ILKOM Jurnal Ilmiah Vol. 15, No. 2, August 2023, pp. 215-221
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Diabetes Mellitus Early Detection Simulation using The K-Nearest Neighbors Algorithm with Cloud-Based Runtime (COLAB)

Mohamad Jamil [a,1,*]; Budi Warsito [b,2]; Adi Wibowo [b,3] ; Kiswanto [b,4]
[a] *Universitas Khairun, Jl. Jusuf Abdulrahman Kotak Pos 53 Gambesi, Ternate, 97719, Indonesia*
[b] *Universitas Diponegoro, Jl. Imam Bardjo SH No.5, Pleburan, Kota Semarang, 5024, Indonesia*
[1] *jamilkhairun@gmail.com;* [2] *budiwarsitoundip@gmail.com;* [3] *bowo.adi@undip.ac.id;* [4] *kiswanto@students.undip.ac.id*
** Corresponding author*

Article history: Received October 30, 2022; Revised November 17, 2022; Accepted June 10, 2023; Available online August 16, 2023

Abstract

Diabetes Mellitus is a genetically and clinically heterogeneous metabolic disorder with manifestations of loss of carbohydrate tolerance characterized by high blood glucose levels as a result of insulin insufficiency. Public knowledge of diabetes mellitus 39.30% is influenced by public health education and information about diabetes mellitus that the public has ever received. Early detection of diabetes mellitus can prevent the development of chronic complications and allow timely and rapid treatment. The aim of this study is to simulate the early detection of diabetes mellitus with the K-Nearest Neighbors (K-NN) algorithm using Cloud-Base Runtime (COLAB). The highest accuracy is 76% in k=3, the highest precision is 68% in k=3, and the highest recall is 60% in k=3. The researchers used K-NN as a method to classify data from the Pima Indians Diabetes Database and obtained a fairly good accuracy value of 76% with a value of k = 3.

Keywords: Classification; COLAB; Diabetes; K-Nearest Neighbors.

Introduction

Diabetes Mellitus (DM) is a metabolic disorder that is genetically and clinically heterogeneous with manifestations in the form of loss of carbohydrate tolerance characterized by high blood sugar levels as a result of insufficient insulin function. Diabetes mellitus has become a global health problem. The prevalence and incidence of diabetes mellitus are increasing drastically all over the world including Indonesia [1]. Southeast Asia ranks as the world's third-largest region for diabetes mellitus prevalence. The prevalence of diabetes mellitus in 2019 in individuals aged 20-79 years was 8.3%. There is an increase in the prevalence of diabetes mellitus in people aged 65-79 years, which is 19.9% or 111.2 million. It is estimated that in 2045 the prevalence of people with diabetes mellitus will increase by 700 million [2]. This number is predicted to grow to 700 million people by 2045. Most people with diabetes live in low and middle-income countries, while 1.6 million deaths are caused directly by diabetes each year. That makes diabetes one of the ten leading causes of death worldwide [3].

Diabetes can be caused by various factors, including poor diet, chemicals or toxic content in food, environmental pollution, bacteria, virus infections, eating habits, lifestyle changes, and obesity [4]. The high number of people with DM is due to sufferers not realizing the initial symptoms that appear, including frequent urination (polyuria), frequent thirst (polydipsia), and a lot of eating easy hunger (polyphagia). In addition, there are often symptoms of blurred vision, tingling in the hands or feet, itching that is often very disturbing (pruritus), and weight loss for no apparent reason. For further detection, laboratory tests are needed to check the blood sugar level, fasting blood sugar level, and plasma glucose content after 2 hours [5].

Research on diabetes mellitus shows that late and improper treatment of people with diabetes mellitus results in uncontrolled blood sugar levels over a long period of time, a condition that causes serious changes in the heart, blood vessels in the brain and legs, nerves, kidneys, and eyes. Public knowledge about diabetes mellitus 39.30% is influenced by public health education and information about diabetes mellitus received by the community [6], [7]. Early detection of diabetes mellitus events can prevent the onset of chronic complications and allow appropriate and rapid treatment. The increasing amount of health data and the need for fast and accurate presentation of information is encouraging

http://dx.doi.org/10.33096/ilkom.v15i2.1510.215-221

the application of technology in various aspects of the health sector [8]. The application of this technology certainly requires new methods of processing and presenting information so that it can be used by different groups such as academics, government, and the general public. Different data mining methods can be used to solve classification problems. The number of attributes can affect the performance of an algorithm, several algorithms can be used to perform classification tasks, one of the best data mining classification algorithms is K-NN.

