Classification of diabetes mellitus disease at Rato Ebuh Hospital-Indonesia using the K-Nearest neighbors method based on missing value

Sigit Susanto Putro[1*], *Moh Abdan Syakura* Putra[1], *Doni Abdul* Fatah[1], *Yuli Panca* Asmara[2], *Hermawan Bin* Fauzan[1], *Eka Mala Sari* Rochman[1] and *Aeri* Rachmad[1]

[1]Departemen of Informatics, Faculty of Engineering, University of Trunojoyo Madura, Kamal, Bangkalan, Indonesia
[2]Faculty of Engineering and Quantity Surveying, INTI International University, Negeri Sembilan 71800, Malaysia

Abstract. Diabetes mellitus is a chronic disease often caused by high blood glucose levels and insufficient insulin production. This research aims to address the classification problem of diabetes mellitus using the K-Nearest Neighbor (K-NN) method. The aim of this research is to create a machine learning model that can detect diabetes early. The study was conducted at Syarifah Ambami Rato Ebu Hospital in Bangkalan, utilizing data from 120 patients in 2019, employing data mining techniques to classify diabetes mellitus patients. Additionally, the steps in data mining involve determining significant variables or features for classification Cleansing and normalization and transformation. The research compares training test results with ratios of 90:10, 80:20, and 70:30. Experimental results show that K-NN with a neighbor value of K=11 achieves the highest accuracy rate of 83% a reduced error rate of 16.67%, and the highest AUC value of 0.7407. These results indicate that the 90:10 data split ratio yields the best model performance in terms of accuracy and class differentiation for diabetes mellitus, as well as the lowest error rate compared to other data split ratios. This study provides a better understanding of diabetes mellitus and demonstrates that K-NN is effective in addressing classification problems, focusing on specific variables that influence the disease. Therefore, it can be concluded that K-Nearest Neighbor (K-NN) is a suitable algorithm for classifying diabetes mellitus.

1 Introduction

Diabetes mellitus (DM), commonly known as diabetes, is a collection of metabolic disorders marked by persistently high blood sugar levels. These elevated levels lead to symptoms like frequent urination, excessive thirst, and heightened hunger [1]. People's habit of consuming MBDK excessively is an unhealthy behavior which is one of the main causes of diabetes. As a result, Indonesia was recorded as being in fifth place with the highest number of diabetes cases in the world after China, India, Pakistan and America in 2021.

Diabetes Mellitus is categorized into two types: (1) Type I Diabetes Mellitus, which is caused by the pancreas producing very little or no insulin at all, typically occurring in children and non-obese adolescents, and (2) Type II Diabetes Mellitus, which occurs when the pancreas does not produce enough insulin or when the body's fat and muscle cells become resistant to insulin. Type II diabetes is generally associated with obesity, physical activity levels, diet, and other factors [2].

Several studies have shown that diabetes mellitus not only affects blood sugar levels but also has significant consequences on various other aspects of health. For example, it increases the risk of chronic kidney disease, mental health disorders, and hearing impairment [3-6].

The World Health Organization (WHO) forecasts that the number of individuals with diabetes in Indonesia will rise from 8.4 million in 2000 to around 213 million by 2030. WHO data indicates that the global burden of diabetes mellitus was 135 million people in 2000, and this burden is expected to rise to 366 million people within the next 25 years (by 2025) [7].

Considering the increasing prevalence of diabetes and its serious consequences, as predicted by the WHO, effective planning and medical interventions become crucial. A growing approach in medical research involves utilizing machine learning algorithms to aid in the diagnosis and management of diseases. Recently, numerous studies have explored the application of different algorithms for the classification of diabetes mellitus.

For example, research by Dewi Ratnasari and Iswanto, which studied diabetes mellitus and stroke using the K-Nearest Neighbor (KNN) model with Euclidean and Minkowski distance methods, demonstrated accuracy levels ranging from 85% to 96.03% [8, 9]. Other studies have even achieved higher accuracy, reaching 93.58%, with data variables nearly identical to those used in this research [10].

