Bit-Tech (Binary Digital - Technology)

Vol.6, No.2, December 2023
Available online at: http://jurnal.kdi.or.id/index.php/bt

Design of Diabetes Prediction Application Using K-Nearest Neighbor Algorithm

Alvin Gunawan[1)*], Indah Fenriana[2)]

[1)2)]*Buddhi Dharma University*
Imam Bonjol, Tangerang, Indonesia
[1)]alvingunawantkj1@gmail.com
[2)]indah.f88@gmail.com

Article history:

Received 20 Sept 2023;
Revised 30 Sept 2023;
Accepted 14 Oct 2023;
Available online 28 Dec 2023

Keywords:

Algorithm
Data Mining
Diabetes Prediction
Euclidian Distance
K-Nearest Neighbor

Abstract

The development of diabetes continues to increase accompanied by an increase in unhealthy lifestyles with a high number of cases, making diabetes need to be continuously researched and developed to obtain useful information in terms of research related to diabetes. This study aims to predict diabetes using the K-Nearest Neighbor Algorithm and make a simulation of checking the disease and test the quality of the K-Nearest Neighbor Algorithm for diabetes and make comparisons with the Naïve Bayes algorithm. The K-Nearest Neighbor algorithm is the method used in this study because it has the advantage of being able to train data that is fast, simple, and easy to learn. The way this algorithm works is by calculating the distance between each row of training data and test data based on a predetermined K value. In the process of using the K-Nearest Neighbor, there is a Z-Score normalization stage which is used to adjust the values for each attribute of diabetes symptoms so that they have a range of values that are not too far away. Based on the results of the research and testing of the K-Nearest Neighbor that has been carried out, an accuracy of 97.12% is obtained and the Area Under Curve value is 0.872 which is included in the good classification category and these results have a greater accuracy value compared to previous studies on the same disease, namely Diabetes with the Naïve Bayes algorithm which produces the most optimal accuracy of 87.69%.

I. INTRODUCTION

As time progresses, there are more and more innovations that make human life easier in all aspects of life. This rapid development often makes a person's lifestyle start to become unhealthy and can cause various health problems. Health is an important component in human life and greatly influences all aspects of activities. One disease that can arise due to an unhealthy lifestyle is diabetes. Diabetes is a condition when sugar levels rise and there is a lack of insulin produced by the body's pancreas to regulate blood sugar levels[1]. The number of cases of death due to diabetes is 1.5 million deaths to 2.2 million deaths if accompanied by other symptoms related to diabetes[2]. According to the International Diabetes Federation, the number of diabetes sufferers could reach 783.7 million people in the world in 2045. This number is up 46% compared to the number of sufferers in 2021 which will reach 536.6 million[3]. This shows that diabetes continues to into a dangerous condition that can cause many deaths and needs to be studied further. Diabetes is a chronic disease with symptoms that appear slowly and are mild so that sufferers will feel healthy and will not increase their knowledge about diabetes independently[4].

Seeing that the number of cases of the disease continues to increase, diabetes is a dangerous disease and requires a system that can help in developing early detection of diabetes. Much research related to diabetes has been developed and one of them is research on early symptoms of diabetes using the Naïve Bayes algorithm carried out by Hosea Adrianus and Desiyanna Lasut which produced an accuracy value for diabetes prediction of 87.69% [5]. However, in this study no comparison was made with other algorithms to compare the best algorithm in predicting diabetes. So a comparison is needed to see whether Naïve Bayes is the best algorithm or whether there is a better algorithm for detecting early symptoms of diabetes. From this problem, data mining techniques are applied to help predict diabetes.

Data mining is the process of taking and processing existing data to obtain important information so that it can be used for various purposes related to the data that has been processed. In this research, the K-Nearest Neighbor

* Corresponding author

ISSN 2622-2728 (online) 2622-271X (print)
doi: 10.32877/bt.v6i2.939

Algorithm was used to compare the resulting accuracy with the Naïve Bayes Algorithm which was carried out in previous research.

K-Nearest Neighbor was chosen considering the existence of research on the Classification of Naïve Bayes and K-Nearest Neighbor Algorithms for Diabetes Patients, it shows that K-Nearest Neighbor is better than Naive Bayes[6]. So based on previous research K-Nearest Neighbor was chosen in this study.

