Journal of Software Engineering, Information and Communication Technology (SEICT)

Journal homepage: https://ejournal.upi.edu/index.php/SEICT

Implementation of the K-Neighbors Algorithm to Detect Diabetes Web Based Application

*Faris Huwaidi, Hibar Taufikurachman, Mohammad Farrel Nur Rilwanu**

Universitas Pendidikan Indonesia, Indonesia

Correspondence: E-mail: farrelnr@upi.edu*

ABSTRACT

Indonesia is the fifth country with the most diabetes sufferers in the world. This is influenced by an unhealthy lifestyle and then coupled with a lack of public awareness to check whether he has diabetes or not. The KNN (K-Nearest Neighbors) algorithm can be used to pr edict whether a person has diabetes. By using a dataset from the Pima Indian Diabetes Database, the data training process was carried out using the KNN algorithm and obtained decent accuracy results using a Jupyter notebook. From the results of the trained data set, it is then exported to be used in website development using the Python programming language. In the web application developed, the user is asked to input data on pregnancies (a person's pregnancy rate as long as he is alive), insulin levels, glucose levels, BMI, blood pressure, family history of diabetes, skin thickness, and age in the form of a slider. The input data is processed by the KNN algorithm to determine the outcome in the form of a positive or negative diabetes result based on the proximity of the new data entered with other data that has been trained.

ARTICLE INFO

Article History:
Submitted/Received 02 Apr 2022
First Revised 26 Apr 2022
Accepted 12 May 2022
First Available online 25 May 2022
Publication Date 01 Jun 2022

Keyword:
Dataset,
Diabetes,
KNN,
Prediction,
Web Application.

1. INTRODUCTION

Diabetes is a chronic disease which happens when pancreases are no longer able to produce insulin or when the body does not able to utilize produced insulin. Insulin is a hormone produced by the pancreas, which acts as a key to allow glucose from the food we eat to pass through the bloodstream into the cells in the body to produce energy. All carbohydrate-containing foods are broken down into glucose in the blood (Englyst et al, 1996). Insulin helps glucose enter the cells.

Indonesia is one of the top 10 countries with the highest number of diabetes patients in the world (Suryasa et al, 2021). In 1995, Indonesia, which was still considered as a developing country, ranked 7th with a total of 4.5 million diabetes patients (Amos et al, 1997). It is predicted that this ranking will rise to 5th place by 2025, with an estimated number of patients reaching 12.4 million (Guariguata et al, 2014). In 2021, the International Diabetes Federation (IDF) reported that Indonesia has already occupied the 5th position with a total of 19.47 million individuals suffering from diabetes, considering a population of 179.72 million (Paramita et al, 2022). This indicates that the prevalence of diabetes in Indonesia is 10.6% **(see Figure 1)** (Yosmar et al., 2018).