* Corresponding author: sigit.putro@trunojoyo.ac.id

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

In another study conducted by Daut Sohot Siregar and colleagues, the F1 Score achieved was 0.907, indicating a good balance between precision and recall [11]. In another study, it was claimed that applying min-max scaling to data before inputting it into the KNN model improved performance. The accuracy increased by 5.29%, resulting in a final accuracy of 82.5% [12].

Another study conducted by Umikulsum Indah Lestari, experimented with K values up to 26 and achieved the highest accuracy at K=23, with a very high accuracy rate of 96% [13]. Another study conducted by Dwi Fasnuari and colleagues also achieved a high accuracy of 93%. In this study, they used 108 training data and 27 testing data, resulting in an accuracy of 93% at K=9, with precision of 100%, recall of 60%, and an F1-Score of 75% [14].

In another study conducted by Susanto and colleagues, which focused on heart disease data using the K-Nearest Neighbor (KNN) algorithm, high accuracy was also achieved. This research utilized a dataset of 200 samples, with the testing conducted using a split validation method, where 180 samples were used for training and 20 for testing. With a K value of 3, the study achieved an accuracy of 95% [15].

Based on these findings, it can be concluded that the use of the KNN algorithm in diabetes mellitus classification shows significant potential. Through further examination and the application of appropriate techniques, such as min-max scaling, the performance of the model can be enhanced to provide greater benefits in the diagnosis and management of diabetes.

Based on the existing background, the contribution of this research is to classify diabetes using the KNN method. This is of course related to early treatment of diabetes patients.

2 Methods

Data mining, often known as knowledge discovery in databases (KDD), involves the collection and utilization of historical data to identify patterns, regularities, or relationships within large datasets [16]. The results of the data mining process can be used to improve decision-making in the future. Additionally, the steps in data mining involve determining significant variables or features for classification and regression. Therefore, data mining plays a crucial role in diagnosis and healthcare [17]. Data mining focuses on processing large-scale data, and the stages of the process are illustrated below.

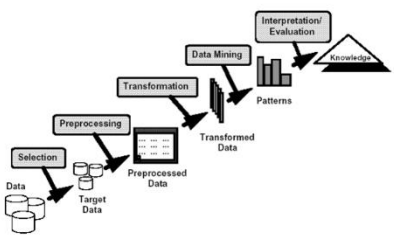

Fig. 1. Stages of data mining.

According to Jack Billie Chandra [10], Fig. 1, the KDD process can be broadly explained as follows:

1. Data Cleaning

The process of cleaning data to remove noise and inconsistent data.

2. Data Integration

The integration of data where multiple data sources can be combined.

3. Selection

Data selection, where relevant data for the analysis task is retrieved from the database.

4. Data Transformation

Data transformation, where data is transformed and consolidated into a suitable format for mining by performing summary or aggregation operations.

5. Data Mining

The essential process where intelligent methods are applied to extract data patterns.

6. Patern Evaluation

Pattern evaluation to identify truly interesting patterns that represent knowledge.

7. Knowledge Presentation

Knowledge presentation, where visualization and knowledge representation techniques are used to present the mined knowledge to users.

2.1 Data

In this study, the dataset used comes from RSUD Syarifa Ambami Rato Ebhu, Bangkalan Regency. This dataset is a recapitulation of diabetes patient data, consisting of 120 records. The data is presented in a table that includes information such as age, glucose level, blood pressure, skin thickness, insulin, BMI, and the final outcome. All of this data will serve as the basis for analysis in this study.

The attributes or criteria utilized for the classification process are described as follows:

1. Age

This feature may have characteristics that influence whether a person contracts diabetes mellitus. For example, individuals over the age of 30 may experience a decline in insulin production and irregular or improper dietary habits.