II. Related Works/Literature Review

This research takes inspiration from various studies that have been carried out by several previous researchers who studied the K-Nearest Algorithm and who carried out comparisons, such as research by Annisa Nurba Iffah'da dan and Anita Desiani with the title "Implementation of the K-Nearest Neighbor Algorithm (K-NN) and Single Layer Perceptron (SLP) in Predicting Primary Biliary Cirrhosis". This study conducted trials between the 2 algorithms and concluded that both algorithms can be used well in primary biliary cirrhosis but the k-nearest neighbor algorithm shows different results. is better than the Single Layer Perceptron (SLP) algorithm which is interesting to try to apply to diabetes [7].

Then in research with a different algorithm carried out by Norma Ningsih, Aprianto, and Angeline entitled "Data Science Approach for Early Detection of Diabetes Using Naive Bayes Classifier" This research made predictions using the naive Bayes algorithm and added the Laplacian smoothing method for early diabetes symptoms which resulted in the conclusion that the algorithm could be applied to diabetes and showed an accuracy value of 70%[8].

In research conducted by Taufiq, Erfan Jasmin, Cucut Susanto, Komang Aryasa with the title "Expert System for Predicting Diabetes Using the Android-Based K-NN Method". This research used data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey that they used and had cleaned, and concluded that the K-Nearest Neighbor algorithm could be applied to diabetes prediction activities with quite good performance with an accuracy value of 72%[9].

In research conducted by Muhammad Naja Maskuri, Harliana, Kadek Sukerti, and R.M Herdian Bhakti with the title "Application of the K-Nearest Neighbor (KNN) Algorithm to Predict Stroke" the accuracy value of the algorithm was 95% when using K=9 which is the most optimal value and conclude that K-Nearest Neighbor has a good level of accuracy in carrying out these prediction activities[10].

III. Methods

A. Data Mining

Data Mining is an analytical technique that relies on technological capabilities in processing data and using statistical methodology to determine patterns and relationships in data[11]. Data Mining is a series of processes related to searching for patterns in data which functions to explore added value from data using pattern recognition technology, statistics and mathematics[12]. Data Mining will help the process of processing data so that the information produced can be maximized. When processing data using Data Mining, there are various types of algorithms. The type of algorithm used in this research is a classification algorithm because the information that will be sought is in the form of predictions of label events based on the attribute columns that influence them.

B. Algorithm

Algorithm is the activity of taking a collection or several input values which are referred to as input which are then processed into output using a sequence of computational steps[13].

C. Classification

Classification is the grouping of several parameters into one of the previously determined categories [14]. Classification is a method that involves investigating information to reveal a model that describes the classes it contains[15]. The classification algorithm used is the K-Nearest Neighbor algorithm.

D. CRISP-DM

CRISP-DM is a method that is often used by experts to overcome problems in data development[16]. This research process follows the six stages of the CRISP-DM model, with the following stages:

1) *Business Understanding*

   This stage determines the research objectives to be achieved to predict diabetes and compares the algorithms between K-Nearest Neighbor and Naïve Bayes. which aims to find out a better algorithm to apply to this disease by comparing the level of accuracy that has been produced by previous research using the Naive Bayes algorithm, with the accuracy in this study using the K-Nearest Neighbor algorithm.

2) *Data Understanding*

   The dataset used in conducting this research is a dataset taken through a secondary data provider site called Kaggle, with the dataset name being Early Stage Diabetes Risk Prediction Dataset. With complete sources

are https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset. This dataset source collected data using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. The dataset consists of 520 records consisting of 17 attributes.

