Journal of Information Technology and Cyber Security 1(2) July 2023: 76-84 EISSN 2987-386X | PISSN 2987-3878
DOI: 10.30996/jitcs.9945

# Prediction of Women's Potential Type 2 Diabetes with Similarity Classifier Based on P-Probabilistic Extension

Ratih Kartika Dewi [1,*] and Shinta Kusuma Wardhani [2]

[1] Department of Informatics Engineering, Universitas Brawijaya, Indonesia
[2] Department of Medicine, Universitas Brawijaya, Indonesia
* Corresponding author: ratihkartikad@ub.ac.id

Received: 4 December 2023
Accepted: 31 December 2023
Revised: 26 December 2023
Available online: 24 January 2024

**To cite this article:** Dewi, R. K., & Wardhani, S. K. (2023). Prediksi Potensi Diabetes Tipe 2 pada Wanita dengan Similarity Classifier Berbasis P-Probabilistic Extension. *Journal of Information Technology and Cyber Security, 1*(2), 75-84. https://doi.org/10.30996/jitcs.9945

## Abstract

Diabetes is a chronic disease that occurs when the pancreas can't produce enough insulin or when the insulin hormone can't be used effectively by the body. The condition of the increased blood sugar, known as hyperglycemia, is a short-term impact that often occurs in uncontrolled diabetes. Meanwhile, the long-term impact of uncontrolled diabetes can cause damage to various body systems, especially blood vessels and nerves. Early detection of diabetes in individuals who are susceptible to diabetes is the main key to control diabetes issues. Artificial intelligence can help this issue. Early diabetes detection with artificial intelligence can predict whether a person in the next 5 years has the potential to suffer from diabetes type 2 or not, based on six variables including 2-hour plasma glucose concentration in the oral glucose tolerance test, diastolic blood pressure, fold thickness triceps, body mass index, diabetes pedigree function, and age. The prediction was built by using similarity classifier based on p-probabilistic extension, trained with the Pima Indian Diabetes dataset with women as research subjects. The contribution of this research is to select representative features in the Pima Indian diabetes dataset then implement them with similarity classifier based on P-Probabilistic Extension. The aim of this study is to compare similarity classifier algorithm with K-nearest neighbor as classifier that widely used in Pima Indian diabetes dataset. The test scenario is carried out by dividing 70% of the training data and 30% of the testing data, then the accuracy for the Pima Indian diabetes data will be compared with K-nearest neighbor and the similarity classifier. Accuracy shows a success value of 75.38%, so the similarity classifier that is built can be used to predict potential diabetes with better performance than K-nearest neighbor.

Keywords: Artificial intelligence, diabetes, P-probabilistic extension, similarity classifier.

## 1. Introduction

Diabetes is a chronic disease that can occur when the concentration of blood sugar in the body increases due to the body's inability to produce the hormone insulin optimally or the insulin hormone in the body cannot be used effectively (Magliano & Boyko, 2021). Diabetes mellitus (DM) is a group of metabolically related diseases characterized by elevated blood sugar conditions that arise due to abnormalities in the production of the hormone insulin or the absorption system of the hormone insulin, which is not optimal (PERKENI, 2021). Diabetes is a chronic disease characterized by three typical symptoms: easy hunger, easy thirst, and easy urination. These three things are caused by increased blood glucose levels and abnormalities in the body's metabolic functions. Untreated diabetes can lead to multiorgan damage, including the heart, kidneys, nerves, and blood vessels, which impairs quality of life and increases mortality caused by diabetes complications (Li et al., 2019). Diabetes is one of the four major types of non-communicable diseases (NCDs) that cause the most significant contribution to morbidity and mortality worldwide. According to data from the World Health Organization (WHO) in 2016, around 422 million people worldwide have diabetes, most of whom live in developing countries, and more than 80% of deaths from diabetes occur in low- and middle-income countries (Mohiuddin, 2019; Suryasa et al., 2021).

The incidence of diabetes is predicted to increase consistently by 600 million people by 2035 (Safitri et al., 2022). The proportion of diabetes in Indonesia amounted to 8.5% or around 20.4 million people diagnosed with diabetes (Kementerian Kesehatan RI, 2018). Based on the East Java Health Profile (Dinkes Jatim, 2022), the estimated number of people with DM in East Java in 2021 is 2.6 million of the Java Island population in the age group of more than 15 years. The most common type of diabetes found in health services is type 2 diabetes, with about 90%-95% of cases of the world's diabetes incidence, with the highest prevalence in middle and low-income countries (World Health Organization, 2019). There are certain types of diabetes due to other causes, such as monogenic diabetes syndrome (i.e., diabetes that occurs in young adulthood or neonatal diabetes), exocrine pancreatic diseases (such as cystic fibrosis and pancreatitis), and chemical or drug-induced diabetes (such as glucocorticoid consumption, patients on HIV/AIDS treatment, or post organ transplant)(American Diabetes Association Professional Practice Committee, 2022).