2. Glucose

The relationship with diabetes is closely linked to the significant role of glucose in regulating blood sugar levels. Glucose, as the primary form of sugar providing energy to body cells, plays a crucial role in diabetes conditions. The regulation of blood glucose levels should remain stable, controlled by the insulin hormone produced by the pancreas.

3. Blood Pressure

An increase in blood pressure in hypertensive conditions is closely related to the neglect of salt and water retention or increased internal pressure in the peripheral blood vessels.

4. Skin Thicness

Skin thickness is not generally considered a major risk factor or direct cause of diabetes mellitus. However,

certain skin conditions may be associated with diabetes, such as discolored, dry, and infected skin.

5. Insulin

Typically, individuals with diabetes mellitus exhibit a tendency towards insulin resistance or a disruption in insulin secretion. This leads to elevated blood glucose levels or hyperglycemia.

6. BMI (Body Mass Index)

While BMI is not a direct cause of diabetes mellitus, several symptoms may lead to its development. For instance, excess fat can cause insulin resistance, which is an early symptom of diabetes mellitus.

7. Outcome

After considering the results of all parameters, it is determined whether an individual has diabetes mellitus or not.

2.2 Data Mining

2.2.1 Pre-processing

Data pre-processing consists of a series of steps aimed at cleaning, transforming, and preparing raw data for subsequent analysis or modelling [18]. The main goals of data pre-processing are to enhance data quality, reduce noise or uncertainty, and eliminate potential inconsistencies within the dataset. During the cleaning stage, some irrelevant data, such as features like Number and Name, are removed. Next, normalization is performed using the Min-Max method to ensure that the data falls within an appropriate range. This process entails scaling the data values to fall within a specific range, usually between 0 and 1. This transformation enhances the performance of machine learning algorithms by ensuring that all features have an equal impact on the analysis[19]. Additionally, handling missing values is a crucial step in pre-processing. Techniques such as K-Nearest Neighbor (KNN) imputation or replacing missing values with the group-based mean can be used to address incomplete data, ensuring the dataset is robust and reliable for subsequent analysis.

The aim of this research is to predict the initial possibility that a person will be diagnosed with diabetes based on the diabetes dataset.

2.2.2 Data transformation

Data transformation entails altering the scale of data to a different form to ensure that the data distribution aligns with expectations. Each data value undergoes consistent mathematical operations applied to the original data. These modifications aim to preserve the relative differences between data points. When dealing with multiple data variables, all variables are transformed to maintain the unchanged relative relationships among the data. In this study, data transformation includes converting categorical values in the dataset into numerical data, focusing primarily on three parameters: glucose, blood pressure, and BMI.

2.2.3 Imputation of missing values

Missing values refer to situations where information for a given subject is absent. The causes of missing values can vary, such as errors in data collection or data being unreadable by the system, leading to values being considered missing [20]. Several methods exist to handle missing values, one of which is imputation using the K-Nearest Neighbor (KNN) method. The KNN imputation method can be used to predict two types of data, discrete data using the mode value and continuous data using the mean value.

The KNN imputation technique is a common and effective method for addressing missing values in multivariate data. The K-Nearest Neighbor Imputation method has become a popular choice for handling missing data, especially in datasets with multiple incomplete variables. This KNN approach utilizes similar or comparable observations to the ones with missing values for imputation.

Besides the KNN method, another approach for handling missing values is the group-based mean concept. This technique, commonly used in statistical analysis of collective data, involves replacing missing values within a group with the mean value of the non-missing data within that group[19].

The closest distance is calculated using the Euclidean formula, as follows

$$d(a,b) = \sum_{i=1}^{n} (Xi - Yi)^2 \quad (1)$$

Describes:
$d(a, b)$: Euclidean distance
X_i : Value of the attribute in the first data point,
Y_i : Value of the attribute in the second data point,
I : Attribute index,
n : Number of attributes.