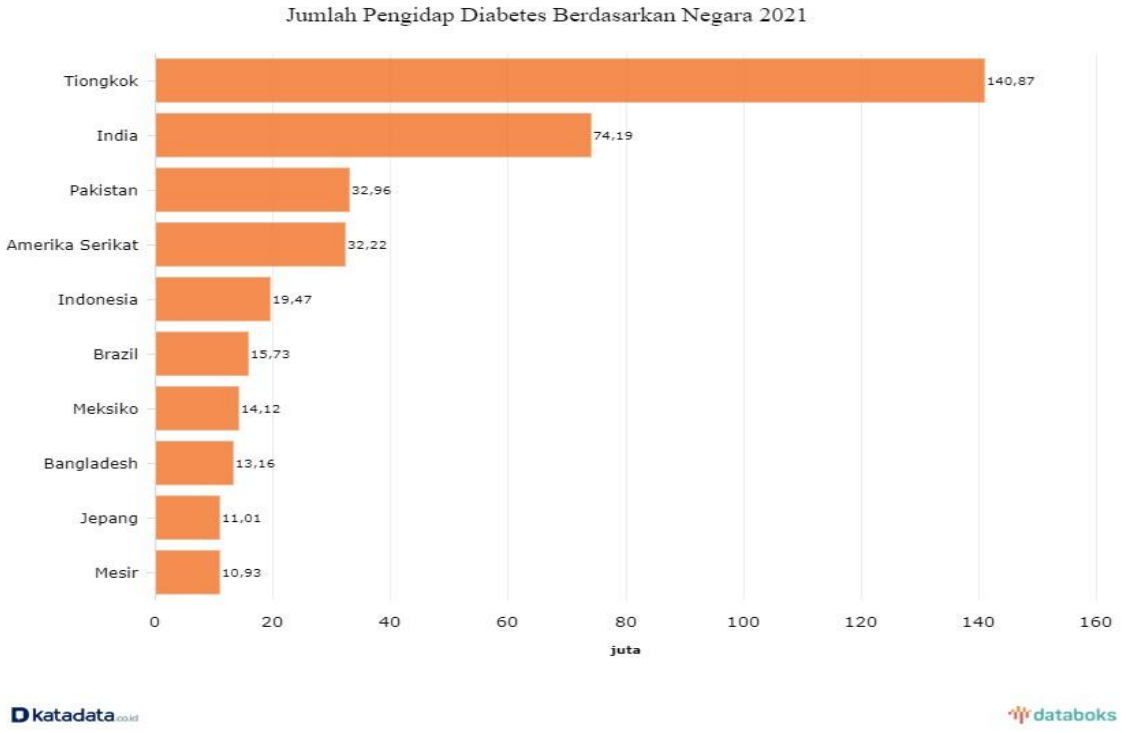

Figure 1. Number of People with Diabetes by Country 2021

In reality, the general population is often unaware of being affected by diabetes (Moore et al, 2000), leading to delayed early diagnosis. Consequently, individuals may fail to maintain a healthy lifestyle and neglect self-care due to their lack of awareness. Additionally, an unhealthy lifestyle (Farhud, 2015) can contribute to the development of diabetes at a young age. As a result, when individuals seek healthcare services, they may already be in a severe condition. Therefore, there is a need for a system that can detect whether someone has been affected by diabetes or not (Luthfa, 2019).

The K-Nearest Neighbour (KNN) algorithm is one of the lazy learning techniques that employs a classification method for objects based on data points that are closest in distance

DOI: https://doi.org/10.17509/seict.v3i1.42347
p- ISSN 2744-1656 e- ISSN 2774-1699

to the object. KNN is also categorized as an instance-based learning approach (Leidiana, 2013).

In the q-dimensional dataset of KNN algorithm (Norouzi et al, 2012), the distance between data points can be calculated. The values of these distances are used as proximity measures between test data and training data. Various methods can be applied in KNN to calculate the proximity of objects with the nearest data, but the most commonly used method is the Euclidean Distance, which is even set as the default in the SKLearn library in the Python programming language when using the KNN algorithm. Other commonly used distance calculation methods in KNN include the Minkowski Distance and the Manhattan Distance (Ooi, dkk., 2013).

The dataset used is the Pima Indians Diabetes Database, which was published by UCI Machine Learning on the Kaggle website (Ul Hassan et al, 2022). We chose this dataset because it has been widely used and has received nearly 3,000 votes from users on Kaggle, which is a global community portal for data science and machine learning.

To detect whether a person has diabetes or not, a diabetes detection system is developed using the K-Nearest Neighbor (KNN) classification algorithm.

2. BASE CONCEPT

2.1. Assesing k Value

Determining the value of k, or the number of neighbors (Zhang dan Song, 2017), is an initial step in constructing a learning model based on the K-Nearest Neighbor algorithm. The value of k represents the number of data points considered when determining the classification for a new input data. For example, if we want to classify a value represented on a Cartesian plane and we specify the value of k as three (three nearest neighbors) (Wilson, 1972), this value will define the three closest neighbors **(see Figure 2)**.