DM that is not managed correctly can lead to various other types of disease severity. The type of severity of diabetes that often occurs is abnormalities in large blood vessels that can trigger the narrowing of blood vessels due to the accumulation of clumps of residual fat and cholesterol. This severity can lead to an increase in the incidence of other diseases, such as blockage of heart vessels, blockage of cerebral vessels, wounds that are difficult to heal, and tissue damage to the feet of diabetics. Abnormalities of the blood vessel wall and other blood components in blood clotting, red blood cells, and fat metabolism can trigger the onset of other vascular disease activities. Therefore, the level of knowledge regarding DM management is the main focus to avoid the severity of the disease (J, 2019). DM is a significant risk factor for various heart diseases, including heart failure. Epidemiological and observational studies show that DM increases the risk of developing heart failure. The prevalence of heart failure in type 1 diabetes is 8%, while type 2 diabetes is 90%. The inability of the heart to work usually and effectively in type 1 and 2 diabetics is 14.5% and 35%, respectively (Triposkiadis et al., 2021). Long-term complications that often occur in diabetics are diabetic nephropathy, diabetic retinopathy, and neuropathy. This is due to high blood sugar levels in the long term (Alam et al., 2021).

The condition of elevated blood sugar, known as hyperglycemia, is a short-term impact that often occurs in uncontrolled diabetes. Meanwhile, the long-term impact of uncontrolled diabetes is damage to various body systems, especially blood vessels and nerves. Delayed diagnosis of diabetes will exacerbate these effects. Early detection of diabetes in individuals who are prone to diabetes is the key to successful diabetes control. Artificial intelligence (AI) can help with this. Early detection of diabetes built with AI can predict whether a person in the next five years has the potential to suffer from type 2 diabetes or not, based on six variables, namely 2-hour plasma glucose concentration on the oral glucose tolerance test, diastolic blood pressure, triceps thickness, body mass index, diabetes pedigree function, and age. The prediction was built using a similarity classifier based on p-probabilistic extension trained using the Pima Indian Diabetes dataset with women as research subjects, a population with a high risk of diabetes. Research on predicting the potential for diabetes using the Pima Indian diabetes dataset developed using the K-nearest neighbor (KNN) algorithm is also in line with several studies, such as the research of Huwaidi et al. (2022), who made an application to predict the potential for diabetes with KNN using the Pima Indian diabetes dataset. Research by Perdana et al. (2023) stated that Diabetes pedigree function is not a significant variable in KNN. Research by Azizah et al. (2023), which examines the implementation of KNN in Pima Indian diabetes, obtained the optimal k value of 11.

This study aims to compare the KNN-based classifier with the Similarity classifier algorithm for the new data in this study. A similarity classifier based on P-probabilistic extension is one of the classification methods based on similarity measurement used to predict the potential of type 2 diabetes. The similarity classifier that is built produces a positive ideal vector that states an individual has a positive potential for diabetes and a negative vector that states otherwise. The reason for using a similarity classifier based on P-probabilistic extension is that this classifier can learn patterns from the dataset, which then produces positive and negative ideal vectors, making it suitable for predicting potential diabetes, which produces prediction outputs in the form of two binary classes (in this case positive potential diabetes or harmful potential diabetes), so it is expected to outperform the performance of classifiers that are currently widely used for the Pima Indian Diabetes dataset, namely KNN.

The contribution of this research is to select representative features in the Pima Indian diabetes dataset and implement them in the similarity classifier based on P-Probabilistic Extension. The selection of representative features is further explained in the Methods section. The test scenario is done by dividing 70% of the training data and 30% of the testing data, and then the accuracy for KNN and the similarity classifier will be compared. This research is divided into four sections: background in the first section, literature study in the second section, methodology in the third section, and results and discussion in the last section.

## 2. Literature Review

This section contains literature studies on diabetes and similarity classifiers based on P-probabilistic extension, as follows:

### 2.1. Diabetes Mellitus

Diabetes mellitus (DM), or diabetes, is a group of metabolic diseases in which high blood sugar levels exceed normal limits. When blood sugar levels exceed normal limits for an extended period, it can cause symptoms of frequent urination, easy thirst, and easy hunger (Kumar et al., 2020). Diabetes is a chronic disease that can occur when the concentration of blood sugar in the body increases due to the body's inability to produce the hormone insulin optimally or the insulin hormone in the body cannot be used effectively (Magliano & Boyko, 2021). DM is a group of metabolically related diseases characterized by elevated blood sugar conditions arising from abnormal production or synthesis of the hormone insulin, which is not optimal. Based on PERKENI (2021), various symptoms can generally arise in people with DM, namely urine produced in 24 hours increases beyond normal limits or 1500cc every day (polyuria), an unnatural feeling of continuous thirst (polydipsia), accessible feelings of hunger and weakness (polyphagia), weight loss drastically in a few months. Other complaints include body weakness, itching, blurred vision, tingling and abnormal erectile function in men, and redness and itching in the vulva in women (PERKENI, 2021).

