Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.9 No.2 July 2024, P-ISSN : 2502-3470, E-ISSN : 2581-0367

K-Nearest Neighbor Method for Early Detection of Diabetes Patients Based on Symptoms and Clinical Data

Nindynar Rikatsih[1], Mochammad Anshori[2], Risqy Siwi Pradini[3], Faurika[4]

[1,2,3,4]*Informatics Department, Institut Teknologi, Sains, dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW, Malang, Indonesia*

[2]moanshori@itsk-soepraoen.ac.id (*)

[1,3][nindynar,risqypradini]@itsk-soepraoen.ac.id

[4]safaurika@gmail.com

Received: 2024-07-09; Accepted: 2024-08-01; Published: 2024-08-07

Abstract—Diabetes is a chronic disease rarely detected and develops quickly. Diabetes can trigger other chronic diseases such as kidney failure and heart disease. Early detection is necessary to help patients treat diabetes before the disease becomes more severe. Various health examination methods to detect diabetes, but these examinations require medical expert action and cannot be carried out by anyone. In addition, examination costs are often unaffordable. This research aims to apply data mining methods, especially k-Nearest Neighbor (KNN), for early detection of diabetes patients based on disease symptoms and patient clinical data. KNN is used to classify patient symptoms and clinical data into two classes, diabetes and non-diabetes, calculating the distance between test data and training data using Euclidean Distance. The research results show that a lower k-value provides a higher accuracy value. However, accuracy at low k-values is insufficient to conclude the performance of KNN for early diabetes detection. High accuracy at low k-values has the potential for overfitting, and the model is not generalizing well. Apart from that, if you use a low k-value, the model only sees patterns from 1 or a few neighbors, which results in the pattern of the data not being captured by the KNN model using a k-value that is too high also risks the model becoming underfitting. The model is too general, which makes the model unreliable. This research made use of the k-fold cross-validation technique to circumvent these issues. It is possible to avoid overfitting in the constructed KNN model by employing this method. The researchers are employing *k-fold=10* and *k-fold=20* in their investigation. KNN This research carried out this analysis by looking at the accuracy of each iteration of the k and k-fold values. The higher the k-fold value, the more accuracy the KNN produces. Inversely proportional to the k-fold cross-validation value, the higher the k-value in KNN, the decreases the accuracy. The KNN method applied in this research provides an accuracy of 98.2692% with higher precision than recall. These findings suggest that KNN can be an effective and efficient tool for early diabetes detection.

Keywords—Classification Techniques; Early Detection of Diabetes; K-Nearest Neighbor Method; Clinical Data; Symptoms Data.

I. INTRODUCTION

Diabetes is a chronic disease that is one of the types of disease with the fastest growing rate throughout the world. Diabetes is predicted to affect 693 million adults by 2045 [1] - [3]. Diabetes can threaten health conditions [4] and trigger disorders of body functionality, such as kidney failure and heart disease [5], making diabetes a disease that needs attention. According to WHO data regarding diabetes sufferers, the number of diabetes patients increased significantly from 314 between 1980 and 2014 [6], [7]. The most worrying facts emerge from low and upper-middle-income countries, which had more than 80 % of people living with diabetes in 2013, with the number always increasing [8], [9]. According to data from the International Diabetes Federation (2019), there are around four million diabetes patients in the world with an age range of 20-79 years, and this number is predicted to continue to increase [10], [11].

To treat diabetes effectively, early detection and therapy are both necessary. It was necessary to conduct a clinical examination to obtain relevant results in the early identification of diabetes. Diabetes often has a lengthy period without symptoms, leading to about half of all individuals with the condition remaining undiagnosed. Various methods are applied to detect diabetes in patients, such as OGTT (Oral Glucose Tolerance Test), HbA1c examination, and blood sugar test [12], [13]. However, a series of tests to detect diabetes is not cheap. In digital era technology, various methods are often proposed to solve prediction problems, such as predicting diabetes [14].

A method that has been proven to be capable of developing a disease detection system is the data mining method. Previous research was carried out to analyze data mining methods with an ensemble approach to diabetes analysis and prediction, namely using random forest, KNN, Naïve Bayes, and J48 methods. The research results show that the KNN method applied in this case did not analyze large datasets well. The proposed method gives better results on small data, while on large data, the proposed method gives relatively poor results [15]. Another study was conducted by applying classification techniques to predict diabetes mellitus. The methods proposed in this research are SVM, Decision Tree, and KNN. KNN involves two experiments: the first is conducted on data that has not been changed, and the second is conducted on data that has been modified using scaling to improve its accuracy. Compared to data that has not been converted, data that has been transformed has a higher level of accuracy, which indicates that there is an influence on the shape of the data through transformation [16].