TABLE 1
DATASET DESCRIPTION

Attribute	Description	Attribute Type
Age	The ages of the patients ranged from 16 to 90 years	Binomial
Gender	Gender of the patient	Binomial
Polyuria	A condition where you urinate more frequently	Binomial
Polydipsia	Conditions when experiencing excessive thirst	Binomial
Sudden Weight Loss	Accidental weight loss without changing your diet or routine	Binomial
Weakness	The body is weaker than usual	Binomial
Polyphagia	Feeling excessively hungry and getting bigger	Binomial
Genital Thrush	Fungal infections of the genitals, pain, burning, swelling in the genital area	Binomial
Visual Blurring	Blurry or blurry vision	Binomial
Itching	The itching sensation in the skin area is more than usual	Binomial
Irritability	A person's condition becomes more emotional (angry, anxious, impatient)	Binomial
Delayed Healing	Slower wound healing	Binomial
Partial Paresis	Nerve disorders that make it difficult to move certain parts of the body	Binomial
Muscle Stiffness	Body condition becomes stiff or uncomfortable	Binomial
Alopecia	Condition Hair loss or baldness on the skin	Binomial
Obecity	Excess fat accumulation	Binomial
Class	Indicators that someone has diabetes	Label

3) Data Preparation

At the data preparation stage, we carry out the process of ensuring that the dataset we will use in the analysis is in a condition that is ready to be used for modeling. Activities carried out in this stage include:

a) Initial Data Precheck

In this process, an initial check is carried out on the dataset that we obtain to ensure that the data is in the correct format and is not damaged.

b) Data Selection

in this process decided to use all 17 attributes in this dataset because they are all relevant to our analysis. Therefore, no additional feature selection was carried out.

TABLE 2
THE SELECTED DATASET

Age	Gender	Polyuria	Polydipsia	Sudden Weight Loss	Weakness	Polyphagia	…	Class
16	Male	Yes	No	Yes	No	Yes	…	Positive
25	Female	No	No	No	Yes	Yes	…	Positive
25	Male	Yes	Yes	No	No	Yes	…	Positive
26	Male	No	No	No	No	No	…	Negative
27	Male	No	No	No	No	No	…	Negative
…	…	…	…	…	…	…	…	…
90	Male	No	Yes	Yes	No	No	…	Positive

Based on table 2, all attributes in the data have been selected, the data used is 17 attributes with a total of 520 data. In this case, the dataset we obtained was in clean condition without any missing values or other data problems, so there is no need to carry out the data cleaning process and can proceed to the modeling stage.

4) Modelling

This stage carries out the modeling process with a predetermined algorithm, namely the K-Nearest Neighbor algorithm in the Rapidminer application, such as the retrieve operator to retrieve the dataset and import it into Rapidminer, setting roles to determine labels, normalization to carry out normalization, cross validation which includes K-Nearest Neighbor.

5) Evaluation

At this stage, we carry out an evaluation process of the algorithm used, namely K-Nearest Neighbor, which has gone through a training process and compares it with Naïve Bayes in previous research from the Confusion Matrix value and the Area Under Curve value.

6) Deployment

This stage applies the prediction model that has been created to a website and creates a feature to enter data on diabetes symptoms experienced and then processed to provide prediction results according to the symptoms that have been entered.

E. K-Nearest Neighbor

The K-Nearest Neighbor algorithm is a supervised learning data mining technique used in the classification process in training datasets with nearest neighbor distances[17]. K-Nearest Neighbor has the advantage that the algorithm is simple so it is easy to implement, and has the disadvantage that it requires determining the appropriate K parameters to provide maximum results. And here is the formula for determining the Euclidian distance in the K-Nearest Neighbor Algorithm to calculate the nearest neighbor distance in the data.

$$Euclidian\ Distance = \sqrt{\sum_{i=0}^{n}(x_i - x_i)^2 + (y_i - y_i)^2 + \cdots} \quad (1)$$

Information:

$(x_i - x_i)^2$ = distance from x_i point

$(y_i - y_i)^2$ = distance from y_i point

IV. RESULTS

A. K-Nearest Neighbor

The K-Nearest Neighbor algorithm is an algorithm that functions to assist in carrying out the prediction process to find a particular class. The K-Nearest Neighbor algorithm works by calculating the distance between the training data and the test data. In K-Nearest Neighbor without normalization, it is enough to directly calculate the distance of each existing attribute from the training data to the training data, and the existing distance results are then searched for the closest distance, namely the smallest Euclidian distance. When doing K-Nearest Neighbor with normalization, you need to change the existing attributes into Z-score values for each attribute first, then calculate the distance.