### 2.2. Type of Diabetes

According to the American Diabetes Association Professional Practice Committee (2022), diabetes can be categorized into four general groups. Type 1 diabetes can be caused by absolute insulin deficiency due to excessive autoimmune destruction of beta cells. Type 2 diabetes occurs when the body experiences insulin resistance due to impaired secretion of the hormone insulin in pancreatic beta cells. Gestational diabetes is usually diagnosed in pregnant women in the second or third trimester of pregnancy who have no history of diabetes before pregnancy. Certain types of diabetes due to other causes, such as monogenic diabetes syndrome (i.e., diabetes that occurs in young adulthood or neonatal diabetes), exocrine pancreatic diseases (such as cystic fibrosis and pancreatitis), and chemicals or drugs that are triggered by diabetes (such as glucocorticoid consumption, patients under HIV/AIDS treatment, or post organ transplantation)(American Diabetes Association Professional Practice Committee, 2022).

### 2.3. Causes of Diabetes

Risk factors for DM include a combination of genetic, metabolic, and environmental factors that interact with each other to contribute to the incidence of DM. Although individual factors in DM, including non-modifiable risk factors (such as ethnicity and family history/genetic predisposition), have a strong genetic basis, evidence from epidemiological studies suggests that many cases of type 2 DM can be prevented by improving the major modifiable risk factors including obesity, low physical activity, and unhealthy diet (Galicia-Garcia et al., 2020). Unhealthy lifestyles, consumption of high-sugar foods, urbanization, and climate change are global risk factors for type 2 diabetes. The combination of physical inactivity, unhealthy diets, and increased greenhouse emissions due to climate change characterize modern urban society. These factors also increase the risk of type 2 diabetes (Zilbermint, 2020).

### 2.4. Diagnostic criteria

Table 1
Concentration at the time of blood laboratory tests for the diagnosis of diabetes (PERKENI, 2021).

| | HbA1c (%) | Fasting blood sugar (mg/dL) | Blood sugar 2 hours after the Test of Oral Glucose (mg/dL) |
|---|---|---|---|
| Normal | < 5.7 % | <100 | 70-139 |
| Pre-Diabetes | 5.7-6.4 % | 100-125 | 140-199 |
| Diabetes | ≥ 6.5 % | ≥126 | ≥ 200 |

The following is the DM Diagnosis according to the Indonesian Endocrinology Society (PERKENI), which is also illustrated in Table 1 (PERKENI, 2021):

a) Fasting blood sugar ≥ 126 mg/dL at the time of examination. Fasting is a state of the body in which there is no consumption of food or drink for at least 8 hours.
b) Blood sugar 2 hours after the Test of Oral Glucose (TTGO) with a sugar concentration of 75 grams ≥ 200 mg/dl at the time of examination.

c) Current blood sugar ≥ 200 mg/dl with classic symptoms of diabetes or elevated blood sugar ≥ 300 mg/dl.
d) HbA1c value ≥ 6.5% using a standardized method during examination.

The diagnosis of DM can be established when the results of the blood sugar examination meet one of the above criteria. Meanwhile, the DM criteria are classified in the prediabetes group if they do not meet the normal and one of the above criteria.

Criteria for screening for diabetes or prediabetes in asymptomatic adults (American Diabetes Association, 2022; ElSayed et al., 2023) as follows:

1. Screening can be done in overweight or obese adults (body mass index ≥ 25 kg/m2) who have one or more of the following risk factors:
   - Presence of family or offspring who have been diagnosed with diabetes,
   - High-risk races include African Americans, Latinos, Native Americans, Asian Americans, Pacific Islanders,
   - History of heart disease,
   - Hypertension (140/90 mmHg or on therapy for high blood pressure),
   - HDL (High Density Lipoprotein) cholesterol levels <35 mg/dL (0.90 mmol/L) and/or triglyceride levels >250 mg/dL (2.82 mmol/L),
   - Women with a history of polycystic ovary syndrome,
   - Physical inactivity, and
   - Other clinical conditions associated with insulin resistance (such as severe obesity and acanthosis nigricans).
2. Patients with prediabetes who have Hba1c ≥ 5.7% or 39 mmol/mol, impaired glucose tolerance, or impaired fasting glucose should be tested annually.
3. Women with a history of gestational diabetes should still be routinely checked every three years.
4. Patients with a history of other diseases should start early detection of diabetes from age 35.
5. If you have had a blood sugar lab test and the result is typical, the test must be repeated at least within three years.
6. People with HIV.

## 2.5. Diabetes treatment

The principle of management in patients with DM is to eliminate the cause and lower blood sugar so that it can reach typical values. The three goals of DM management are to control the disturbed metabolic system of diabetics to approach normal. The second is to prevent or delay the progression of the disease in the short and long term. The third is to provide patients with information, education, motivation, and means to control their blood sugar independently (Kumar et al., 2020). DM management consists of 4 essential pillars helpful in controlling disease progression and preventing disease severity. Education, dietary or nutritional regulation, physical activity regulation, and medication therapy are four essential pillars to control blood sugar levels (Putra & Berawi, 2015).

