mean}{stdev} \qquad (2)$$

C. K-Means Clustering

The K-Means Clustering method was first introduced in 1976 by MacQueen JB. This is the most commonly used method [9]. This method divides or separates the K-th motorcycle taxi into separate sections. In this method, each object will belong to a certain group in a certain process.

The K-means algorithm is a centroid model. This centroid model uses the midpoint of a cluster which is a value. The centroid is used in calculating the distance of an object to the center point. A data object is included in a cluster if it has the shortest distance from the cluster centroid.

The steps in the K-Means algorithm for classifying data are [10]:

1. Determine the number of clusters (k)
2. Determine the centroid
3. Calculate the distance of the object to the centroid
4. Data is grouped by the closest distance
5. Check if there are objects that move:
    a. If so, return to step 2
    b. If not, you're done

D. Euclidean Distance

Euclidean distance is a distance calculation method used to measure the distance of two points in the Euclidean space (covering two-dimensional or more Euclidean planes). To measure the level of data similarity with the Euclidean distance formula, formula [11] is used.

$$d\,(x,y) = |x\text{-}y| = \sqrt{\textstyle\sum_{i=1}^{n}(xi - yi)^2} \qquad (3)$$

Where,

d = distance between x and y

x = cluster center data

y = data on attributes
i = each data
n = amount of data,
xi = data at the center of the ith cluster
yi = data on each data ith

E. Classification

Classification is the grouping of an object into a certain class. Various cases related to the grouping of objects can be solved by applying classification techniques [12].

The classification algorithm uses training data to create a model. The model that has been built is then used to predict the labels of new unknown data classes. But the principle of each algorithm is the same, namely to carry out a training so that at the end of the training the model can predict each input vector to the output class label accurately [13].

A system in performing classification is expected to be able to classify all data sets correctly, but it cannot be denied that the performance of a system cannot be 100% accurate.
To measure the level of accuracy, the following formula can be used:

$$\text{Accuracy} = \frac{\sum correct\ test}{\sum total\ data\ amount\ of\ test\ data}\ x\ 100\% \quad (4)$$

III. RESEARCH DESIGN

In principle, there are three steps in the diagnosis of a disease using machine learning, namely data collection, pre-processing, and diagnosis of a disease with an appropriate and appropriate classification model. The flowchart flow (Figure 3.1) of the diagnosis process of diabetes mellitus in the completion of the diabetes classification design is as follows:.

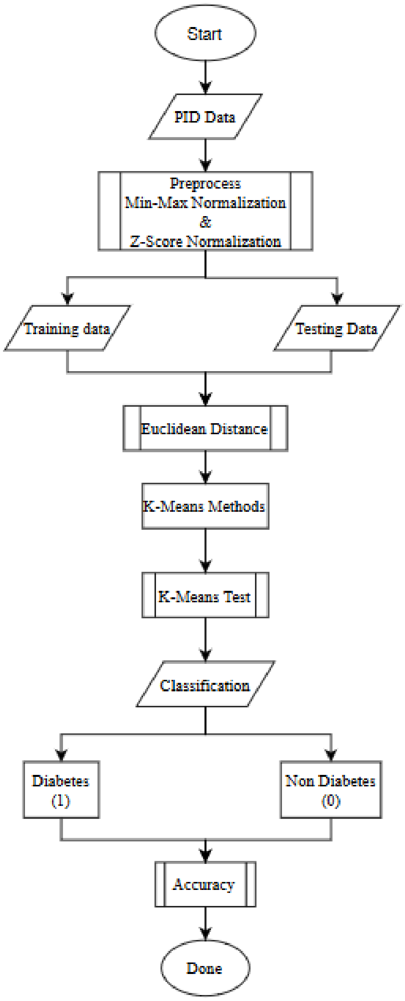

Figure 3.1 Flowchart of Diabetes Diagnosis Methodology

IV. TESTING AND ANALYSIS

At this stage, the results of data testing on the system will be described and their explanation. Testing is done by normalizing the data and the K-Means method which aims to test and evaluate the accuracy of the system based on test data against training data.

A. *Testing Non-Reduction Data and Data Reduction with Min-Max Normalization*

Normalization At this stage the test is carried out by changing the amount of training data four times with the min-max normalization method. The number of training data used was 50, 100, 150, and 200 with a range of 50 training data. The number of test data used was still yaitun768 test data. The test results are shown in Figure 4.1.