The following is a display of the flow in the Rapidminer application and the operators used. The testing process was carried out using 520 diabetes patient data with 320 diabetes positive patient data and 200 diabetes negative patient data.

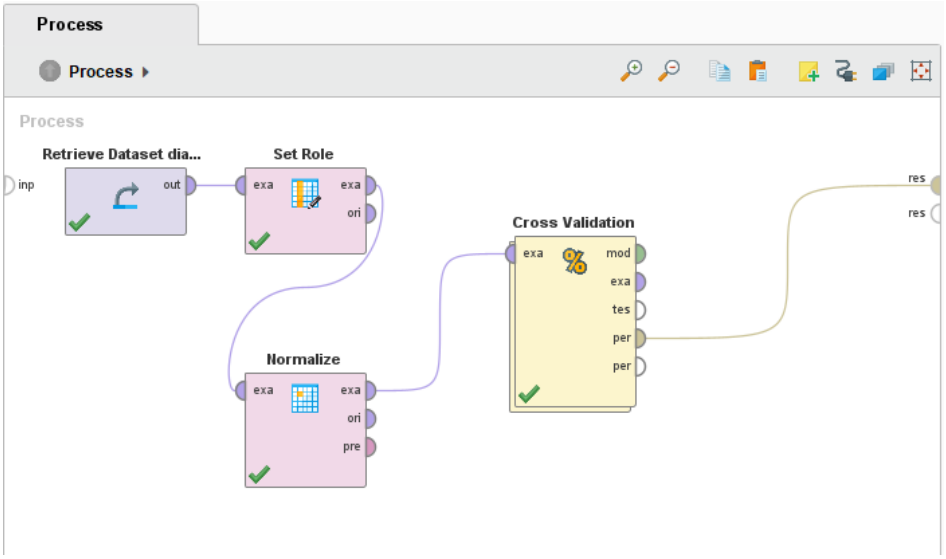

Fig. 3 Operator Display in the Rapidminer application with K-Nearest Neighbor

From Figure 3, there is a diabetes dataset retrieve operator which functions to retrieve the existing dataset so that it can be read by Rapidminer. Then there is the Set Role operator to select the class attribute to be the label to be searched for, the normalize operator is used to carry out normalization, in the normalize operator it is carried out in the method section selecting Z-transformation to normalize with Z-Score. The Cross Validation operator

functions to carry out the testing process by dividing training data and test data to obtain maximum accuracy results.

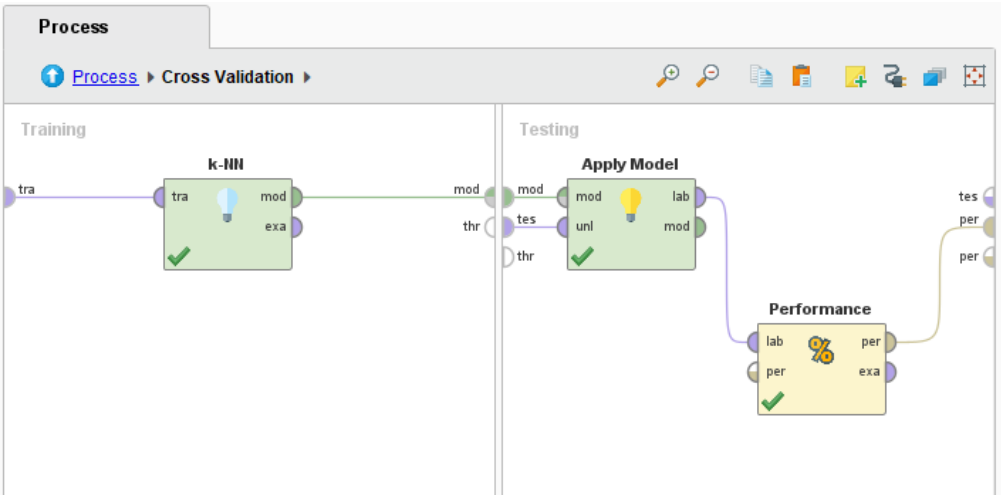

Fig. 4 K-Nearest Neighbor Model

From Figure 4, the K-Nearest Neighbor model is contained in the Cross Validation operator, namely there is the k-NN operator which is the algorithm itself, the apply model which functions to apply the model, and the Performance operator which functions to measure the performance of the model such as the confusion matrix and area under curve. Then the model is implemented on a website page to carry out prediction activities. Prediction activities are carried out using questions related to the symptoms experienced and can be answered with yes or no. Then the number of yes or no fields that have been filled in will be processed using K-Nearest Neighbor and the percentage of someone affected by diabetes will be displayed.