The K-Nearest Neighbors or K-NN technique is a classification model that has several advantages, its application is simple but effective in many cases [9], [10], [11]. Training on K-NN is very fast and robust, even in the presence of noise data. K-NN also performs well in systems where a sample has many class labels. K-Nearest Neighbors (K-NN) is a lazy learning algorithm that uses a classification method for objects based on the data closest to the object. K-NN is also part of the instance-based learning group [12], [13]. K-Nearest Neighbors or K-NN is an algorithm that classifies data based on training data sets taken from the K nearest neighbors. Several studies have been conducted by previous researchers related to the identification of diabetes mellitus using K-Nearest Neighbors (K-NN), Yunita, et al [14] in their research found that K-NN is able to classify diabetes disease data based on age, diet influence, and other factors. Furthermore, research of Asmarani [15] explained that K-NN has good accuracy for diabetes disease data. The highest accuracy obtained was 66.6667% with a value of k = 5. Mahalisa, et al [16] also explained in their research that object classification is very important. The number of attributes can affect the performance of an algorithm, in the simulation results a good accuracy value of 76% was obtained.

Method

The process of discovering a model (or function) that describes and distinguishes classes of data or concepts, to be used to predict the class of an object whose class label is unknown [17]. Widely used classification algorithms include decision/classification trees, Bayesian classifiers/Naïve Bayes classifiers, neural networks, statistical analysis, genetic algorithms, rough sets, K-Nearest Neighbors, rule-based methods, memory-based reasoning, and support vector machines (SVM).

Our research methodology consists of several stages: first, we collect the dataset. Second, we pre-process the dataset to build the model. Thirdly, we find the prediction model of the K-NN classification algorithm, fourthly, we run the K-NN algorithm to predict the diabetes indication. **Figure 1** shows a diagram of the stages of the K-Nearest Neighbors (K-NN) process.

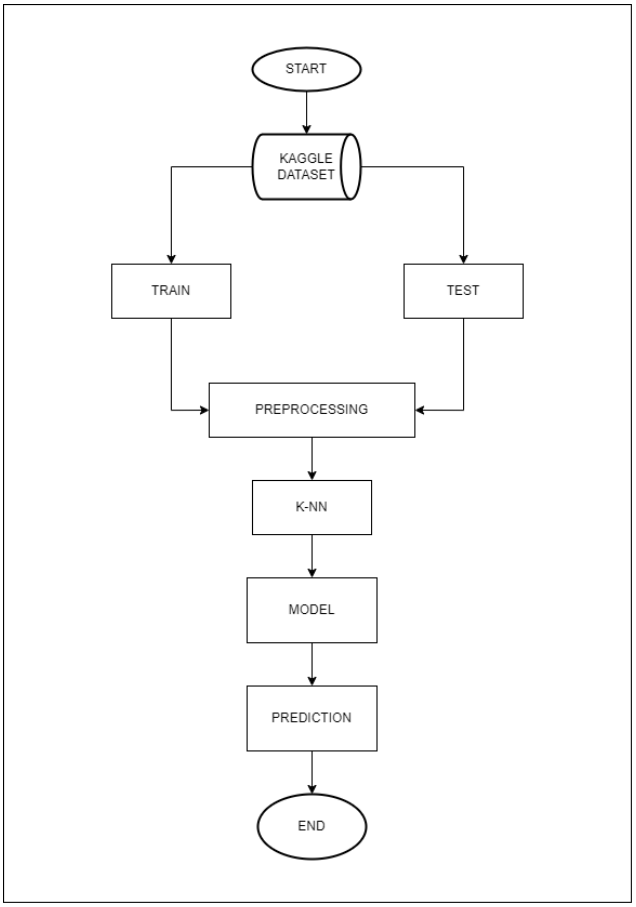

Figure 1. Flowchart of the K-NN stages

A. Classification

Classification is the process of grouping an object into a particular class. Various cases related to object grouping can be solved by applying classification techniques [18]. Classification algorithms use training data to build a model. The built model is then used to predict the class label of new unknown data.