2.2.4 Normalization

Data normalization is an initial step in the classification stage, where the attribute values of the data are adjusted to fall within a specific range. The normalization process in the dataset is implemented to achieve the goal of standardizing data distribution and enhancing the accuracy of the system. The method used in this study is the Min-Max technique, which is commonly applied in computer programming and artificial intelligence to determine the minimum (min) and maximum (max) values of a variable or function within a given range. The main objective of the Min-Max algorithm is to find the best move that will provide the maximum result (maximum gain or value) for the player using the algorithm, while also anticipating the move that could be the most detrimental (minimum gain or value) taken by the opponent. The formula is as follows:

$$z = \frac{x - \min(x)}{[\max(x) - \min(x)]} \quad (2)$$

Describes:
z : Normalized value,
x : Original value x,
min(x) : Minimum value for the variable x,
max(x) : Maximum value for the variable x.

2.2.5 Feature selection

Feature selection using the chi-square (χ^2) method is a technique used in data analysis to identify features that have a significant relationship with the target variable in categorical or frequency data. This method compares the observed and expected frequency distributions of each feature against the target variable, then calculates the chi-square value to reflect its significance. Features with high chi-square values are considered more relevant and are deemed to contribute significantly to class separation in the data. Thus, feature selection using chi-square helps identify the most informative subset of features to enhance model performance.

2.3 KNN Algorithm (K Nearest Neighbors)

The K-Nearest Neighbor (K-NN) algorithm is a classification method that uses training data to determine the class of an object based on its proximity to other objects [21]. The accuracy of the K-NN algorithm is greatly influenced by the presence of irrelevant features or the unequal weighting of features for classification [22]. Most research related to this algorithm focuses on selecting and determining feature weights to improve classification performance. K-NN can also be categorized as an example of lazy learning, where the algorithm waits for a new query before performing the learning process, making it similar to the training data.

The steps of the K-Nearest Neighbor (K-NN) algorithm are as follows:

1. Prepare training data and testing data.
2. Determine the value of k.
3. Calculate the distance of the testing data to each training data point using the Euclidean distance formula as follows:

$$d(x1,x2) = \sqrt{\sum_{i=1}^{n}(x1i - x2i)^2} \quad (3)$$

Describes:
X1i = Training Data
X2i = Testing Data

4. Determine the value of k based on the training data points that are closest to the testing data.
5. Check the labels of the k nearest training data points.
6. Determine the label that occurs most frequently.
7. Assign the testing data to the class with the most frequenct label.
8. Termination condition.

K-Nearest Neighbor (K-NN) has the advantage of being able to produce robust or clear data and is effective when used on large datasets. However, K-Nearest Neighbor also has some drawbacks: it requires a value for k, the distance from the test data is unclear with the type of distance used, and to obtain the best results, all attributes or only one definite attribute must be used. The K-Nearest Neighbor (K-NN) method is quite simple, there are no assumptions about data distribution, and it is easy to apply. The selection of the K value (the number of nearest neighbors) is determined by the researcher. The choice of K value can affect the accuracy of the predictions.

2.4 Evaluation

When modeling the classification process, it is crucial to assess the system's performance to determine the effectiveness of the applied method. A widely used approach for this evaluation is the confusion matrix. This matrix helps evaluate the accuracy, precision, recall, and F-measure of the algorithm used in the study, based on the predicted results from the test data.

Accuracy represents the effectiveness of the overall classification process. The formula for accuracy is as follows:

1. Accuracy represents the effectiveness of the overall classification process, calculated with the formula[17]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

2. Precision is the percentage of correctly classified positive labels, calculated with the formula:

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (5)$$

3. Recall is the effectiveness of the classification process in identifying positive labels, calculated with the formula:

$$Recall \quad = \frac{TN}{TN + FN} \times 100\% \quad (6)$$

4. The F1-measure is the harmonic mean of precision and recall, with a range from 0 to 1:

$$F - measure = 2 \times \frac{Recall\ x\ Precision}{Recall + Precision} \times 100\% \quad (7)$$

Describes:
TP: the number of data points predicted as positive correctly,
TN: the number of data points with an actual positive class but predicted as negative,
FN: the number of data points predicted as negative correctly,
FP: the number of data points with an actual negative class but predicted as positive.