Diabetes management is also the same as other chronic diseases, which is influenced by the characteristics of each individual with diabetes. Awareness and vigilance are strongly associated with better self-care for diabetics, especially in terms of controlling blood sugar levels. Consistency in self-care behaviors such as glucose monitoring, physical activity regulation, regular check-ups with the doctor, and a healthy and regular diet are the primary keys to successful therapy. Patients who have abnormalities in the psychiatric part are seen from an unstable emotional state and excessive feelings of worry. Stress, anxiety, excessive fear, and erratic mood, if they occur continuously without being controlled, will essentially interfere with therapy management and adherence and can also contribute to the increased development of comorbidities in the psychiatric or psychiatric field (Bhat et al., 2020).

## 2.6. Similarity classifier based on P-probabilistic extension

Similarity classifier based on P-probabilistic extension introduces a more generalized way for similarity measurement, namely with generalized mean as listed in Eq. (1). Generalized mean combines arithmetic, harmonic, and geometric mean (Luukka, 2009). Generalized mean combines arithmetic, harmonic, and geometric mean (Luukka, 2009). It is used to improve arithmetic, harmonic, and geometric mean shortcomings. For $a_1$,$a_2$,...,$a_n$ and $p$ are non-zero real numbers, and they can be formulated (Raïssouli et al., 2009),

$$Generalized\ mean = \left(\frac{a_1^p+a_2^p+\cdots+a_n^p}{n}\right)^{\frac{1}{p}} \quad (1)$$

If you want to classify, a set of $X$ of an object with $N$ are different classes $C_1, C_2, \ldots, C_n$, if the value of the distance on each characteristic is normalized with min-max normalization such as Eq. (2), then the value is represented with the range [0,1]. $v'$ is the normalized value calculated from the variable value $v$, the smallest value of all variables $v(min_a)$, the greatest value of all variables $v(max_a)$, and $newmax_a$ value is 1 and $newmin_a$ value is 0.

$$v' = \left(\frac{v - min_a}{max_a - min_a}\right) \times (newmax_a - newmin_a) + newmin_a \tag{2}$$

The first thing that can be done is to determine the positive and negative ideal vector values $V_i = (V_i(1), \ldots, V_i(d))$ is shows a class $i$ whereas $d$ that the identification of attribut $f_1, f_2, \ldots, f_d$ is the object measures. This vector can be calculated from several sample sets $X_i$ from $X_1 = (X_i(1), \ldots, X_i(d))$ which is known from the class $C_i$. Eq. (3) is used to calculate $V_i$, where $V_i$ is the ideal vector for the class $i$, $X$ is an object (data), $x$ is an attribute of the object $(X)$, $m$ the value 9,9 is the same value for all $d$ (Luukka, 2009).

$$V_i(d) = \left(\frac{1}{\#X_i} \sum_{x \in X_i} x(d)^m\right)^{\frac{1}{m}} \tag{3}$$

After the ideal vector of each class is determined, the decision of which class to choose is implemented by comparing the values of each vector. The comparison can be made using the previous equation, which is shown in Eq. (3). Therefore, a *similarity classifier* based on P-*probabilistic extension* is able to identify in Eq. (4), where $S(x,y)$ is a similarity between data $x$ and $y$, $x$ is data test, $y$ is an ideal vector (positive or negative value), $D$ is different attribute total, $d$ is a *different* attribute, $W$ is weight, we choose $weight(W)$ 0,1 because this value is the default value used in Luukka research (2009) and $m$ is the same value for all $d$.

$$S(x,y) = \left(\frac{1}{D} \sum_{d=1}^{D} W_d (1 - x(d) - y(d) + 2x(d)y(d))^m\right)^{\frac{1}{m}} \tag{4}$$

The weight value $w_d \in [0,1]$ for Eq. (4) is a set value. We determine that a data $(x)$ is assigned to a class $C_m (x \in C_m)$ if the entered data value is more towards the positive or negative ideal vector, for example, the similarity value of data $X$ with $V_{i+}$ is greater than the similarity value of data $X$ with $V_{i-}$ then the data is classified into the positive class, which is denoted by $S(X, V_i)$ as Eq. (5). If $S(x, v_{i+}) > S(x, v_{i-})$ then the data is predicted to be in a positive class. If $S(x, v_{i+}) < S(x, v_{i-})$, the data is predicted to be in the negative class (Luukka, 2009).

$$S(X, V_m) = \max_{1=1,\ldots,N} S(X, V_i) \tag{5}$$

Similar literature that discusses the prediction of potential diabetes in the next five years is the first study using the Pima Indian Diabetes data set with input or input in the form of Pima Indian Diabetes data containing eight variables in the form of the number of times pregnant, 2-hour oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, blood sugar 2 hours after eating, body mass index, diabetes pedigree function, and age. The output of the study by Smith et al. (1988) is a prediction of the first appearance of diabetes (onset). Prediction is done with the ADAP (Adaptive Algorithm) algorithm based on the Artificial Neural Network (JST) of Smith et al. (1988).