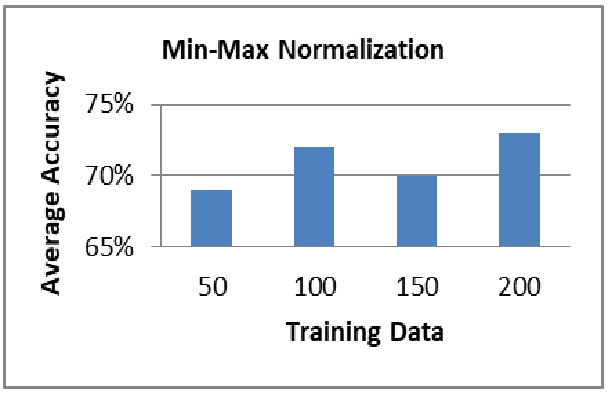

Figure 4.1 Testing of Training Data with the Z-Score Method (Non-Reduction)

The first test is conducted to determine the comparison of the average accuracy rate of the min-max normalization and k-means methods. In this section, each process in each algorithm uses 50, 100, 150, 200 training data and 568 and 233 training data.

Based on Figure 4.1 it can be seen that when the increase in the amount of training data will affect the average accuracy which tends to go up and down as in the data 100 has an increase of 72% but has decreased in the training data 150 is 70% and will experience an increase in average accuracy. on the training data 200 that is 73%. This happens because the greater the amount of training data used, the system will recognize the existing patterns in the training data when testing so that the accuracy obtained will be better.

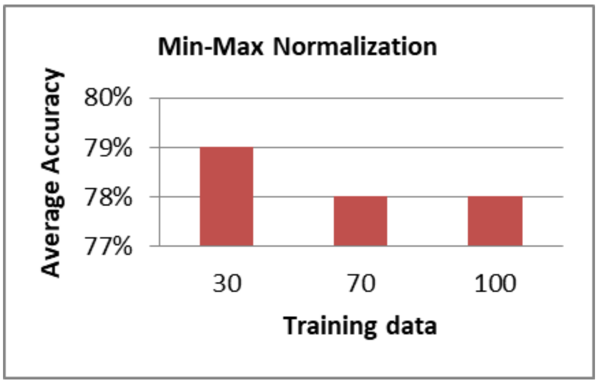

Figure 4.2 Testing Training Data with Z-Score (Data Reduction)

Whereas in Figure 4.2 on the contrary, the more training data, the smaller the average accuracy. The highest average accuracy value is found in the number of training data 30, namely 79%, but there is a decrease in the number of training data 40 and 100. This is because the range of training data is not the

B. *Testing Non-Reduction Data and Reduction Data Using Z-Score Normalization*

In the second test, the same test was carried out by changing the training data four times, namely 50,100,150, 200 training data with a fixed amount of test data, namely 568 and 233 training data using the z-score and k-means methods. The test results are shown in Figures 4.3 and 4.4.

From Figure 4.3 it can be seen that, the more the amount of training data, the better the average accuracy value as shown in the 200 training data with an average accuracy of 73%. However, like the previous test, this test also experienced a decrease in the average accuracy of the 150 training data, which was 67% smaller than the min-max normalization method. As previously noted, the average accuracy value is influenced by the distribution of data from each class and the amount of training data used.

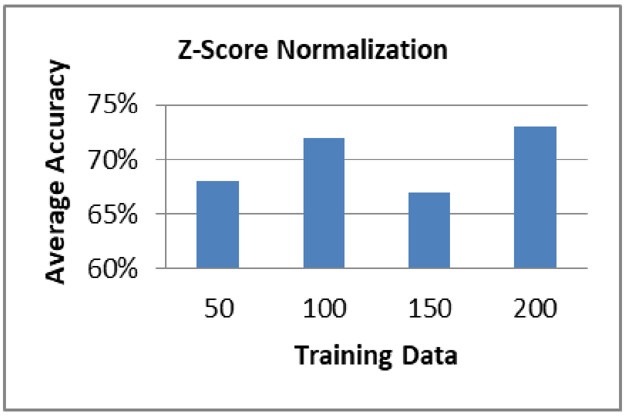

Figure 4.3 Testing Training Data with the Z-Score Method (Non-Reduction)

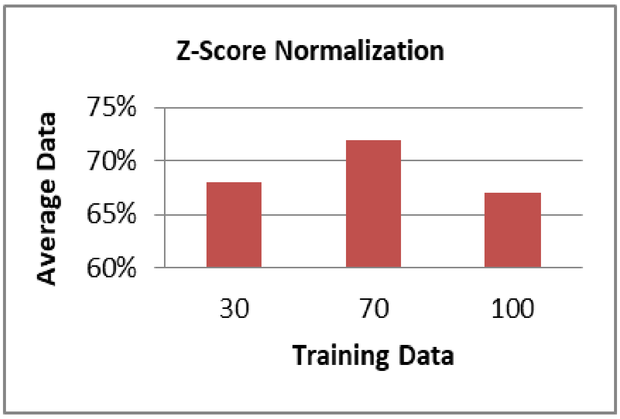

Figure 4.4 Testing of Training Data with Z-Score (Data Reduction)