3 Result and discussion

3.1 Data collection

The data used in this classification process is diabetes data from residents of Bangkalan, consisting of 120 records and 7 attributes: age, glucose, blood pressure, skin thickness, insulin, BMI, and outcome. There are no missing values in the data related to diabetes mellitus. Table 1 shows the sample dataset

3.2 Analysis

3.2.1 Data input process

In this process, the data used for classification consists of information from diabetes mellitus patients, with a total of 7 features. Table of Diabetes Mellitus Dataset: 10-15 Records.

Table 1. Sample dataset.

age	glucose	blood pressure	skin thickness	insulin	BMI	Outcome
54	148	72	35	0	33,6	Diabetes Melitus
43	85	66	29	0	26,6	Diabetes Melitus
27	183	64	0	0	23,3	Diabetes Melitus
45	89	66	23	94	28,1	Diabetes Melitus
55	137	40	35	168	43,1	Diabetes Melitus

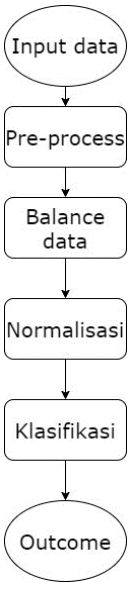

Fig. 2. Classification process.

3.2.2 Data Pre-processing

At this stage, a series of steps are taken to clean and organize the data that will be analyzed, ensuring that the results are significantly better than those from the raw data. The goal is to address missing values. Missing values are handled using the K-Nearest Neighbor (KNN) method with the nearest neighbor's value (k) set to 5, as this produces better results. Consequently, this approach affects the accuracy of the results.

	Usia	Glukosa	Tekanan darah	Ketebalan Kulit	Insulin	BMI	Outcome
0	54	148.0	72.0	35	89.0	33.6	1
1	43	85.0	66.0	29	568.0	26.6	1
2	27	183.0	64.0	33	90.0	23.3	1
3	45	89.0	66.0	23	94.0	28.1	1
4	55	137.0	40.0	35	168.0	43.1	1
5	57	116.0	74.0	28	87.0	25.6	1
6	56	78.0	50.0	32	88.0	31.0	1
7	55	115.0	67.0	37	875.0	35.3	1
8	67	197.0	70.0	45	543.0	30.5	1
9	55	125.0	96.0	44	65.0	41.0	1
10	51	110.0	92.0	16	76.0	37.6	1
11	49	168.0	74.0	19	34.0	38.0	1
12	48	139.0	80.0	35	57.0	27.1	1
13	40	189.0	60.0	23	846.0	30.1	1
14	48	166.0	72.0	19	175.0	25.8	1
15	60	100.0	77.0	37	54.0	30.0	1
16	65	118.0	84.0	47	230.0	45.8	1
17	34	107.0	74.0	36	678.0	29.6	1
18	57	103.0	30.0	38	83.0	43.3	1
19	58	115.0	70.0	30	96.0	34.6	1
20	45	126.0	88.0	41	235.0	39.3	1

Fig. 3. Data after pre-processing.

3.2.3 Data normalization

Data normalization using MinMaxScaler is a preprocessing technique aimed at transforming feature values into a specific range, typically between 0 and 1. This is achieved by adjusting each feature value so that the smallest value becomes 0 and the largest value becomes 1.