B. K-Nearest Neighbor

The K-Nearest Neighbor (K-NN) algorithm is an algorithm that determines the distance value in test data with training data based on the smallest value of the nearest neighbor value. The purpose of this algorithm is to classify new objects based on attributes and training samples [19]. The K-NN algorithm is a method that uses a supervised algorithm [20], [21]. The difference between supervised and unsupervised learning is that supervised learning aims to find new patterns in the data by associating existing data patterns with new data. In unsupervised learning, the data does not yet have a pattern, and the purpose of unsupervised learning is to find patterns in the data. The stages of the K-NN algorithm are explained as follows:

1. Preparation of training data and test data
2. Determine the value of k
3. Calculate the distance of the test data to each training data.
4. The training data is calculated using the Euclidean distance formula as follows Equation 1:

$$d(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (X_{1i} - X_{2i})^2} \quad (1)$$

5. Determine the value of k training data that is closest to the test data.
6. Check the labels of the k nearest training data
7. Identify the label with the highest frequency
8. Put the test data into the class with the highest frequency
9. Condition stop.

Results and Discussion

In this study we present the stages in the simulation process of early detection of diabetes mellitus (DM). The steps are explained as follows:

A. Dataset

The dataset used in this study is from the Pima Indians Diabetes Database, published by UCI Machine Learning on the website https://www.kaggle.com/. The dataset consists of several medical predictor variables (independent). There are 9 variables that determine whether someone has diabetes or not. The variables are shown in **Table 1**.

Table 1. Predictor Variables

Variable	Description
Pregnancies	The percentage of times a woman has been pregnant in her lifetime
Insulin	Insulin level in 2-hour serum insulin in units of mu U/ml
Glucose	Percentage of blood glucose at 2 hours in the glucose tolerance test
BMI	Body mass index (weight in kg/height in metres)
Blood Pressure	Blood pressure in mm/hg
Diabetes Pidegree Function	Indicators of a family or hereditary history of diabetes
Skin Thickness	A measurement of body fat taken on the arm midway between the olecranon process of the elbow and the acromial process.
Age	Age of a sample

where the dependent variable is the output of the prediction results with a class variable of 0 or 1, 0 if a sample does not have diabetes and 1 if a sample has diabetes.

B. Preprocessing

At this stage, missing data values are checked as the dataset may contain incomplete data. Missing data values are replaced by the median value of each variable, so that each data in the dataset variable has an absolute value. **Table 2** shows a preprocessing data.

Table 2. Preprocessing

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	0.288	33	1

C. Modeling

In this stage, the application of models and techniques is adjusted from several parameters to obtain an optimum value. **Figure 2** shows a modeling implementation on COLAB.

```
[19] from sklearn.model_selection import train_test_split
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/3, random_state = 42, stratify=y)

     from sklearn.neighbors import KNeighborsClassifier

     train_scores = []
     test_scores = []

     for i in range(1,15):
         knn = KNeighborsClassifier(i)
         knn.fit(X_train, y_train)
         train_scores.append(knn.score(X_train, y_train))
         test_scores.append(knn.score(X_test, y_test))
```

Figure 2. Modeling Implementation on COLAB

Before proceeding with the prediction process, the author will standardize the scales of all variables so that they can be scored uniformly. If one of the variables has a distance with a large range (different scales), the value must be adjusted. So, the distance between all variables must be standardized so that the final distance obtained is proportional. Next, we start the prediction with k=3 and continue with k=5 (where k is determined by each researcher) as show in **Figure 3**.

```
[68] from sklearn.neighbors import KNeighborsClassifier

     train_scores = []
     test_scores = []

     for i in range(1,6):
         knn = KNeighborsClassifier(i)
         knn.fit(X_train, y_train)
         train_scores.append(knn.score(X_train, y_train))
         test_scores.append(knn.score(X_test, y_test))
```

```
[69] max_test_score =max(test_scores)
```

```
[70] test_score_index = [i for i, v in enumerate(test_scores) if v== max_test_score]

     print('Max test score {} % and k = {}'.format(max_test_score*100,list(map(lambda x: x+1, test_score_index))))

     Max test score 74.609375 % and k = [3]
```

Figure 3. Determination of K Value on COLAB

The next step is to apply the K-NN method, the selection of k values in this study is the value of k = 3 and k = 5. **Table 3** shows the experimental results of the K-NN method at k = 3 and k = 5.