Further research was carried out using the KNN method on a dataset of diabetes sufferers. This research uses small data, which causes KNN to provide accuracy that is not good enough [17]. Previous research was conducted by Delvika, which

DOI : https://doi.org/10.25139/inform.v9i2.8582

compared the KNN method with Naïve Bayes and gave results that the KNN method produced lower accuracy than Naïve Bayes based on the accuracy results obtained by KNN of 74.48% with a value of *k=25*. In comparison, Naïve Bayes produced an accuracy of 75.78 % with a value of *k=10* [18]. Another research was conducted by Anthony who compared KNN with fuzzy c-means and obtained the results that the fuzzy c-means method was better than KNN with an accuracy of 96% and used a dataset obtained from observations and interviews with diabetes experts at Tanjung Health Center and obtained 120 data with a total of 7 attributes, namely patient, often feel tired, wounds hard to cure, blurred vision, often feel hungry (polyphagia) A history of descendants, as well as a status that includes both positive and negative results, which indicates that no presence of diabetes was found. In the dataset used, the value of each attribute is in the form of a weighing scale on a scale of 0 to 3 and a scale of 0 for no, scale 1 for rarely, scale 2 for often, and scale 3 for very often [19]. Naïve Bayes is good for data with a moderate number of features, and the assumption of feature independence is quite acceptable. Still, this assumption of independence is not always acceptable, and this method is very suitable for irregular data containing noise. KNN is good for structured data and features that can be normalized well and data free from noise. Based on these previous studies, the KNN method produces fairly good accuracy only when using small datasets, while the KNN method produces poor accuracy for large datasets. In this research, the researcher intends to modify previous research by using a large dataset and applying cross-validation to optimize the accuracy produced by the KNN method. The main contribution of this research is testing and evaluating the modified KNN method with cross-validation techniques on a large dataset to increase accuracy in diabetes detection. The novelty of this research lies in the approach that uses cross-validation techniques and data transformation simultaneously on large datasets, which is expected to provide more accurate and consistent results in early diabetes detection using the KNN method.

One data mining method that can be used to predict or detect diabetes is k-Nearest Neighbor (KNN). KNN has been proven to be effective in carrying out classification [20] - [23]. KNN is a simple method, has good resistance to noisy training data, and is effectively used in cases with large training data [24] - [26]. In this research, KNN uses classification techniques to analyze the results of diabetes detection. Detecting diabetes requires patient data from the patient's symptoms and clinical data. This research uses secondary data from 520 instances, 16 patient data features, and one class feature. KNN classifies data based on the nearest neighbor distance, represented by the k-value. Analyzing the k-value to determine the best performance KNN provides when classifying is important. In addition, the distance calculation results are determined using the Euclidean Distance formula. In this research, the method used is k-Nearest Neighbor, which, based on its accuracy, can show whether the data obtained can be used to detect diabetes. This research was conducted to analyze the performance of the k-nearest Neighbor method in the early detection of diabetes based on symptoms and clinical data.

II. Research Methodology

The methodology used during research. The stage begins with identifying disease problems, especially diabetes, followed by conducting a literature study to explore insights from previous research on diabetes, data mining, the application of data mining in the health sector, and cross-validation evaluation techniques. After conducting a literature study, then start collecting data. The data used in this study used secondary data obtained from Sylhet Diabetic Hospital, Bangladesh. After obtaining the dataset, the next stage is to carry out an analysis of KNN. This analysis was carried out using WEKA tools. KNN was analyzed using cross-validation techniques with a value of *k=10*. The methodology of this research is in Figure 1.