	Usia	Glukosa	Tekanan darah	Ketebalan Kulit	Insulin	BMI	Outcome
0	56	0.604651	0.228571	39	741	0.375465	1
1	48	0.550388	0.571429	35	57	0.27881	1
2	60	0.248062	0.528571	37	54	0.386617	1
3	44	0.186047	0.314286	7	258	0.297398	1
4	51	0.75969	0.457143	19	176	0.238483	1
5	81	0.511628	0.485714	33	60	0.234201	0
6	55	0.581395	0.771429	33	146	0.63197	0
7	58	0.589147	0.6	26	285	0.468967	1
8	44	0.992248	0.714286	32	54	0.758929	0
9	63	0.255814	0.514286	48	180	0.494424	0
10	47	0.736434	0.457143	41	114	0.791822	1
11	51	0.674419	0.514286	28	150	0.589294	1
12	23	0.697674	0.514286	36	245	0.446897	0
13	50	0.744186	0.6	43	67	0.490706	0
14	48	0.403101	0.571429	37	150	0.643866	1
15	55	0.44186	0.8	44	65	0.795539	1
16	46	0.410605	0.542857	31	94	0.297398	0
17	39	0.44186	0.428571	26	115	0.427509	1
18	52	0.837209	0.657143	27	196	0.589294	1
19	52	0.379845	0.742857	18	978	0.539033	0
20	51	0.875969	0.628571	21	192	0.605948	1

Fig. 4. Data after normalization.

3.2.4 Train-test split

Train-test split is a model validation technique where the dataset is divided into two subsets: training and testing. The goal of train-test split is to measure how well the model can generalize from training data to previously unseen data. The training-testing scenarios in this study are 70:30, 80:20, and 90:10. Accuracy results can be seen in table 2.

Data splitting occurs when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other part to train the model. Data separation is an important aspect of data science, especially for creating models based on data.

Table 2. Accuracy from split data.

	70:30	**80:20**	**90:10**
Training	84	96	108
Testing	36	24	12

3.2.5 Classification process

At this stage, the learning process focuses on developing a classification model using the K-Nearest Neighbors (K-NN) algorithm. The performance of this model will be evaluated to determine its effectiveness in classification tasks.

3.2.6 Output

After completing the entire process, the output will consist of class predictions for the KET attribute or outcome based on the modeling with the methods proposed in this study.

3.3 K-Nearest neighbors

In the experiment using the K-Nearest Neighbors (KNN) method, the initial step involves exploring kkk values to determine which yields the highest accuracy. After identifying several promising kkk values, the next step involves validating the data using k-fold cross-validation. The flowchart for the KNN process can be seen in the KNN process diagram.

In this study, the collected data amounts to 120 sets. Before conducting validation tests, an initial test with varying parameter k was performed. The parameter k that provided the highest accuracy was then selected for further testing using data validation. The testing results showed the system's accuracy for several values of k, specifically k=2 with an accuracy of 75% and k=11 with an accuracy of 83.33%. Based on the experiments with varying k values, the highest accuracy was achieved at k=11. Therefore, k=11 was chosen for further testing using data validation with the application of a confusion matrix. The comparison of K-NN accuracy results was conducted to evaluate the performance of K-Nearest Neighbor using all attributes in the dataset. The values of k used were k=2 and k=12, where the nearest neighbors included k=12 with both odd and even numbers of neighbors. Calculations were performed to obtain accuracy, precision, and recall. Detailed evaluation results can be seen in the accompanying figure of this study. And after reviewing the results from the confusion matrix above, the accuracy results will be displayed in the table below. Table 2: Classification Results of K-Nearest Neighbor without SMOTE.

Table 3. Accuracy results.

Train Test	Percentage						
	acc	prcs		rcl		F1-S	
		0	1	0	1	0	1
70:30	63.89	0.38	0.71	0.27	0.80	0.32	0.75
80:20	66.67	0.40	0.74	0.29	0.82	0.33	0.78
90:10	83.33	1.00	0.82	0.33	1.00	0.50	0.90

The table 3 shows the average results for accuracy, precision, and recall. These parameters are used to assess performance accuracy. As a result, the best value for k in the 90:10 scenario is k=11, with an accuracy performance of 83.33%.