Table 3. Results of Installing Class k

K=3, K=5 (TP=True Positive, FN= False Negative)		
1	1	TP
0	0	FN
1	1	TP
0	0	FN
0	0	FN
-----	-----	-----
-----	-----	-----
0	0	FN
0	0	FN

Based on Table 3, the next step is to translate the results into a confusion matrix. Table 4 shows the confusion matrix for k=3. Figure 4 show the confusion matrix process on COLAB.

Table 4. Confusion Matrix k=3

N=120	Predicted: 1	Predicted: 0
Actual : 1	37	52
Actual : 0	138	29

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, fbeta_score
y_pred = knn.predict(X_test)

cnf_matrix = confusion_matrix(y_test, y_pred)
```

Figure 4. Confusion Matrix Process on COLAB

Once applied to the confusion matrix, the performance of the method can be measured at a value of k = 3 as shown in Table 5, in this study the performance measured is accuracy, precision, and recall.

Table 5. K-NN Simulation Results of k=3

No	Model	Accuracy	Precission	Recall	F1 Score	F2 Score
0	KNN	0.765625	0.683544	0.606742	0.642857	0.622069

The next step continues with the confusion matrix experiment at k=5, based on the data previously presented in Table 3. Table 6 and Table 7 shows the results of the confusion matrix at k=5.

Table 6. Confusion Matrix of k=5

N=120	Predicted: 1	Predicted: 0
Actual : 1	37	55
Actual : 0	138	23

Table 7. K-NN Simulation Results of k=5

No	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
0	KNN	0.742188	0.626374	0.640449	0.633333	0.637584

Conclusion

From the results of the calculation and simulation of the K-Nearest Neighbors (K-NN) algorithm above, it has obtained the highest accuracy result of 76% at k = 3, the highest precision is 68% at k = 3, the highest recall is 60% at k = 3. Researchers using K-NN as a method to classify data from the Pima Indians Diabetes Database obtained a fairly good accuracy value of 76% with a value of k = 3. Suggestions for further research are to conduct the same experiment by adding the amount of data and applying cross-validation.

References

[1] E. Irawaty, N. Hendry, P. Sunardi, and F. Muatiara, "Skirining Faktor Risiko Penyakit Diabetes Melitus sebagai Upaya Pencegahan di Kelurahan Tomang Jakarta Barat pada masa Pandemi COVID 19," pp. 889–896, 2020.

[2] V. Agustina *et al.*, "Deteksi Dini Penyakit Diabetes Melitus," *Magistrorum Sch. J. Pengabdi. Masy.*, vol. 02, no. 02, pp. 300–309, 2021.

[3] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, "Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 11, no. 2, pp. 88–96, 2022.

[4] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of Diabetes using Machine Learning Algorithms," *Mater. Today Proc.*, no. 80, 2021, doi: 10.1016/j.matpr.2021.07.196.

[5] R. Sofiana, "Sistem Deteksi Dini Diabetes Mellitus menggunakan Jaringan Saraf Tiruan Backpropagation Dengan Optimasi Adaptive Learning Rate dan Momentum," pp. 1–10, 2016.

[6] P. Permatasari and N. Fajrin, "Perbedaan Pengetahuan dan Sikap Deteksi Dini Diabetes Melitus Sebelum dan sesudah diberikan Promosi Kesehatan di Wilayah Kerja Puskesmas Pasar Rebo," *J. Ilm. Kesehat. Masy. Media Komun. Komunitas Kesehat. Masy.*, vol. 12, no. 2, pp. 56–61, 2020, doi: 10.52022/jikm.v12i2.61.