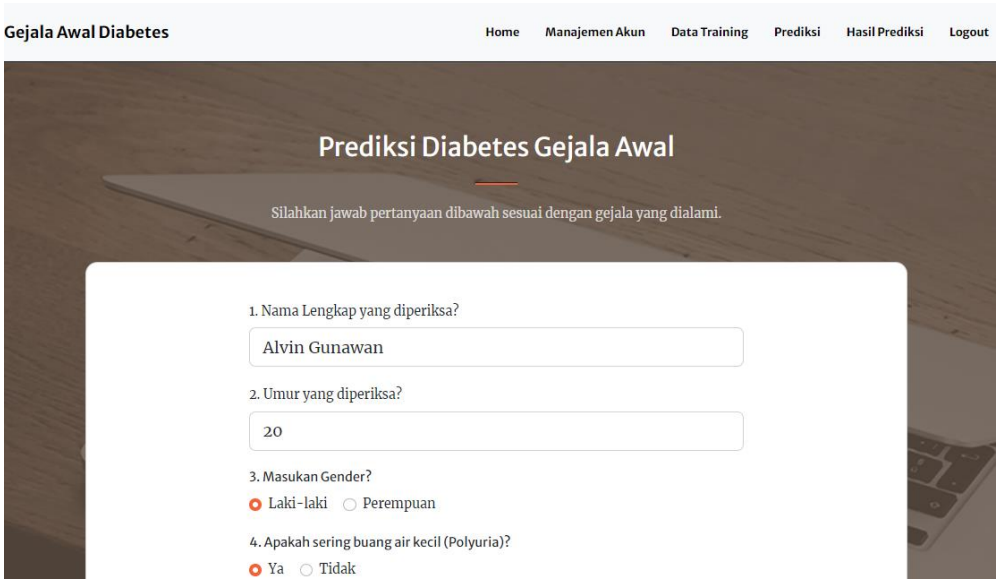

Fig. 5 K-Nearest Neighbor Prediction

From Figure 5 is the implementation of K-Nearest Neighbor in an application to carry out prediction activities. From the picture, fill in the name question column according to your name, then fill in the columns for age, gender, symptoms of polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genitals. mouth ulcers, blurred vision, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia and obesity. In this experiment, we used information about a 20 years old person, male, experiencing symptoms of polyuria and polydipsia, but not experiencing other symptoms.