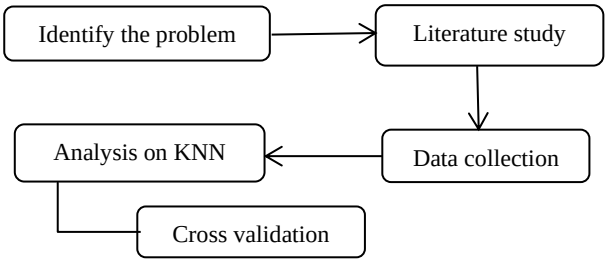


Figure 1. The Methodology of This Research

A. Dataset

Diabetes, especially Diabetes Mellitus or DM, is a metabolic disorder characterized by hyperglycemia due to abnormalities in insulin—high blood sugar results from relative or absolute insulin deficiency [27], [28]. Diabetes consists of a collection of existing conditions categorized more generally based on a single diagnosis [1], [29]. DM is a disease that patients are rarely aware of, and often, when it is discovered, it is at a stage where complications have occurred. This is caused by the relatively long asymptomatic phase of diabetes [14]. Examples of symptoms that diabetes patients can experience include visual disturbances and impaired kidney function [30], [31].

This study used data obtained from Sylhet Diabetic Hospital, Bangladesh, by giving questionnaires directly to patients. The data set consists of 520 instances determined from the number of patients with 17 attributes consisting of age, gender, and symptoms indicating diabetes [14]. The data attributes used in this research are in Table I, the attributes and values contained in each attribute. The patient's age ranges from 25 to 90 years in the table. Patients are male and female. The other attributes consist of 2 value categories, namely no and yes. The class consists of class no, which represents non-diabetes, and class yes, which represents diabetes.

Table I
Attribute Data Set

Attribute	Value
age	aged 25-90 years
gender	0 (male), 1 (female)
poliuria	0 (no), 1 (yes)
polidipsia	0 (no), 1 (yes)
significant weight loss	0 (no), 1 (yes)
weak	0 (no), 1 (yes)
polifagia	0 (no), 1 (yes)

DOI : https://doi.org/10.25139/inform.v9i2.8582

Attribute	Value
Fungal infections of the genital organs	0 (no), 1 (yes)
Blurred vision	0 (no), 1 (yes)
itchy	0 (no), 1 (yes)
irritability	0 (no), 1 (yes)
slow healing	0 (no), 1 (yes)
paresis partial	0 (no), 1 (yes)
stiffness in muscles	0 (no), 1 (yes)
alopecia	0 (no), 1 (yes)
obesity	0 (no), 1 (yes)
class	0 (no), 1 (yes)

Diabetes is detected based on the symptoms experienced by the patient. Each patient from 520 instances had different symptoms. The patient's symptoms are utilized to diagnose whether or not the patient has diabetes. Examples of data sets used to detect diabetes are in Table II.

TABLE II
EXAMPLES OF DATASET

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17
40	M	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
69	F	1	1	1	1	0	0	1	1	1	0	0	1	0	1	1
40	M	0	1	1	1	0	0	1	1	0	0	1	1	0	0	0
29	M	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
90	F	0	1	1	0	0	1	1	1	0	0	0	1	1	0	1

Marks A1, A2, A3 to A17 indicate the attributes and classes of the data set. The first attribute (A1) shows the patient's age, A2 shows the patient's sex or gender, which is represented in the form m for men and f for women, A3 represents symptoms of polyuria, A4 represents symptoms of polydipsia, and so on. Values 0 and 1 indicate whether the patient experiences these symptoms or not. For example, a value of 0 in 3 and 1 in A4 indicates that the patient does not have polyuria but has polydipsia. The values 0 and 1 also apply to other symptoms.

B. K-Nearest Neighbor

K-nearest neighbor (KNN) is a classification technique that uses the k-value as a representation of the number of close neighbors in determining the class or group that corresponds to the object being classified [32] - [34]. The distance of the data to each neighboring k-value is determined using Euclidean Distance as the distance calculation method most commonly used in KNN [35], [36]. Euclidean Distance is the most general and easy formula for calculating distances for classification problems. It is proven to provide higher accuracy than other distance calculation methods such as Hamming, Jaccard, Cosine Distance, and so on [37]. Euclidean Distance is used in Equation (1). Where the n variable is several attributes, the *x* variable is a vector of real attributes of data, the y variable is a vector of attributes resulting from the calculation (output) of data, and the *d(x, y)* variable is a euclidean distance of *x* and *y*.

$$ED = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (1)$$

The *k-value* in KNN is analyzed to determine the optimal value based on the resulting accuracy value. Apart from the *k-value*, there are other considerations in determining how well KNN performs in classification, such as the cross-validation process, precision, and recall. The *k-value* in cross-validation used is *k=10*. Research shows that k-10 in cross-validation is a fairly good model [38]. In this research, the k-fold cross-validation value uses *k=10*.