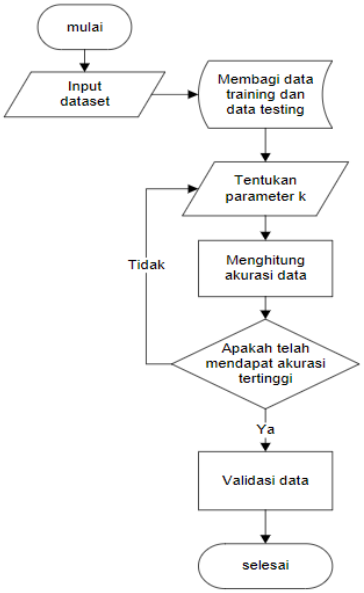

Fig. 5. KNN process.

4 Conclusion

Based on the data analysis and exploration in this study, an effective model was successfully implemented to address the issue of missing values in the diabetes mellitus dataset at RSUD Syarifah Ambami Rato Ebu Bangkalan, using the K-NN method approach.

During the performance evaluation stage, parameters such as accuracy, error, and Area Under the Curve (AUC) were used to compare the effectiveness of each model in the classification of diabetes mellitus. The performance comparison results of each model are presented in a table that visualizes the differences in performance.

Table 4. Comparison results.

	70 : 30	80 : 20	90 : 10
Accuracy	63.89%	66.67%	83.33%
Error	36.11%	33.33%	16.67%
AUC	0.5018	0.5462	0.7407

Based on the table of model performance evaluation, the following conclusions can be drawn. With a 70 : 30 data split ratio, the model achieved an accuracy of 63.89%, an error rate of 36.11%, and an Area Under the Curve (AUC) value of 0.5018. When the data ratio was changed to 80: 20, there was an improvement in accuracy to 66.67%, a reduction in the error rate to 33.33%, and an increase in the AUC value to 0.5462.

However, the 90 :10 data split ratio provided the most optimal results, with an accuracy of 83.33%, a reduced error rate of 16.67%, and the highest AUC value of 0.7407. These results indicate that the 90 : 10 data split ratio yields the best model performance in terms of accuracy and class differentiation for diabetes mellitus,

as well as the lowest error rate compared to other data split ratios.