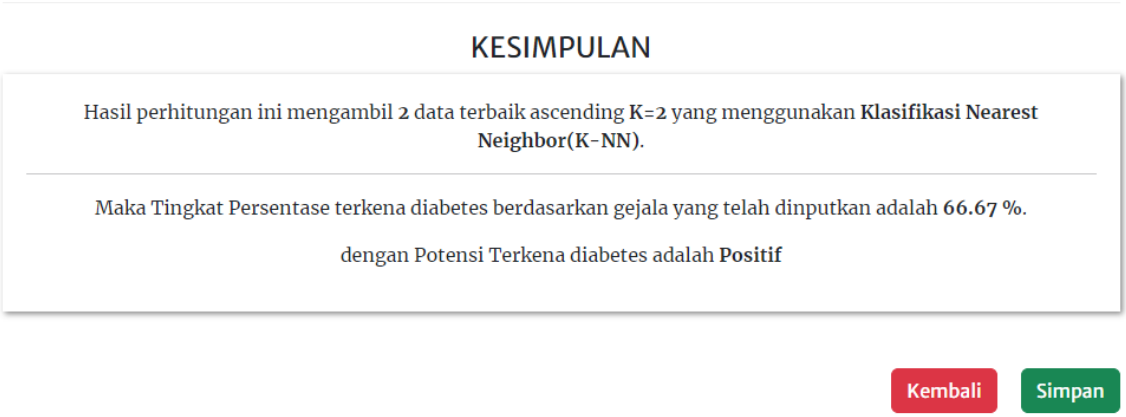

Fig. 6 K-Nearest Neighbor Prediction Result

From Figure 6, the results obtained are based on the answers to the columns regarding diabetes symptoms that were previously filled in, based on the previous information that was filled in: 20 years old, male, experiencing symptoms of polyuria and polydipsia symptoms but not experiencing other symptoms. The results produce a percentage value of 66.67% of a person with positive potential.

B. Confusion Matrix Testing

To test the performance of the K-Nearest Neighbor algorithm, a Confusion Matrix is used. In the Confusion Matrix there are 4 categories, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive (TP) refers to the model predicting the positive class and it is a correct prediction. False Positive (FP) refers to the model predicting the negative class and it is a correct prediction. False Positive (FP) refers to a model predicting a positive class but the actual result is a wrong prediction. False Negative (FN) refers to a model predicting a negative class but the actual result is a wrong prediction and here is the Confusion Matrix display in the K-Nearest Neighbor algorithm by experimenting with various K different Parameters:

TABLE 8
CONFUSION MATRIX TABLE

K	Status	Accuracy
2	Without Normalization	93.65% +/-3.01% (micro average: 93.65)
2	Using Normalization	97.12% +/-2.08% (micro average: 97.12%)
3	Without Normalization	94.04% +/-4.10% (micro average: 94.04%)
3	Using Normalization	96.15% +/-3.27% (micro average: 96.15%)
4	Without Normalization	92.12% +/-4.39% (micro average: 92.12%)
4	Using Normalization	95.96% +/-3.79% (micro average: 95.96%)
5	Without Normalization	89.62% +/-3.87% (micro average: 89.62%)
5	Using Normalization	93.46% +/-4.64% (micro average: 93.46%)
6	Without Normalization	89.62% +/-3.87% (micro average: 89.62%)
6	Using Normalization	93.65% +/-4.71% (micro average: 93.65%)
7	Without Normalization	87.88% +/-5.29% (micro average: 87.88%)
7	Using Normalization	92.31% +/-4.25% (micro average: 92.31%)
8	Without Normalization	86.54% +/-4.71% (micro average: 86.54%)
8	Using Normalization	92.50% +/-3.45% (micro average: 92.50%)
9	Without Normalization	85.19% +/-4.71% (micro average: 85.19%)
9	Using Normalization	90.96% +/-3.15% (micro average: 90.96%)
10	Without Normalization	84.81% +/-4.75% (micro average: 84.81%)
10	Using Normalization	91.35% +/-4.18% (micro average: 91.35%)

From Table 8 is the test to determine the optimal K parameters in the K-Nearest Neighbor Algorithm, from experiments from K = 2 to K = 10, it is known that the value of K = 2 with normalization shows higher accuracy results compared to K other. The following are calculations for the Confusion Matrix with normalization and K = 2 with the following details:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \; x \; 100\ \% \qquad (2)$$

$$= \frac{(307+198)}{(307+13+2+198)} \; x \; 100\ \%$$

$$= \frac{(505)}{(520)} \text{x } 100\ \% \ = \ 0.9712 \text{ x } 100\ \% \ = \ 97.12\ \%$$

From the results of the K-Nearest Neighbor test, an accuracy of 97.12% was obtained and this result had a greater accuracy value compared to previous research conducted by Hosea Adrianus and Desiyanna Lasut on the same disease, namely Diabetes, with the Naïve Bayes algorithm which produced the most optimal accuracy of 87.69 %. So it shows that the K-Nearest Neighbor Algorithm is better to use.