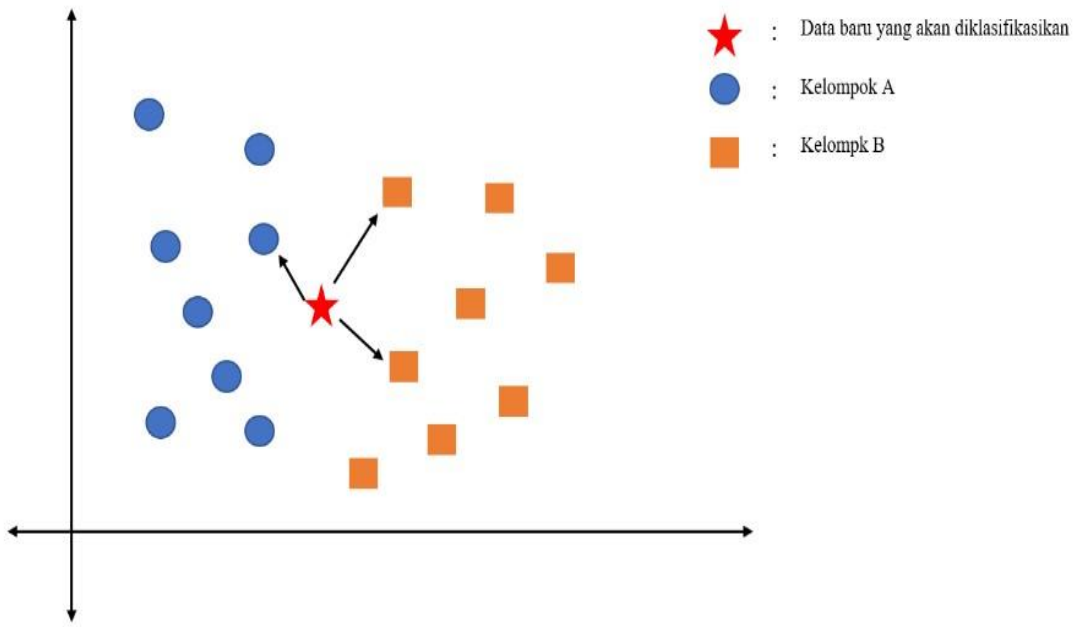

Figure 2. Example algorithm K-NN with 3 objects

DOI: https://doi.org/10.17509/seict.v3i1.42347
p- ISSN 2774-1656 e- ISSN 2774-1699

The new data/object will be classified as Group B because it has a higher number of neighbors in Group B (two neighbors) compared to Group A (one neighbor). If we change the value of k, the determination of the number of neighbors will also slightly change. For example, if we change the value of k to 5 **(see Figure 3)**.

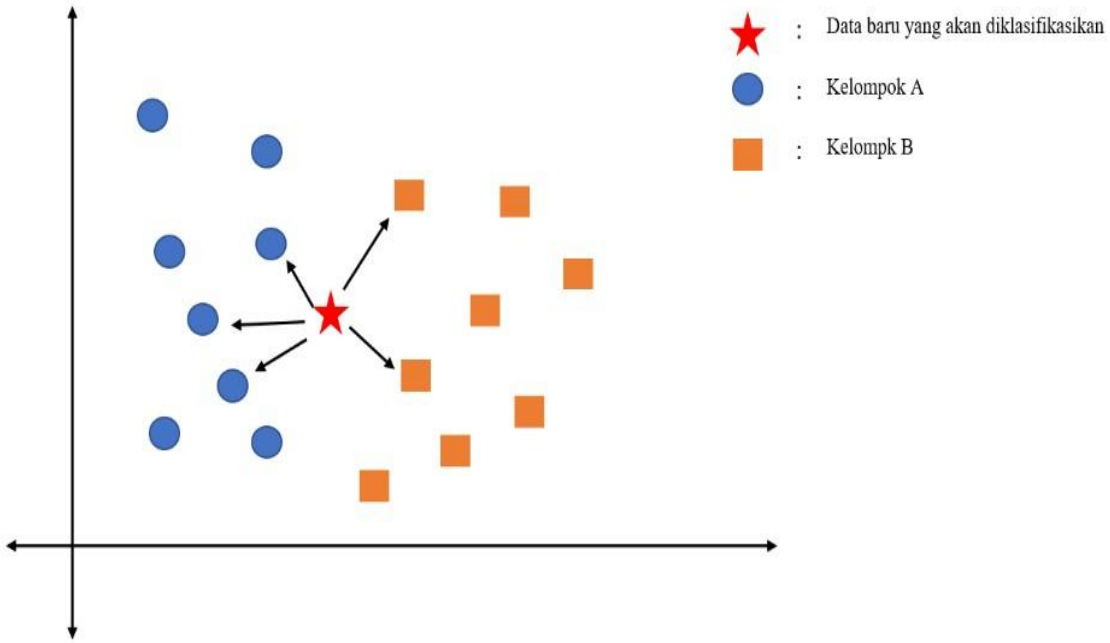

Figure 3. Example algorithm K-NN with 5 objects

The data will be classified as Group A because it has 3 neighbors belonging to Group A, compared to its proximity to only 2 neighbors from Group B. The determination of k or the number of nearest neighbors is indeed an important concept in the K-Nearest Neighbor algorithm.