KNN implementation is carried out using the WEKA application. The WEKA analysis begins by loading the data into WEKA in the "Preprocess" tab and then clicking "Open file". After the file is loaded into WEKA, go to the "Classify" tab, and under "Classifier," click the "Choose" button. Then, select lazy classifier and IBK. After determining the classifier, the next step is determining the evaluation technique. In this study, the cross-validation technique was used with a k-fold value = 10, then under the "test option", select cross-validation with folds-10. Once finished then, click "Start".

C. Model Evaluation

KNN performance is generally assessed based on accuracy. However, accuracy cannot always be a benchmark in determining how well KNN performs, especially if the data is unbalanced. Therefore, KNN performance is measured in accuracy, precision, recall, and f-measure [39]. Accuracy compares the amount of predicted data to the total data. Precision is the ratio of data predicted to be truly positive to all data predicted to be positive [40]. Recall is the comparison of the proportion of predicted data that is truly positive to all data that is actually positive [41]. F-measure is the average between precision and recall and is also called the F1-Score [32]. Accuracy, Precision, and Recall calculations are based on Equations (2), (3), and (4), respectively.

$$Accuration = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Where TP (True Positive): Both the actual and predicted outcomes are positive. TN (True Negative): Both the actual and predicted outcomes are negative. FP (False Positive): The prediction is positive, but the actual outcome is negative. FN (False Negative): The prediction is negative, but the actual outcome is positive.

III. RESULT AND DISCUSSION

Accuracy value testing was performed using 520 instances and 17 attributes in each instance. The accuracy value testing results are in Figure 2. The k-value test was conducted to determine the optimal KNN performance based on accuracy. In Figure 1, the highest accuracy of KNN is obtained at the value *k=1*. *B*ecause *k=1* gives the highest accuracy, we use *k = 1* to test the value of k in cross-validation. The k-fold test shows that

DOI : https://doi.org/10.25139/inform.v9i2.8582

the greater the k-value, the greater the accuracy provided. Figure 1 shows that *k-fold=20* provides an accuracy value often better than *k-fold=10* at every k-value in KNN. A high k-fold value can increase the accuracy value. However, even though the k-fold value increases, the accuracy will decrease as the k-value in KNN increases.

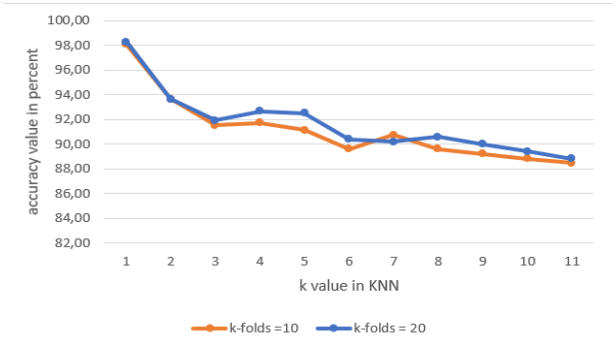


Figure 2. Accuracy Value

The k and k-fold values in KNN have an important role in the performance of the KNN model. A small k-value in KNN has the potential to cause the model to overfit because the model becomes sensitive to noise, which means the model tends to capture small details in the training data, including noise, which causes the model to overfit. Apart from overfitting, a small value of k also causes data variability to be very high because the model completely depends on one or several nearest neighbors. The decisions obtained will also change with small changes in the training data. A small k-value affects the KNN model, and a large k-value also has the potential for underfitting the model. This is because a large value of k makes the model work too generally. After all, decisions are based on many neighbors. This is what causes the model to be unable to capture finer data structures. Essential patterns in the data are overlooked. The advantage of using a large k-value is that it makes the model more general, which makes the KNN model stable and able to handle noise better to make the results more general. The number of folds also influences the performance of the KNN model, namely that a small number of folds has the potential for high bias in the data and low variance. Each fold has more data, which can cause biased model performance estimates. After all, there is less variation in the training and test data. Also, with a small number of folds, the validation process is faster because the number of iterations is smaller, but the results are less stable. Many folds will overcome bias and variance in the data in the KNN model. With many folds, each fold has a higher data variance, which can provide a more accurate estimate of model performance and reduce bias. Many folds also increase the variance because the validation process takes longer. After all, more iterations are required, making the classification results more stable and reliable.