C. Area Under Curve Testing

The AUC or Area Under Curve test is a test that measures the area under the ROC curve, which describes the model's ability to differentiate between positive and negative classes correctly. The higher the AUC value, the better the model is at differentiating between the two classes. The following is the AUC graph:

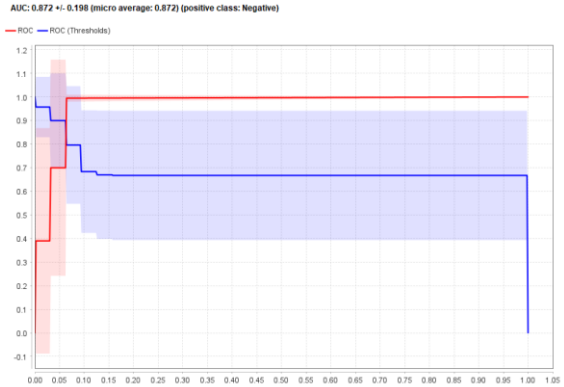

Fig. 7 AUC graph with normalization and K=2

In the test results from figure 7, it is known that the AUC value is 0.872 using the K-Nearest Neighbor algorithm which is included in the good classification category between the range 0.8-0.9 of the resulting AUC value[18].

V. DISCUSSION

This K-Nearest Neighbor algorithm model will further increase accuracy if the number of rows of data obtained increases so that it can produce a Euclidean distance that is not too far between each piece of data. You can use other datasets to compare and see differences in results so that they can be used as a reference in predicting diabetes. When carrying out prediction activities using K-Nearest Neighbor, to get results with good accuracy it is necessary to determine the most optimal K value. During the research process to find the K value, several experiments are needed using different K to find out the best K parameter value. So that it can be used during the prediction process, it would be good if there was a formula that makes it easier to determine the K parameter value so that you don't need to experiment and can find out the best K value faster and better. The dataset currently used is a dataset obtained from a dataset provider so you can also compare it with datasets received directly from hospitals or other health agencies. Algorithm comparison activities other than those carried out in this research with the K-Nearest Neighbor algorithm can also be carried out to compare again in terms of accuracy to find a better model.

VI. CONCLUSIONS

Based on the results of the research that has been carried out, the author concludes that the K-Nearest Neighbor algorithm can be applied and carry out prediction activities well using the secondary dataset "Early Stage Diabetes Risk Prediction" obtained from the dataset provider site called Kaggle. The use of the K-Nearest Neighbor algorithm on this dataset when using the most optimal K parameter value obtained in the confusion matrix test is K=2 which produces an accuracy value of 97.12% and an AUC value of 0.872. With these high results, the K-Nearest Neighbor Algorithm can be implemented well compared to the Naïve Bayes Algorithm to predict diabetes in research on “Designing a Diabetes Prediction Application Using the Naïve Bayes Algorithm” with an accuracy of 87.69%.

REFERENCES

[1] H. Sitorus, V. Yasin, and A. B. Yulianto, "Perancangan sistem pakar diagnosis penyakit diabetes berbasis web menggunakan algoritma naive bayes," *J. Sains dan Teknol. Widyaloka*, vol. 1, pp. 135–144, 2022.

[2] D. A. Agatsa, R. Rismala, and U. N. Wisesty, "Klasifikasi Pasien Pengidap Diabetes Metode Support Vector Machine," *e-Proceeding of Enginering*, vol. 7, no. 1, pp. 2517–2525, 2020, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/11924/11786

[3] L. Pebrianti, F. Aulia, H. Nisa, and K. Saputra, "Informasi Implementasi Metode Adaboost untuk Mengoptimasi Klasifikasi Penyakit Diabetes dengan Algoritma Naïve Bayes," *J. Sist. dan Teknol.*, vol. 7, no. 2, pp. 122–127, 2022, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO/article/view/8627%0Ahttp://jurnal.unmuhjember.ac.id/index.php/JUSTINDO/article/download/8627/4296

[4] R. Oktorina, R. Sitorus, and L. Sukmarini, "Pengaruh Edukasi Kesehatan dengan Self Instructional Module Terhadap Pengetahuan Tentang Diabetes Melitus," *J. Endur.*, vol. 4, no. 1, pp. 171–183, 2019, doi: 10.22216/jen.v4i1.2995.

[5] H. Adrianus and D. Lasut, "Penerapan Data Mining Untuk Memprediksi Penyakit Diabetes Menggunakan Algoritma Naive Bayes," *Algor*, vol. 4, no. 2, pp. 75–85, 2023.