Figure 4.4 can be seen that the graph shows different results from Graph 4.3 with the same method and amount of training data. The training data 30 to 70 experienced an increase in the average accuracy value from 68% to 72%, but there was a decrease in the training data of 100 with an average accuracy of 67%. The addition of training data to this method affects the average accuracy tends to rise and fall so that it cannot form a pattern. This is influenced by the distribution of data from each class and the amount of training data used

CONCLUSIONS

1. This study can classify diabetic and non-diabetic patients by calculating the distance between the training data and the test data using the min-max normalization and z-score normalization and the k-means method.
2. The highest average accuracy lies in the Pima Indian Diabetes (PID) dataset which uses the min-max normalization method on training data 30 by 79%
3. The lowest accuracy is found in the Pima Indian Diabetes (PID) dataset using the z-score normalization method, with an average accuracy of 67%.
4. The choice of data preprocessing method on PID data affects the accuracy of the results of data classification.

REFERENCES

1) American Diabetes Association, Diagnosis and classification of diabetes mellitus, Diabetes Care 32 (Supplement1) (2009) S62–S67
2) "About diabetes". World Health Organization. Retrieved 4 April 2014
3) Yilmaz N., Inan O., Uzer M.S., " A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases," J Med Syst, vol. 38, no. 5 2014.
4) Lowongtraool C., Hiransakolwong N., "Noise filtering in unsupervised clustering using computation intelligence," International Journal of Math, vol. 6, no. 59,pp. 2911-2920,2012

5) Nirmala Devi M.,Appavu alias Balamurugan S.,Swathi U.V., 2013.", An amalgam KNN to predict Diabetes Mellitus", IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology(ICECCN), pp 691-695.

6) Panwar, Madhuri, et al. "K-nearest neighbor based methodology for accurate diagnosis of diabetes mellitus." *2016 Sixth International Symposium on Embedded Computing and System Design (ISED)*. IEEE, 2016.

7) Malley B, Ramazzotti D, Wu J T-y. Data pre-processing. Secondary analysis of electronic health records. Springer; 2016. p. 115–41.

8) Fikriya, Arina Ashfa, and Sanny Hikmawati. "Support Vector Machine Predictive Analysis Implementation: Case Study of Tax Revenue in Government of South Lampung." *Proceeding International Conference on Science and Engineering*. Vol. 3. 2020.

9) T. T. Hanifa, S. Al-faraby, F. Informatika, and U. Telkom,"*Analisis Churn Prediction pada Data Pelanggan PT . Telekomunikasi dengan Logistic Regression dan Underbagging*," vol. 4, no. 2, pp. 3210–3225, 2017.

10) Azzahra, Darnisa Nasution dkk. *Perbandingan Normalisasi Data Untuk KlasifikasiWine Menggunakan Algoritma K-NN*. Journal of Caomputer Engineering System and Science Vol. 4 No. 1. Januari 2019.

11) Noor Fitriana, Perbandingan Kinerja Metode Lingkage, Metode Average Lingkage, dan Metode K-Means Dalam menentukan hasil analisis cluster, (Jogyakarta: UNY, 2014), h.22

12) Praja, Abdi, Chairisni Lubis, and Dyah Erny Herwindiati. "Deteksi Penyakit Diabetes dengan Metode Fuzzy C-means Clustering dan K-means Clustering." *Computatio: Journal of Computer Science and Information Systems* 1.1 (2017): 15-24.

13) Garcia-Carretero, Rafael, et all. 2020. *Use of a K-Nearest Neighbors Model to Predict The Development of Type 2 Diabetes Within 2 Years in An Obese, Hypertensesive Population*. International Federation for Medical and Biological Engineering.

14) N. Syafitri. Perbandingan metode KNearestNeighbor (KNN) dan Meode Nearest Cluster Classifier (NCC) Dalam Pengklasifikasian kualitas batik tulis. Teknologi Informasi & pendidikan, Vol 2, no. 1, pp. 45-46, 2010

15) B. A. Muktamar.Implemenasi dengan Naive Bayes Classifier untuk mendukung Strategi pemasaran di Bagian Humas STMIK AMIKOM Yogyakarta [Laporan Penelitian], Yogyakarta, 2013.