Measuring the success of a similarity classifier can be measured by accuracy testing. The accuracy of a system can be tested with accuracy testing using testing data (Appavu & Rajaram, 2009). The calculation for accuracy testing can be described by weighted accuracy as in Eq. (6) Eq. (6) (Tan et al., 2006), where $w_1$ is a weight $TP$ the value is 1, $w_2$ is weight $FP$ the value is 1, $w_3$ is weight FN with values is 1, $w_4$ is weight $TP$ values is 1, $TP$ is a true positive or positive data prediction and value is true, $FP$ is false positive or data prediction positive with false value, $FN$ is negative (false) or negative data prediction, and $TN$ is a true negative or data prediction with a negative and the value is true.

$$Weighted\ Accuracy = \frac{w_1TP + w_4TN}{w_1TP + w_2FP + w_3FN + w_4TN} \tag{6}$$

## 3. Methods

AI-based diabetes prediction can be built using a classifier trained using the Pima Indian Diabetes Data dataset to obtain patterns of individuals likely to develop type 2 diabetes within the next five years. Pima Indian Diabetes Data is a validated dataset because the Pima Indian population in Arizona, which is a population with a high risk of diabetes, has been studied by the National Institute of Diabetes, Digestive and

Kidney Diseases since 1965. This dataset becomes the training data for the Similarity Classifier based on the P-Probabilistic extension. As for testing the prediction system, the testing data uses data from Indonesia. The built classifier is able to predict whether a woman in the next five years has the potential to suffer from type 2 diabetes or not, based on several variables, namely 2-hour plasma glucose concentration on oral blood sugar tolerance test, diastolic blood pressure, triceps fold thickness, body mass index, diabetes pedigree function, and age.

Algorithm 1. Similarity Classifier based on P-Probabilistic Extension/SCPPE.

```
Declaration:
String glu,dia,tri,bmi,dpf,age ➔ variable to hold the input, in the form of a textfield
Double glu1,dia1,tri1,bmi1,dpf1,age1 ➔ conversion variable for reading textfield (string) to double.
Double nglu,ndia,ntri,nbmi,ndpf,nage ➔ variable for normalization
Double snglu,sndia,sntri,snbmi,sndpf,snage ➔ sn = negative similarity (similarity data with ideal vector -)
Double spglu,spdia,sptri,spbmi,spdpf,spage ➔ sp = positive similarity (similarity data with  ideal vector +)
Double simneg, simpos ➔ variable SCPPE (data,vektor ideal -/+)
Double pre ➔ variables that predict potential diabetes
Algorithm:
Input: glu,dia,tri,bmi,dpf,age.
Process:
1. Reading textfield with data type string to double
2. Data normalization using Eq. (6).
3. Calculation of similarity (data, ideal vector -) stage 1 using Eq. (7).
4. Calculation of similarity (data, ideal vector -) stage 2 using Eq. (3)..
5. Calculation of similarity (data, ideal vector +) stage 1 using Eq. (7).
6. Calculation of similarity (data, ideal vector +) stage 2 using Eq. (3).
7. simpos >= simneg?
    If yes, pre=1.
    If no, pre = 0.
8. pre =1?
    If yes, the result will be:
    Prediction ← The data entered is predicted to be potentially POSITIVE for diabetes.
    Suggestion ← Control your blood sugar, with oral medication & insulin (only if absolutely necessary). Control your blood
    pressure. Take care of the food you eat (reduce consumption of sugar & carbohydrates.
    If no, the result will be:
    Prediction ← The data entered predicts potentially NEGATIVE diabetes.
    Suggestion ← Maintain a healthy lifestyle: ideal body weight, exercise at least 30 minutes a day, healthy diet (consuming
    3-4 vegetables & fruit a day, reducing sugar & fat consumption), don't smoke.
Output: prediction and suggestion
```