[6] F. Irawan, T. Suprapti, and A. Bahtiar, "KLASIFIKASI ALGORITMA NAIVE BAYES DAN K-NEAREST NEIGHBOR," *J. Tek. Elektro dan Inform.*, vol. 18, no. 1, pp. 80–86, 2023.

[7] A. N. Iffah'da and Anita Desiani, "Implementasi Algoritma K-Nearest Neighbor (K-NN) dan Single Layer Perceptron (SLP) Dalam Prediksi Penyakit Sirosis Biliari Primer," *J. Ilm. Inform.*, vol. 7, no. 1, pp. 65–74, 2022, doi: 10.35316/jimi.v7i1.65-74.

[8] N. Ningsih, Aprianto, and Angeline, "Pendekatan Data Science untuk Deteksi Dini Diabetes Menggunakan Naive Bayes Classifier," *J. Inf. Syst. Graph. Hosp. Technol.*, vol. 05, no. 01, pp. 26–31, 2023.

[9] C. Susanto, T. Taufiq, E. Hasmin, and K. Aryasa, "Sistem Pakar Prediksi Penyakit Diabetes Menggunakan Metode K-NN Berbasis Android," *CogITo Smart J.*, vol. 8, no. 2, pp. 359–370, 2022, doi: 10.31154/cogito.v8i2.406.359-370.

[10] M. N. Maskuri, K. Sukerti, and R. M. H. Bhakti, "Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke," vol. 4, no. 1, pp. 130–140, 2022.

[11] A. K. Hermawan and A. Nugroho, "Analisa Data Mining Untuk Prediksi Penyakit Ginjal Kronik Dengan Algoritma Regresi Linier," *Bull. Inf. Technol.*, vol. 4, no. 1, pp. 37–48, 2023.

[12] G. Setiawan, D. S. D. Putra, and H. Wijaya, "Aplikasi Data Mining Berbasis Web Menggunakan Algoritma Apriori Untuk Analisa Pola Pembelian Barang Pada PT Menara Bahagia Bersama," *Algor*, vol. 3, no. 2, pp. 1–11, 2022, doi: 10.31253/algor.v3i2.1020.

[13] S. N. Sari, R. Kaban, A. Khaliq, and A. Andari, "Sistem Penjadwalan Mata Pelajaran Sekolah Menggunakan Metode Hybrid Artificial Bee Colony (HABC)," *J. Nas. Teknol. Komput.*, vol. 2, no. 1, pp. 20–32, 2022, [Online]. Available: https://publikasi.hawari.id/index.php/jnastek/article/view/21

[14] E. Budiarto, R. Rino, S. Hariyanto, and D. Susilawati, "Penerapan Data Mining Untuk Rekomendasi Beasiswa Pada SD Maria Mediatrix Menggunakan Algoritma C4. 5," *Algor*, vol. 2, 2022, [Online]. Available: https://jurnal.buddhidharma.ac.id/index.php/algor/article/view/1019%0Ahttps://jurnal.buddhidharma.ac.id/index.php/algor/article/download/1019/638

[15] K. Christoper, D. Lasut, and L. W.Kusuma, "Aplikasi Klasifikasi Kepribadian Manusia Menggunakan Algoritma Tree (C4.5) Berbasis Web," *J. Algor*, vol. IV, no. 1, pp. 79–86, 2022.

[16] D. Astuti, "Penentuan Strategi Promosi Usaha Mikro Kecil Dan Menengah (UMKM) Menggunakan Metode CRISP-DM dengan Algoritma K-Means Clustering," *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 1, no. 2, pp. 60–72, 2019, doi: 10.20895/inista.v1i2.71.

[17] A. Syarifah, A. A. Riadi, and A. Susanto, "Klasifikasi Tingkat Kematangan Jambu Bol Berbasis Pengolahan Citra Digital Menggunakan Metode K-Nearest Neighbor," vol. 7, no. 1, pp. 27–35, 2022.

[18] R. Annisa, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," *J. Tek. Inform. Kaputama*, vol. 3, no. 1, pp. 22–28, 2019, [Online]. Available: https://jurnal.kaputama.ac.id/index.php/JTIK/article/view/141/156