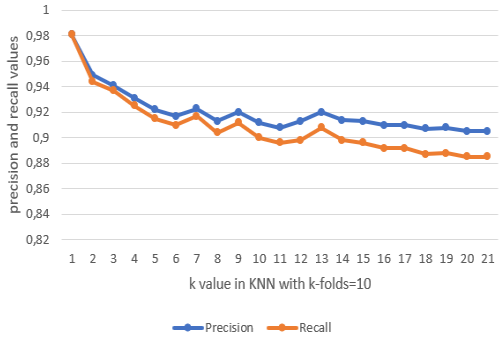
Analyzing the performance of each k and k-fold value to determine the optimal k and k-fold values in the KNN model aims to avoid overfitting and underfitting in the KNN model so that the selected k-value can produce an accurate and reliable model.

The value of k increases the smaller the accuracy provided, as shown in Figure 2. The highest accuracy value is *k=1*, so we can conclude that k = 1 is the k-value that can provide the best results. However, the high accuracy of the *k=1* is due to the test data obtained from training data. The distance calculation results, which represent the prediction results, have values that are not too far away. The value *k=1* can produce the highest accuracy due to various factors, namely, adjusting to the training data. When *k=1*, each data point is classified exactly as its nearest neighbor, which means the model fits the training data very well, resulting in high accuracy because each data point only sees one neighbor in the training data. The second factor is avoiding classification errors. In the training data, if there is a lot of data representing each class, a model with a value of *k=1* will tend to avoid classification errors because each point is classified based on its nearest neighbor. The value *k=1* also affects the generalization of the model in various ways. The first is its sensitivity to noise. k-1 makes the model very sensitive to noise or outliers. If there are incorrect data points or outliers, these points will greatly influence the classification results because the decision only depends on one nearest neighbor. The second is overfitting. k-1 tends to make the model overfitting and has high variability, which causes the model to perform well on training data and produce poor performance on test data. Also, in its performance, the model will produce variable performance with only small changes due to its variability. The third is the lack of generalization.

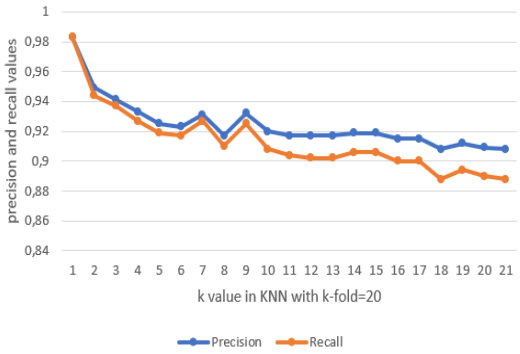
A model with a value of *k=1* becomes very specific to the training data, and its ability to generalize more general data patterns becomes very limited, which causes the model to perform poorly when dealing with new data because it fails to capture broader patterns. Therefore, the high accuracy value given by k = 1 cannot be used to assess KNN in providing optimal performance. A small k-value in KNN risks overfitting because models with small k-values tend to be sensitive to noise or outliers. One data point that is not representative or wrong will cause the classification results to be wrong. Besides being sensitive to noise or outliers, using small k-values causes high variability because the model makes classification decisions using little data. In addition, using a small k-value causes the model to capture fine and specific details of the training data, including noise, and may not capture general data patterns. This is what causes the use of a small k-value to potentially cause overfitting due to the lack of data generalization in the model, which means that the model provides good performance on training data but poor performance on test data or new data. To overcome overfitting in the model, the steps taken in this research are choosing the optimal k-value using a cross-validation technique by trying various k-values to determine the k-value that provides optimal model performance on test data. In this study, *k-fold=10* and *k-fold=20* were used to analyze the performance produced by KNN. Several things must be considered in assessing KNN performance, including k-fold, recall, precision, and f-measure values. Therefore, testing is needed to determine the k-value that can provide the best f-measure value.

DOI : https://doi.org/10.25139/inform.v9i2.8582

Tests that are no less important are precision and recall. Precision states the ratio of existing data predicted to be truly positive with the amount of data that is predicted to be positive. Recall states the ratio of predicted true positive data to the amount of positive data. In this case, precision shows a value that decreases with each k-fold iteration. This also happens to recall, which continues to experience a decrease in value with each k-fold iteration. The precision and recall analysis results indicate that for each k-fold, the precision value is consistently higher than the recall value. The results of precision and recall data testing are in Figure 3. Figures 3(a) and 3(b) illustrate that the precision value consistently exceeds the recall value. The two figures show that the greater the k-value, the smaller the precision and recall values. If the k-value is higher, the discrepancy between the precision and recall values will be much greater.