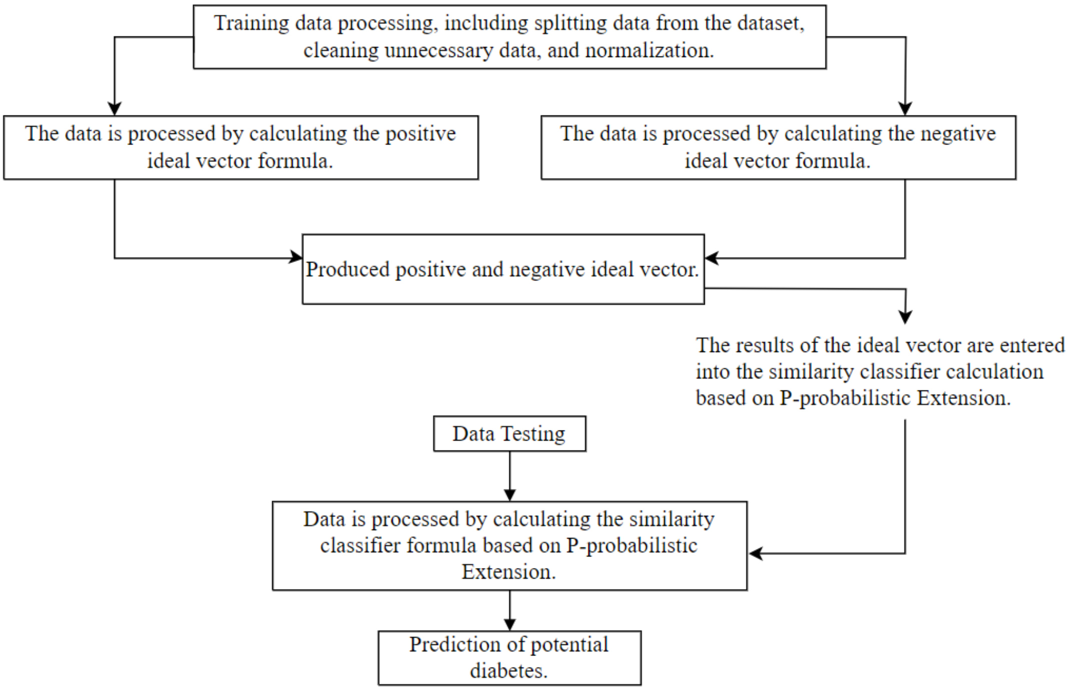
Training data processing, including splitting data from the dataset, cleaning unnecessary data, and normalization.
The data is processed by calculating the positive ideal vector formula.
The data is processed by calculating the negative ideal vector formula.
Produced positive and negative ideal vector.
The results of the ideal vector are entered into the similarity classifier calculation based on P-probabilistic Extension.
Data Testing
Data is processed by calculating the similarity classifier formula based on P-probabilistic Extension.
Prediction of potential diabetes.

Fig. 1. Reseach diagram.

The implementation of the similarity classifier starts from the pseudocode as in Algorithm 1. This algorithm pseudocode consists of 3 parts, namely input, process, and output. The input in this study is individual data with variables of 2-hour plasma glucose concentration on oral blood sugar tolerance test, diastolic blood pressure, triceps fold thickness, body mass index, diabetes pedigree function, and age. The process is data processing with the calculation of the positive ideal vector formula and the negative ideal vector. The output in this study is the result of processing ideal vector calculation data. If the prediction value is equal to one (1), then the data entered is predicted to be potentially positive for diabetes. If the prediction value is equal to zero (0), then the data entered is predicted to be potentially negative for diabetes.

In Fig. 1, the prediction of diabetes potential begins with processing the training data, which is in the form of the Pima Indian Diabetes dataset, and selecting six variables, namely 2-hour plasma glucose concentration on the oral glucose tolerance test, diastolic blood pressure, triceps fold thickness, body mass index, diabetes pedigree function, and age as a pre-processing stage. The next step is to calculate the value of the negative and positive ideal vectors, which shows the reference value for positive or negative predictions of potential diabetes. After getting the positive and negative ideal vector values from the Pima Indian Diabetes dataset, the value is used as a reference for the ideal vector value, which is the central part of the Similarity classifier algorithm based on P-probabilistic extension. Predictions are made by calculating the similarity value of data with positive and negative ideal vectors through a similarity classifier based on P-probabilistic extension. The calculation results in the prediction of diabetes potential, with the decision or prediction result of the class that has a greater similarity value between the data and the positive and negative ideal vectors. The test scenario is carried out by dividing 70% of the training data and 30% of the testing data, and then the accuracy for KNN and the similarity classifier will be compared.

Table 2
Dataset Pima Indian Diabetes.

| No. | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DPF | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 2 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 3 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 4 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 5 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| 6 | 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 7 | 3 | 78 | 50 | 32 | 88 | 31.0 | 0.248 | 26 | 1 |
| 8 | 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 9 | 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |
| 10 | 8 | 125 | 96 | 0 | 0 | 0.0 | 0.232 | 54 | 1 |

Looking at the top 10 Pima Indian Diabetes data as shown in Table 2, 8 variables indicate the potential for type 2 diabetes in women (National Institute of Diabetes & Kidney Diseases, 2017), as follows:

1. Pregnancies state the number of pregnancies or the number of times the woman under study was pregnant,
2. Glucose states the blood sugar concentration 2 hours after eating with an oral glucose tolerance test,
3. Blood Pressure states the diastolic blood pressure at the time the sample was examined,
4. Skin Thickness states the thickness of the triceps skin fold at the time the sample was examined,
5. Insulin states the blood sugar concentration 2 hours after a meal without an oral glucose tolerance test,
6. BMI states the body mass index to determine whether the sample is obese or not,
7. Diabetes Pedigree Function (DPF) states the family history of diabetes. If the DPF value is higher, the higher the history of family members with close genetic proximity who have diabetes, and vice versa. Close genetic proximity, such as a relationship between parents or siblings, and
8. Age states the age at the time of sample data collection.