2.2 Distance Metrics

Previously, it was briefly mentioned about the distance calculation matrix or distance metrics commonly used in the KNN algorithm to measure the proximity of objects to the nearest data, namely Euclidean distance, Minkowski distance, and Manhattan distance. To create a machine learning model with good performance, it is not enough to rely on default parameters or methods alone. With these three different methods of distance calculation matrix, it allows us to obtain the best performing learning model.

2.3 Euclidean Distance

Euclidean distance is a method used to calculate the straight-line distance between two different objects. This method can be applied in 1, 2, and 3-dimensional spaces (Pamungkas, 2019).

DOI: https://doi.org/10.17509/seict.v3i1.42347
p- ISSN 2744-1656 e- ISSN 2774-1699

The calculation of distance in a 1-dimensional space can be illustrated with the following formula.

$$d(x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2}$$

In the implementation of machine learning models, the formula for calculating the Euclidean distance can vary depending on the number of independent variables in the dataset used as training data. If the dataset has two or more independent variables, the dimension of the calculation also increases, and the formula becomes as follows.

$$d = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2 + (y_{1i} - y_{2i})^2 + (z_{1i} - z_{2i})^2 + \ldots}$$

2.4. Manhattan Distance

Manhattan distance is used to select the closest matching case from a case base by calculating the sum of absolute differences between the current case and other cases (Jousselme and Maupin, 2012). The following equation is used to calculate the weight:

$$d_{ij} = \sum W_k \left| x_{ik} - c_{jk} \right|$$

Where dij represents the distance between the i-th and j-th cases with all their parameters. W represents the weight sum. X is the new case subtracted by C, which represents the history (cases present in the Case Base).

2.5. Minkowski Distance

Minkowski distance is a metric in vector space where a norm is defined (normed vector space) and is considered a generalization of both Euclidean distance and Manhattan distance (Mailagaha and Luukka, 2022). In measuring the distance between objects using Minkowski distance, the value of p is typically chosen as either 1 or 2 (Nishom, 2019). The following formula is used to calculate the distance in this method.

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

DOI: https://doi.org/10.17509/seict.v3i1.42347
p- ISSN 2774-1656 e- ISSN 2774-1699

Where.

d = the distance between x and
y x = center data of the cluster
y = data at attribute i for each data
i = each data
n = number of data,
xi = data at the center of the cluster for i
yi = data at each for i
p = power

2.6. Variable Testing

In the given dataset to determine whether a person has diabetes or not, there are 9 variables consisting of 8 independent variables (x) and 1 dependent variable (y). The independent variables (x) are as follows:

a. Pregnancies: the number of times a woman has been pregnant in her lifetime.
b. Insulin: the level of insulin in a 2-hour serum insulin test, measured in units per milliliter (mu U/ml).
c. Glucose: the blood sugar level in a 2-hour glucose tolerance test.
d. BMI: the body mass index, calculated as weight in kilograms divided by height in meters squared.
e. Blood Pressure: the blood pressure measured in millimeters of mercury (mmHg).
f. Diabetes Pedigree Function: an indicator of the genetic history of diabetes in the family.
g. Skin Thickness: a measurement used to estimate body fat, taken on the right half of the forearm, between the olecranon process of the elbow and the acromial process.
h. Age: the age of a sample.

The dependent variable (y) represents the output of the prediction, with a class variable of 0 or 1. 0 indicates no diabetes, while 1 indicates a positive diagnosis of diabetes.

3. METHOD

The research method used is Research and Development (R&D), which aims to develop a product and then test and refine it until the product can function as intended **(see Figure 4)** (Haryati, 2012).

DOI: https://doi.org/10.17509/seict.v3i1.42347
p- ISSN 2744-1656 e- ISSN 2774-1699