(a) precision and recall values based on k-fold value = 10



(b) precision and recall values based on k-fold value = 20

Figure 3. The Results Of Precision And Recall Data Testing

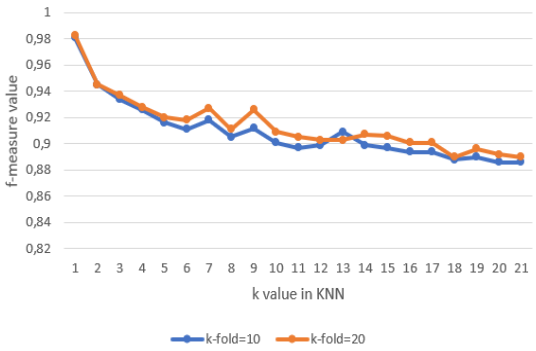


Figure 4. F-measure test

Another test is the f measure. F-measure is the average obtained from precision and recall. F-measure testing can be seen in Figure 4. The F measure value shows a balance between precision and recall. Figure 4 shows that the f-measure value at *k-fold = 20* is not always higher than *k-fold = 10*.

Subsequently, a statistical comparison was conducted to ascertain the difference between *k-fold =10* and *k-fold=20*. The results showed a significance value of 0.019. Since the significance value is below 0.05, it indicates a significant difference between *k-fold=10* and *k-fold=20*, and changes in the k-fold value influence the accuracy provided. The value determined to be optimal is based on the evaluation findings of KNN's performance in classifying diabetic patients according to clinical data and symptoms presented in Table III.

TABLE III
OPTIMAL VALUE FROM THE RESULTS OF KNN PERFORMANCE EVALUATION

KNN performance evaluation	Value
number of instances	520
number of attributes	17
KNN *k-value*	1
k-fold	20
accuracy	98,2692%
precision	0,983
recall	0,983
f-measure (F1-Score)	0,983

It is possible to use the KNN approach to determine whether or not a patient is suffering from diabetes by analyzing the symptoms they experience and the medical records. This classification is based on the results that were obtained. The method of detecting diabetes at an earlier stage will be more effective and efficient because of this data mining software. It has the potential to cut down on the expenses that are required to be expended for carrying out various tests that should not be required.

IV. CONCLUSION

The KNN method can classify diabetes by referring to symptoms and clinical data. In this case, the recall is always lower than the precision. The k-value in cross-validation also influences accuracy. There is a correlation between a larger k-fold cross-validation and increased accuracy. However, the k-value in KNN still has a greater influence, whereas a high k-fold value does not provide better results if the k-value in KNN is larger. In other words, the k-fold value increases with the accuracy value, whereas the k-value in KNN is inversely proportional to the accuracy value. Further research can be carried out by developing the KNN method for early detection of diabetes patients using symptoms and clinical data.

REFERENCES

[1] J. B. Cole and J. C. Florez, "Genetics of diabetes mellitus and diabetes complications," *Nat. Rev. Nephrol.*, vol. 16, no. 7, pp. 377–390, 2020, doi: 10.1038/s41581-020-0278-5.
[2] N. G. Forouhi and N. J. Wareham, "Epidemiology of diabetes," *Med. (United Kingdom)*, vol. 47, no. 1, pp. 22–27, 2019, doi: 10.1016/j.mpmed.2018.10.004.
[3] N. H. Cho *et al.*, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, 2018, doi: 10.1016/j.diabres.2018.02.023.
[4] P. Saeedi *et al.*, "Global and regional diabetes prevalence estimates for