References

1. R. Kumar, P. Saha, S. Sahana, and A. Dubey, *A Review On Diabetes Mellitus: Type1 & Type2*, World J Pharm Pharm Sci, **9**, 10, (2020), doi: 10.20959/wjpps202010-17336
2. M. S. Fiqri and H. Dwi Bhakti, *Klasifikasi Potensi Penyakit Diabetes Mellitus Tipe Ii Pada Pasien Menggunakan Algoritme Knn (K-Nearest Neighbo)r*, (2024)
3. O. Kelly et al., *The impact of diabetes mellitus on the development of psychiatric and neurological disorders*, , Elsevier B.V., (2024), doi: 10.1016/j.dscb.2024.100135.
4. D. Samocha-Bonet, B. Wu, and D. K. Ryugo, *Diabetes mellitus and hearing loss: A review*, Elsevier Ireland Ltd., (2021), doi: 10.1016/j.arr.2021.101423
5. S. Gysling, C. A. Lewis-Lloyd, D. N. Lobo, C. J. Crooks, and D. J. Humes, *The effect of diabetes mellitus on perioperative outcomes after colorectal resection: a national cohort study*, Br J Anaesth, **133**, 1, 67–76, (2024), doi: 10.1016/j.bja.2024.04.010
6. P. Tao et al., *Diabetes mellitus is a risk factor for incident chronic kidney disease: A nationwide cohort study*, Heliyon, **10**, 7, (2024), doi: 10.1016/j.heliyon.2024.e28780
7. S. Syarifuddin, W. Samosir, and U. Efarina, *Characteristics Of Types Of Diabetes Mellitus Ii In Regional General Hospital Than Rondahaim, Simalungun District*, Medical Research, Nursing, Health and Midwife Participation, (2019), Available: https://medalionjournal.com/
8. D. Ratnasari, *Comparison of Performance of Four Distance Metric Algorithms in K-Nearest Neighbor Method on Diabetes Patient Data*, Indonesian Journal of Data and Science, **4**, 2, 101–112, (2023), doi: 10.56705/ijodas.v4i2.71
9. I. Iswanto, T. Tulus, and P. Sihombing, *Comparison of Distance Models on K-Nearest Neighbor Algorithm in Stroke Disease Detection*, Applied Technology and Computing Science Journal, **4**, 1, 63–68, (2021), doi: 10.33086/atcsj.v4i1.2097
10. J. B. Chandra and D. Nasien, *Application Of Machine Learning K-Nearest Neighbour Algorithm To Predict Diabetes*, International Journal of Electrical, Energy and Power System Engineering (IJEEPSE), **6**, (2023), Available: http://www.ijeepse.ejournal.unri.ac.id
11. J. Mantik et al., *Implementation of KNN algorithm in classifying diabetic ulcers in patients with diabetes mellitus*, (2023)
12. T. A. Assegie, T. Suresh, R. Purushothaman, S. Ganesan, and N. K. Kumar, *Early Prediction of Gestational Diabetes with Parameter-Tuned K-Nearest Neighbor Classifier*, Journal of Robotics and Control (JRC), **4**, 4, 452–457, (2023), doi: 10.18196/jrc.v4i4.18412
13. F. Solihin, M. Syarief, E. M. S. Rochman, & A. Rachmad., *Comparison of Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Stochastic Gradient Descent (SGD) for Classifying Corn Leaf Disease based on Histogram of Oriented Gradients (HOG) Feature Extraction,* Elinvo (Electronics, Informatics, and Vocational Education), **8**, 1, 121-129, (2023), http://jurnal.mdp.ac.id
14. H. A. Dwi Fasnuari, H. Yuana, and M. T. Chulkamdi, *Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Penyakit Diabetes Melitus*, Antivirus: Jurnal Ilmiah Teknik Informatika, **16**, 2, 133–142, (2022), doi: 10.35457/antivirus.v16i2.2445
15. H. Susanto, D. Yanto, N. Rusdiana, and D. T. Rahayu, *Prediksi Resiko Penyakit Jantung dan Pembuluh Darah Menggunakan Algoritma K-Nearest Neighbor (KNN)*, Joined Journal (Journal of Informatics Education), **3**, 1, (2020)
16. H. Susanto, D. Yanto, N. Rusdiana, and D. T. Rahayu, *Prediksi Resiko Penyakit Jantung dan Pembuluh Darah Menggunakan Algoritma K-Nearest Neighbor (KNN)*, Journal of Informatics Education, **3**,1, (2020)
17. S. O. Abdulsalam, *A Diabetic Prediction Model using Firefly Algorithm with K-Nearest Neighbor Classifier,* Int J Appl Inf Syst, **12**, (2022), Available: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes
18. M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, *Diabetes prediction using supervised machine learning*, in Procedia Computer Science, Elsevier B.V., 21–30, 2022, doi: 10.1016/j.procs.2022.12.107
19. A. Sumathi and S. Meganathan, *Ensemble classifier technique to predict gestational diabetes mellitus (GDM)*, Computer Systems Science and Engineering, **40**, 1, 313–325, (2022), doi: 10.32604/CSSE.2022.017484
20. E. M. S. Rochman, H. Suprajitno, I. Kamilah, A. Rachmad, & I. Santosa., *Tuberculosis classification using random forest with K-prototype as a method to overcome missing value*, Commun. Math. Biol. Neurosci., (2023)
21. A. Ali, M. Alrubei, L. F. M. Hassan, M. Al-Ja'afari, and S. Abdulwahed, *Diabetes classification based on KNN,* IIUM Engineering Journal, **21**, 1, 175–181, (2020), doi: 10.31436/iiumej.v21i1.1206
22. Rochman, E. M. S., Suprajitno, H., Rachmad, A., & Santosa, I., *Utilizing LSTM and K-NN for Anatomical Localization of Tuberculosis: A*

Solution for Incomplete Data, Mathematical Modelling of Engineering Problems, **10**, 4, (2023)