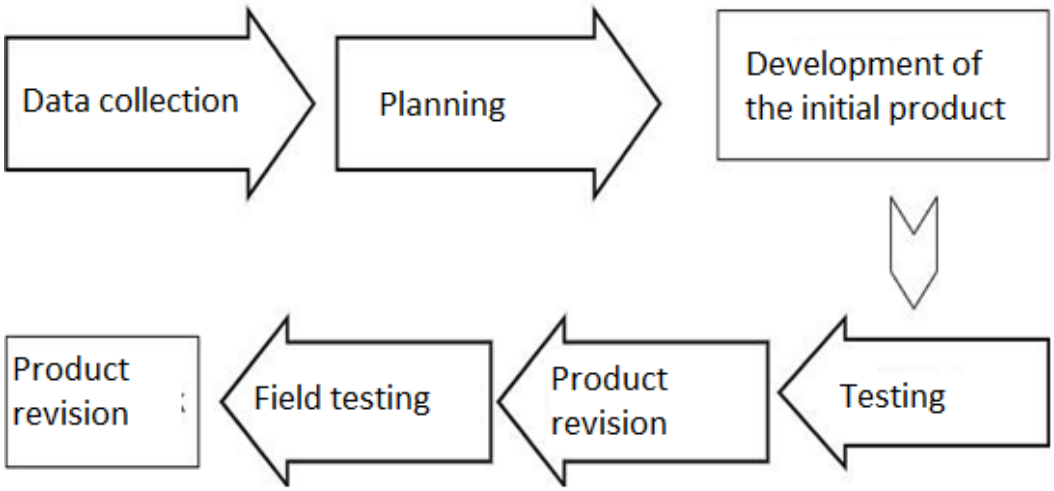

Figure 4. Research process flowchart.

(See Figure 4) it can be seen the stages carried out in the research are as follows..

3.1. Data Collecting

In this stage, several steps are carried out, including gathering information related to diabetes, the K-Nearest Neighbor algorithm and its implementation, as well as searching for a suitable dataset that is relevant to the case and appropriate for training the machine learning model and its implementation. Additionally, this stage involves analyzing the necessary tools required for the development of a web application as the implementation of the machine learning model with the K-Nearest Neighbor algorithm

3.2. Planning

In the planning stage, the product design is created, taking into account important aspects. The primary goal of the product design is to predict whether a person has diabetes or not. Additionally, the target users are identified as adults above 20 years old. The data obtained from the previous stage will be processed to inform decision-making regarding the selection of tools for developing this application. Furthermore, consideration will be given to selecting a programming language that is compatible with the chosen tools and the specific case.

3.3. Prototype Development

The prototype development begins by training and building a machine learning model based on the dataset used, following several stages: importing libraries and the dataset, exploratory data analysis, data pre-processing, splitting, and modelling. Subsequently, a web application is designed using pre-existing templates and the Python programming language. The trained data model using the KNN algorithm is then exported into the web application, enabling it to generate output based on user input.

3.4. Initial Product Development

The initial product development begins by training and building a machine learning model based on the dataset used, going through various stages including importing libraries and dataset, exploratory data analysis, data pre-processing, splitting, and modelling. Afterward, the web application design is created using pre-existing templates and the Python

DOI: https://doi.org/10.17509/seict.v3i1.42347
p- ISSN 2774-1656 e- ISSN 2774-1699

programming language. The trained data results using the KNN algorithm are then exported to the web application, enabling the web to generate outputs based on user inputs.

3.5 Testing

Testing is divided into two parts. The first part focuses on building and developing the machine learning model. This testing involves assessing the model's performance, considering aspects such as model accuracy or score, the processed and organized dataset, and the parameters used in the KNN algorithm. By considering these three aspects, the performance and effectiveness of the machine learning model, which is the core of the product development, can be maximized.

The second part of testing involves the web application development process and deploying the product through a hosting service like Heroku. The aspects considered in this testing include the smoothness of web application development, the success of product hosting, and the ability of the product to make predictions through the hosted web application.

3.6. Revision

Product revision is conducted to enhance the product's quality by improving the performance of the machine learning model. The refinement of the machine learning model involves finding the optimal parameters to achieve a higher score compared to the previous version (model evaluation and hyperparameter tuning). Additionally, revisions are made to the web application development to enhance the UI/UX quality, ensuring that users can easily and comfortably interact with the web application.

3.7. Field Testing

Conducting field testing to predict diabetes diagnosis by inputting data into the form available on the web application and observing the results/output provided by the product.

3.8. Revision Product

Product revision is carried out again based on several considerations, including the analysis of test results, suggestions, and feedback received during the field testing.

4. RESULTS AND DISCUSSION

The application is named according to its main function, which is Diabetes Prediction, and can be accessed at **(see** http://rpl-diabetes-prediction.herokuapp.com**)**.