DOI : https://doi.org/10.25139/inform.v9i2.8582

2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Res. Clin. Pract.*, vol. 157, p. 107843, 2019, doi: 10.1016/j.diabres.2019.107843.
[5] J. L. Harding, M. E. Pavkov, D. J. Magliano, J. E. Shaw, and E. W. Gregg, "Global trends in diabetes complications: a review of current evidence," *Diabetologia*, vol. 62, no. 1, pp. 3–16, 2019, doi: 10.1007/s00125-018-4711-2.
[6] A. A. Kazi and L. Blonde, *Classification of diabetes mellitus*, vol. 21, no. 1. 2019. doi: 10.5005/jp/books/12855_84.
[7] E. Standl, K. Khunti, T. B. Hansen, and O. Schnell, "The global epidemics of diabetes in the 21st century: Current situation and perspectives," *Eur. J. Prev. Cardiol.*, vol. 26, no. 2_suppl, pp. 7–14, 2019, doi: 10.1177/2047487319881021.
[8] W. H. Organization, "Diabetes," *https://www.who.int/news-room/fact-sheets/detail/diabetes.* 2021.
[9] H. T. Cheng, X. Xu, P. S. Lim, and K. Y. Hung, "Worldwide Epidemiology of Diabetes-Related End-Stage Renal Disease, 2000-2015," *Diabetes Care*, vol. 44, no. 1, pp. 89–97, 2021, doi: 10.2337/dc20-1913.
[10] Kemenkes, "Infodatin Pusat Data dan Informasi Kementerian Kesehatan RI Tetap Produktif, Cegah, dan Atasi Diabetes Melitus," *https://pusdatin.kemkes.go.id/resources/download/pusdatin/infodatin/Infodatin-2020-Diabetes-Melitus.pdf*. Pusat Data dan Informasi Kementrian Kesehatan RI.
[11] R. Williams *et al.*, "Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Res. Clin. Pract.*, vol. 162, 2020, doi: 10.1016/j.diabres.2020.108072.
[12] D. Simmons *et al.*, "Treatment of Gestational Diabetes Mellitus Diagnosed Early in Pregnancy," *N. Engl. J. Med.*, vol. 388, no. 23, pp. 2132–2144, 2023, doi: 10.1056/nejmoa2214956.
[13] R. Marium *et al.*, "From Pre-Diabetes to Diabetes : Diagnosis ," *Medicina (Kaunas).*, vol. 55, no. 9, p. 546, 2019.
[14] R. Murugan, *The Retinal Blood Vessel Segmentation Using Expected Maximization Algorithm*, vol. 992. 2020. doi: 10.1007/978-981-13-8798-2_6.
[15] M. Alehegn, R. R. Joshi, and P. Mulay, "Diabetes analysis and prediction using random forest, KNN, Naïve Bayes, and J48: An ensemble approach," *Int. J. Sci. Technol. Res.*, vol. 8, no. 9, pp. 1346–1354, 2019.
[16] A. S. Hassan, I. Malaserene, and A. A. Leema, "Diabetes Mellitus Prediction using Classification Techniques," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 5, pp. 2080–2084, 2020, doi: 10.35940/ijitee.e2692.039520.
[17] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.
[18] P. D. Rinanda, B. Delvika, S. Nurhidayarnis, N. Abror, and A. Hidayat, "Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes pada Ibu Hamil," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 2, pp. 68–75, 2022, doi: 10.57152/malcom.v2i2.432.
[19] A. Anggrawan and M. Mayadi, "Application of KNN Machine Learning and Fuzzy C-Means to Diagnose Diabetes," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 2, pp. 405–418, 2023, doi: 10.30812/matrik.v22i2.2777.
[20] N. Rikatsih and A. A. Supianto, "Classification of Posture Reconstruction with Univariate Time Series Data Type," *3rd Int. Conf. Sustain. Inf. Eng. Technol. SIET 2018 - Proc.*, pp. 322–325, 2018, doi: 10.1109/SIET.2018.8693174.
[21] A. Hamed, A. Sobhy, and H. Nassar, "Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a KNN Variant Algorithm," *Arab. J. Sci. Eng.*, vol. 46, no. 9, pp. 8261–8272, 2021, doi: 10.1007/s13369-020-05212-z.
[22] S. Zhang, "Challenges in KNN Classification," *IEEE Trans. Knowl. Data Eng.*, pp. 1–13, 2021, doi: 10.1109/TKDE.2021.3049250.
[23] S. Zhang, "Cost-sensitive KNN classification," *Neurocomputing*, vol. 391, no. xxxx, pp. 234–242, 2020, doi: 10.1016/j.neucom.2018.11.101.