In this study, the variables used were 2-hour plasma glucose concentration on oral glucose tolerance test (glucose), diastolic blood pressure (Blood Pressure), triceps fold thickness (Skin Thickness), BMI, DPF, and age (Age). The variables pregnancies and insulin were not included in this study because neither variable is urgent in predicting the potential for type 2 diabetes. The pregnancy variable was not used in this study because there were patients who had never been pregnant, and the insulin variable was not used because to diagnose type 2 diabetes, one component of blood sugar 2 hours after eating with an oral glucose tolerance test can be used to predict the potential for diabetes. The authors did not include data with BMI, insulin, Blood pressure, and Glucose values of 0.

## 4. Results and Discussion

The performance of the similarity classifier for predicting diabetes potential can be measured by accuracy testing. The testing procedure is to test the testing data against the similarity classifier, then it will produce prediction results. The prediction results are then matched to the actual data from the class (the name of the variable in Table 2 is outcome). Accuracy shows a success value of 75.38% for the similarity classifier and 74.61% for KNN. There is an increase of 0.77%, this is because the similarity classifier uses an ideal vector value reference to predict the positive or negative potential for diabetes. A description of the accuracy results of Pima Indian Diabetes using KNN and similarity classifier can be seen in Table 3.

Table 3
Comparative algorithms.

| Dataset | KNN | Similarity Classifier |
|---|---|---|
| Pima Indian Diabetes | 74.61% | **75.38%** |

## 5. Conclusions

The similarity classifier based on the P-probabilistic extension that was built is able to predict whether a woman in the next 5 years has the potential to suffer from type two diabetes or not, based on six variables, namely blood sugar two hours after giving an oral glucose tolerance test, diastolic blood pressure, triceps fold thickness, two-hour insulin concentration, body mass index, diabetes pedigree function, and age. The contribution of this research is to select representative features in the Pima Indian diabetes dataset then implement them with similarity classifier based on P-Probabilistic Extension. The aim of this study is to compare similarity classifier algorithm with K-nearest neighbor as classifier that widely used in Pima Indian diabetes dataset. The test scenario is carried out by dividing 70% of the training data and 30% of the testing data, then the accuracy for the Pima Indian diabetes data will be compared with K-nearest neighbor and the similarity classifier. Accuracy shows a success value of 75.38%, so the similarity classifier that is built can be used to predict potential diabetes with better performance than K-nearest neighbor.

In future work, it is possible to test a system for predicting potential diabetes in men because this study uses the Pima Indian Diabetes dataset with women in Arizona as research subjects, where this population has a high risk of diabetes.

## 6. CRediT Authorship Contribution Statement

**Ratih Kartika Dewi:** Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – original draft, and Writing – review & editing. **Shinta Kusuma Wardhani**: Conceptualization, Formal Analysis, Investigation, Resources, Visualization, and Writing – review & editing.

## 7. Data Availability

This study uses an open-source diabetic retinopathy dataset that can be accessed via https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database, an open-source online data repository hosted at Kaggle (www.kaggle.com).

## 8. Funding

No funding was received for this study.

## 9. References

Alam, S., Hasan, M. K., Neaz, S., Hussain, N., Hossain, M. F., & Rahman, T. (2021). Diabetes Mellitus: Insights from Epidemiology, Biochemistry, Risk Factors, Diagnosis, Complications and Comprehensive Management. *Diabetology*, *2*(2), 36–50.

American Diabetes Association. (2022). Standards of Medical Care in Diabetes—2022 Abridged for Primary Care Providers. *Clinical Diabetes*, *40*(1), 10–38.

American Diabetes Association Professional Practice Committee. (2022). Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes — 2022. *Diabetes Care*, *45*(1), 17–38. https://doi.org/https://doi.org/10.2337/dc22-S002

Appavu, S., & Rajaram, R. (2009). Knowledge-based system for text classification using ID6NB algorithm. *Knowledge-Based Systems*, *22*, 1–7. https://doi.org/10.1016/j.knosys.2008.04.006

Azizah, N., Firdaus, M. R., & Indrayatna, R. S. F. (2023). Penerapan Algoritma Klasifikasi K-Nearest Neighborpada Penyakit Diabetes. *Seminar Nasional Statistika Aktuaria 2*.