DOI: https://doi.org/10.17509/seict.v3i1.42347
p- ISSN 2744-1656 e- ISSN 2774-1699

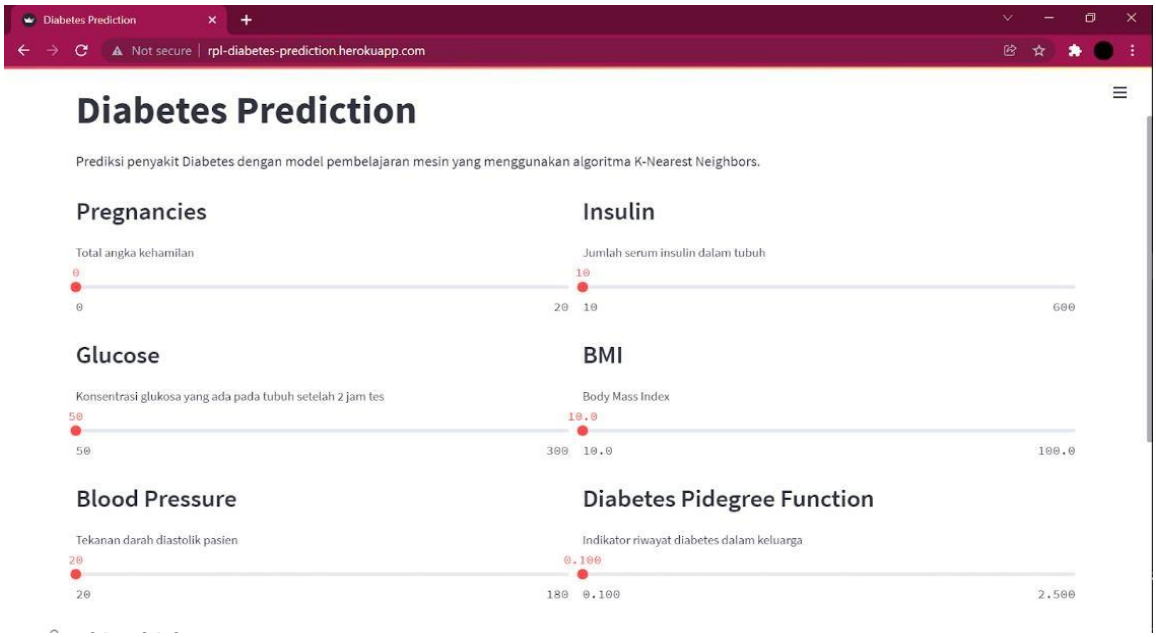

Figure 5. Interface and feature Application Web Diabetes Prediction

The Diabetes Prediction application has reached the finalization stage, as evident from the user interface design **(see Figure 5)**. This application incorporates various features described in the table **(see Figure 5)**. The implementation of the application's usage is now underway.

Table 1. Feature and description application

No.	Feature	Description
1	Diabetes Detection On Main Page	This feature requires several inputs from the user, including the independent variables such as pregnancies (the number of times a person has been pregnant during their lifetime), insulin level, glucose level, BMI, blood pressure, family history of diabetes, skin thickness, and age. After the user inputs all the required information and clicks the prediction button, the input data is analyse using the K-Nearest Neighbor (KNN) algorithm based on its proximity to the other data points in the training dataset. The application then outputs the outcome, indicating whether the user is predicted to be positive or negative for diabetes.
2	Setting	It includes features for configuring the application, which consist of development, appearance (adjusting the width of the display), and theme options to change the visual theme to light, dark, or system-based.
3	Report a Bug	This feature functions to report bugs encountered by users to the developers through GitHub.
4	Get Help	This feature allows users to access help and explanations regarding their needs.

DOI: https://doi.org/10.17509/seict.v3i1.42347
p- ISSN 2774-1656 e- ISSN 2774-1699

4.1. Implementation and Application Usage

After training the machine learning model using the K-Nearest Neighbor algorithm, we can make predictions with the model to assess its accuracy based on the predicted diagnoses (Khorshid dan Abdulazeez, 2021). We can input data containing the independent variables. For example, to test the model's prediction capability, we can input the following data: Pregnancies = 1, Glucose = 202, Blood Pressure = 108, Skin Thickness = 52, Insulin = 131, BMI = 48.5, Diabetes Pidegree Function = 1.114, and Age = 44 **(See Figure 6)**.

```
# Preg: 1, Glu: 202, Bp: 108, SkinT: 52, Insulin: 131, BMI: 48.5, DBF: 1.114, Age: 44
input_data_1 = [[1, 202, 108, 52, 131, 48.5, 1.114, 44]]
```

```
pred = new_knn_model.predict(input_data)
```

```
pred[0]
```

```
1
```

Figure 6. Implementation Using Jupyter Notebook

The prediction result from the model based on the input data is '1' or positive for Diabetes. It's important to note that the diagnosis outcome may vary depending on the input provided. The developed model has also been implemented in a web-based application that is hosted, allowing the machine learning model to be accessed and tested by users **(see Figure 6)**.