[24] A. Rudiyan, A. E. Dzulkifli, and K. Munazar, "Klasifikasi Kebakaran Hutan Menggunakan Metode K-Nearest Neighbor : Studi Kasus Hutan Provinsi Kalimantan Barat," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 3, no. 4, pp. 195–202, 2022, doi: 10.35746/jtim.v3i4.177.
[25] M. Ali, L. T. Jung, A. H. Abdel-Aty, M. Y. Abubakar, M. Elhoseny, and I. Ali, "Semantic-k-NN algorithm: An enhanced version of traditional k-NN algorithm," *Expert Syst. Appl.*, vol. 151, p. 113374, 2020, doi: 10.1016/j.eswa.2020.113374.
[26] Y. Chen *et al.*, "Fast density peak clustering for large scale data based on kNN," *Knowledge-Based Syst.*, vol. 187, p. 104824, 2020, doi: 10.1016/j.knosys.2019.06.032.
[27] ADA, "Ada 2022," *Diabetes Care*, vol. 45, no. Suppl, pp. 17–38, 2022.
[28] Saha et al, "A Review on Diabetes Mellitus : Type1 & Type2," *World J. Pharm. Pharm. Sci.*, vol. 9, no. 10, pp. 838–850, 2020, doi: 10.20959/wjpps202010-17336.
[29] S. Ellahham, "Artificial Intelligence: The Future for Diabetes Care," *Am. J. Med.*, vol. 133, no. 8, pp. 895–900, 2020, doi: 10.1016/j.amjmed.2020.03.033.
[30] S. Alam, M. K. Hasan, S. Neaz, N. Hussain, M. F. Hossain, and T. Rahman, "Diabetes Mellitus: Insights from Epidemiology, Biochemistry, Risk Factors, Diagnosis, Complications and Comprehensive Management," *Diabetology*, vol. 2, no. 2, pp. 36–50, 2021, doi: 10.3390/diabetology2020004.
[31] L. Wen *et al.*, "The Role of Catechins in Regulating Diabetes: An Update Review," *Nutrients*, vol. 14, no. 21, 2022, doi: 10.3390/nu14214681.
[32] A. Deharja, M. W. Santi, M. Yunus, and E. Rachmawati, "Sistem Prototype Klasifikasi Risiko Kehamilan Dengan Algoritma k-Nearest Neighbor (k-NN)," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 4, no. 1, pp. 66–72, 2022, doi: 10.35746/jtim.v4i1.229.
[33] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754.
[34] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci. Rep.*, vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-10358-x.
[35] A. R. Lubis, M. Lubis, and Al-Khowarizmi, "Optimization of distance formula in k-nearest neighbor method," *Bull. Electr. Eng. Informatics*, vol. 9, no. 1, pp. 326–338, 2020, doi: 10.11591/eei.v9i1.1464.
[36] N. Hidayati and A. Hermawan, "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation," *J. Eng. Appl. Technol.*, vol. 2, no. 2, pp. 86–91, 2021, doi: 10.21831/jeatech.v2i2.42777.
[37] E. W. Sholeha, S. Yunita, R. Hammad, V. C. Hardita, and K. Kaharuddin, "Analisis Sentimen Pada Agen Perjalanan Online Menggunakan Naïve Bayes dan K-Nearest Neighbor," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 3, no. 4, pp. 203–208, 2022, doi: 10.35746/jtim.v3i4.178.
[38] Nti Isaac Kofi, O. Nyarko-Boateng, and J. Aning, "Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation," *Int. J. Inf. Technol. Comput. Sci.*, vol. 13, no. 6, pp. 61–71, 2021, doi: 10.5815/ijitcs.2021.06.05.
[39] B. Juba and H. S. Le, "Precision-Recall versus accuracy and the role of large data sets," *33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019*, pp. 4039–4048, 2019, doi: 10.1609/aaai.v33i01.33014039.
[40] A. F. Sadeli and I. I. Lawanda, "Recall, Precision, and F-Measure for Evaluating Information Retrieval System in Electronic Document Management Systems (EDMS)," *Khizanah al-Hikmah J. Ilmu Perpustakaan, Informasi, dan Kearsipan*, vol. 11, no. 2, pp. 231–241, 2023, doi: 10.24252/kah.v11i2a8.
[41] A. Gupta, A. Anand, and Y. Hasija, "Recall-based Machine Learning approach for early detection of Cervical Cancer," *2021 6th Int. Conf. Converg. Technol. I2CT 2021*, pp. 1–5, 2021, doi: 10.1109/I2CT51068.2021.9418099.

DOI : https://doi.org/10.25139/inform.v9i2.8582