Bhat, N. A., Muliyala, K. P., & Chaturvedi, S. K. (2020). Psychological Aspects of Diabetes. *Diabetes*, *8*(1), 90–98. https://doi.org/10.33590/emjdiabet/20-00174

Dinkes Jatim. (2022). *Profil Kesehatan 2021*. https://dinkes.jatimprov.go.id/userfile/dokumen/PROFIL

KESEHATAN 2021 JATIM.pdf

ElSayed, N. A., Aleppo, G., Aroda, V. R., Bannuru, R. R., Brown, F. M., Bruemmer, D., Collins, B. S., Gaglia, J. L., Hilliard, M. E., Isaacs, D., Johnson, E. L., Kahan, S., Khunti, K., Leon, J., Lyons, S. K., Perry, M. Lou, Prahalad, P., Pratley, R. E., Seley, J. J., … Association, A. D. (2023). Classification and Diagnosis of Diabetes: Standards of Care in Diabetes—2023. *Diabetes Care*, *46*(1), 19–40. https://doi.org/https://doi.org/10.2337/dc23-S002

Galicia-Garcia, U., Benito-Vicente, A., Jebari, S., Larrea-Sebal, A., Siddiqi, H., Uribe, K. B., Ostolaza, H., & Martín, C. (2020). Pathophysiology of Type 2 Diabetes Mellitus. *International Journal of Molecular Sciences*, *21*(7), 1–34.

Huwaidi, F., Taufikurachman, H., & Rilwanu, M. F. N. (2022). Implementation of the K-Neighbors Algorithm to Detect Diabetes Web Based Application. *Journal of Software Engineering, Information and Communication Technology*, *3*(1), 71–82.

J, A. (2019). Overview of Knowledge Levels of Type 2 Diabetes Mellitus Patients Regarding Diabetes Management. *Media Keperawatan: Politeknik Kesehatan Makassar*, *10*(2), 19–22.

Kementerian Kesehatan RI. (2018). *Hasil Utama Riskesdas 2018*. https://kesmas.kemkes.go.id/assets/upload/dir_519d41d8cd98f00/files/Hasil-riskesdas-2018_1274.pdf

Kumar, R., Saha, P., Kumar, Y., Sahana, S., Dubey, A., & Prakash, O. (2020). A Review on Diabetes Mellitus: Type1 & Type2. *World Journal of Pharmacy and Pharmaceutical Sciences*, *9*(10), 838–850. https://doi.org/10.20959/wjpps202010-17336

Li, S., Wang, J., Zhang, B., Li, X., & Liu, Y. (2019). Diabetes Mellitus and Cause-Specific Mortality: A Population-Based Study. *Diabetes and Metabolism Journal*, *43*, 319–341.

Luukka, P. (2009). Similarity classifier using similarities based on modified probabilistic equivalence relations. *Knowledge-Based Systems*, *22*(January), 57–62. https://doi.org/10.1016/j.knosys.2008.06.005

Magliano, D. J., & Boyko, E. J. (2021). What is diabetes ? In *IDF Diabetes Atlas* (10th ed., pp. 1–18).

Mohiuddin, A. (2019). Diabetes Fact: Bangladesh Perspective. *Community and Public Health Nursing*, *4*(1), 39–47.

National Institute of Diabetes, & Kidney Diseases. (2017). *Pima Indians Diabetes Database*. Kaggle.

Perdana, A., Hermawan, A., & Avianto, D. (2023). Analyze Important Features of PIMA Indian Database For Diabetes Prediction Using KNN. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *12*(1), 70–75.

PERKENI. (2021). Pedoman Pengelolaan dan Pencegahan Diabetes Melitus Tipe 2 di Indonesia 2021. In *PERKENI*.

Putra, I. W. A., & Berawi, K. N. (2015). Four Pillars of Management of Type 2 Diabetes Mellitus Patients. *Majority: Medical Journal of Lampung University*, *4*(9), 8–12.

Raïssouli, M., Leazizi, F., & Chergui, M. (2009). Arithmetic-Geometric-Harmonic Mean of Three Positive Operators. *Journal of Inequalities in Pure and Applied Mathematics*, *10*(4).

Safitri, N. A. N., Purwanti, L. E., & Andayani, S. (2022). Hubungan Perilaku Perawatan Kaki dengan Kualitas Hidup Pasien Diabetes Melitus di RSU Muhammadiyah dan Klinik Rulia Medika Ponorogo. *Health Sciences Journal*, *6*(1), 67–74.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proc. Annu. Symp. Comput. Appl. Med. Care*, *November*, 261–265.

Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2021). Health and Treatment of Diabetes Mellitus. *International Journal of Health Sciences (IJHS)*, *5*(1), 1–5.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson Addison-Wesley.

Triposkiadis, F., Xanthopoulos, A., Bargiota, A., Kitai, T., Katsiki, N., Farmakis, D., Skoularigis, J., Starling, R. C., & Iliodromitis, E. (2021). Diabetes Mellitus and Heart Failure. *Journal of Clinical Medicine*, *10*(16), 3682. https://doi.org/https://doi.org/10.3390/jcm10163682

World Health Organization. (2019). *Classification of Diabetes Mellitus 2019*.

Zilbermint, M. (2020). Diabetes and climate change. *Journal of Community Hospital Internal Medicine Perspectives*, *10*(5), 409–412. https://doi.org/https://doi.org/10.1080/20009666.2020.1791027