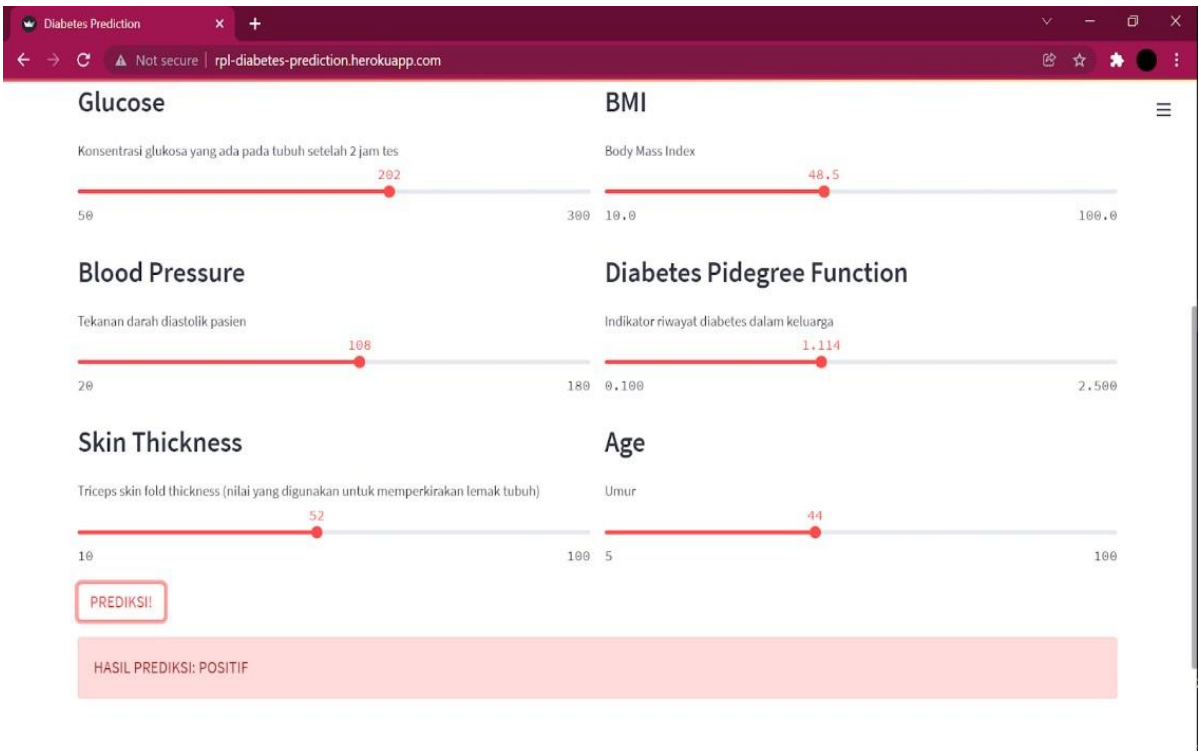

Figure 7. Implementation Using Website Application

The results obtained by inputting data through both Jupyter Notebook and the website application will be the same **(see Figure 7)**.

DOI: https://doi.org/10.17509/seict.v3i1.42347
p- ISSN 2744-1656 e- ISSN 2774-1699

5. CONCLUSION

The Diabetes Prediction application has been successfully developed using the KNN algorithm and the Pima Indian Diabetes Database dataset. This web application is fully functional as intended, where users are prompted to input data such as pregnancies (the number of times a person has been pregnant), insulin level, glucose level, BMI, blood pressure, family history of diabetes, skin thickness, and age using sliders. The input data is processed using the KNN algorithm to determine the Outcome, which indicates whether the result is positive for diabetes or negative based on the similarity of the new input data with the trained data.