[7] N. L. P. S. A. Pancawati and D. Santi, "Pengaruh Pendidikan Kesehatan Terhadap Pengetahuan Deteksi Dini DM pada Masyarakat di Pedukuhan Ngemplak Karang Jati Kelurahan Sinduadi Mlati Sleman Yogyakarta," *J. Keperawatan Respati*, vol. 3, no. 1, pp. 24–34, 2016, [Online]. Available: http://nursingjurnal.respati.ac.id/index.php/JKRY/article/view/171

[8] M. A. Nur, "Pendekatan Teknik Data Mining Pada Pusat Data Kesehatan Nasional Menggunakan Map Visualization," *J. IT Media Inf. STMIK Handayani Makassar*, vol. 14, 2016.

[9] I. W. Gamadarenda and I. Waspada, "Implementasi Data Mining untuk Deteksi Penyakit Ginjal Kronis (PGK) menggunakan K-Nearest Neighbor (KNN) dengan Backward Elimination," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 2, p. 417, 2020, doi: 10.25126/jtiik.2020721896.

[10] P. Sinha and P. Sinha, "Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM," *Int. J. Eng. Res.*, vol. V4, no. 12, pp. 608–612, 2015, doi: 10.17577/ijertv4is120622.

[11] I. M. F. Actor *et al.*, "I j m e r," vol. 514, no. 2, pp. 36–39, 2021.

[12] M. Farrel Nur Rilwanu, H. Taufikurachman, dan Faris Huwaidi, R. Perangkat Lunak, and K. Daerah Cibiru, "Penerapan Algoritma K-Nearest Neighbor untuk Mendeteksi Diabetes Berbasis Web Application," *J. Softw. Eng. Inf. Commun. Technol.*, vol. 3, no. 1, pp. 145–152, 2022.

[13] A. Zainuddin, "Implementasi Metode K-Nearest Neighbor Untuk Klasifikasi Penduduk Miskin Di Desa Ngemplak Kidul Kabupaten Pati Jawa Tengah," *J. Inform. SIMANTIK*, vol. 4, no. 1, pp. 21–28, 2019, [Online]. Available: www.jurnal.stmikcikarang.ac.id

[14] F. Yunita, "Sistem Klasifikasi Penyakit Diabetes Mellitus Menggunakan Metode K-Nearest Neighbor (K-NN)," *Bappeda*, vol. 2, no. 1, pp. 223–230, 2016.

[15] A. Asmarani, M. R. Wijaya, E. Rasywir, D. Meisak, and ..., "Implementasi Algoritma K-Nearst Neighbor Untuk Memprediksi Penyakit Diabetes," *J. Inform. Dan ...*, vol. 2, no. September, pp. 231–239, 2022,

[16] G. Mahalisa and N. Arminarahmah, "Diabetes Classification Analysis Using the Euclidean Distance Method Based on the K-Nearest Neighbors Algorithm," *J. Teknol. Komput. dan Sist. Inf.*, vol. 5, no. 3, pp. 178–182, 2022.

[17] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 160–165, 2018, doi: 10.33096/ilkom.v10i2.303.160-165.

[18] A. Petersmann *et al.*, "Definition, classification and diagnostics of diabetes mellitus," *J. Lab. Med.*, vol. 42, no. 3, pp. 73–79, 2018, doi: 10.1515/labmed-2018-0016.

[19] W. C. Zheng and X. X. Wu, "Investigations of the spin Hamiltonian parameters for the trigonal Co 2+ center in ZnS0.001Se0.999 mixed crystal," *Opt. Mater. (Amst).*, vol. 28, no. 4, pp. 370–373, 2006, doi: 10.1016/j.optmat.2004.12.020.

[20] Hasanuddin, "Perbandingan Algoritma KNN dan KNN-PSO untuk Klasifikasi Tingkat Pengetahuan Ibu dalam Pemberian ASI Eksklusif," *Technol. J. Ilm.*, vol. 7, no. 1, pp. 34–40, 2016.

[21] D. Kurniawan and A. Saputra, "Penerapan K-Nearest Neighbour dalam Penerimaan Peserta Didik dengan Sistem Zonasi," *J. Sist. Inf. Bisnis*, vol. 9, no. 2, p. 212, 2019, doi: 10.21456/vol9iss2pp212-219.