6. AUTHORS' NOTE

The authors declare that there are no conflicts of interest regarding the publication of this article. The authors confirm that this paper is free from plagiarism.

7. REFERENCES

Englyst, H. N., Veenstra, J., and Hudson, G. J. (1996). Measurement of rapidly available glucose (RAG) in plant foods: a potential in vitro predictor of the glycaemic response. *British Journal of Nutrition*, *75*(3), 327-337.

Farhud, D. D. (2015). Impact of lifestyle on health. *Iranian journal of public health*, *44*(11), 1442-1444.

Guariguata, L., Whiting, D. R., Hambleton, I., Beagley, J., Linnenkamp, U., and Shaw, J. E. (2014). Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes research and clinical practice*, *103*(2), 137-149.

Haryati, S. (2012). Research and development (R&D) sebagai salah satu model penelitian dalam bidang pendidikan. *Majalah Ilmiah Dinamika, 37*(1), 12-15.

Jousselme, A. L., and Maupin, P. (2012). Distances in evidence theory: Comprehensive survey and generalizations. *International Journal of Approximate Reasoning*, *53*(2), 118-145.

Khorshid, S. F., and Abdulazeez, A. M. (2021). Breast cancer diagnosis based on k-nearest neighbors: a review. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *18*(4), 1927-1951.

Leidiana, H. (2013). Penerapan algoritma k-nearest neighbor untuk penentuan resiko kredit kepemilikan kendaraan bemotor. *PIKSEL: Penelitian Ilmu Komputer Sistem Embedded and Logic, 1*(1), 65-76.

Luthfa, I. (2019). Implementasi selfcare activity penderita diabetes mellitus di wilayah Puskesmas Bangetayu Semarang. *Buletin Penelitian Kesehatan, 47*(1), 23-28.

Amos, A. F., McCarty, D. J., and Zimmet, P. (1997). The rising global burden of diabetes and its complications: estimates and projections to the year 2010. *Diabetic medicine*, *14*(S5), S7-S85.

Mailagaha Kumbure, M., and Luukka, P. (2022). A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance. *Granular Computing*, *7*(3), 657-671.

DOI: https://doi.org/10.17509/seict.v3i1.42347
p- ISSN 2774-1656 e- ISSN 2774-1699

Moore, P. A., Orchard, T., Guggenheimer, J., and Weyant, R. J. (2000). Diabetes and oral health promotion: a survey of disease prevention behaviors. *The Journal of the American Dental Association*, *131*(9), 1333-1341.

Nishom, M. (2019). Perbandingan akurasi euclidean distance, minkowski distance, dan manhattan distance pada algoritma K-Means clustering berbasis Chi-Square. *Jurnal Informatika, 4*(01), 20-24.

Norouzi, M., Fleet, D. J., and Salakhutdinov, R. R. (2012). Hamming distance metric learning. *Advances in neural information processing systems*, *6-8*.

Ooi, H. L., Ng, S. C., and Lim, E. (2013). Ano detection with k-nearest neighbor using minkowski distance. *International Journal of Signal Processing Systems*, *1*(2), 208-211.

Pamungkas, C. A. (2019). Aplikasi penghitung jarak koordinat berdasarkan latitude dan longitude dengan metode euclidean distance dan metode haversine. *Jurnal Informa: Jurnal Penelitian dan Pengabdian Masyarakat, 5*(2), 8-13.

Paramita, W. K., and Pratiwi, Y. M. (2022). Meta-Analysis Effects of Diabetes Mellitus on Mortality in Patients with Chronic Heart Failure. *Journal of Epidemiology and Public Health*, *7*(1), 92-103.

Ul Hassan, I., Ali, R. H., Ul Abideen, Z., Khan, T. A., and Kouatly, R. (2022). Significance of machine learning for detection of malicious websites on an unbalanced dataset. *Digital*, *2*(4), 501-519.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3), 408-421.

Yosmar, R., Almasdy, D., and Rahma, F. (2018). Survei risiko penyakit diabetes melitus terhadap masyarakat Kota Padang. Jurnal sains farmasi and klinis, 5(2), 134-141.

Zhang, X., and Song, Q. (2014). Predicting the number of nearest neighbors for the k-NN classification algorithm. *Intelligent Data Analysis*, *18*(3), 449-464.

DOI: https://doi.org/10.17509/seict.v3i1.42347
p- ISSN 2744-1656 e- ISSN 2774-1699