SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Information System Exploration and Research

https://shmpublisher.com/index.php/joiser
p-ISSN 2964-1160 | e-ISSN 2963-6361

JOISER
Journal of Information System Exploration and Research
Vol 01, No 01, January 2023

The Influence of Determining the K-Value on Improving the Diabetes Classification Model using the K-NN Algorithm

Nanda Putri Korina[1*], Budi Prasetiyo[2], Ade Anggian Hakim[3], M Rivaldi Ali Septian[4]
[1,2]Department of Computer Science, Universitas Negeri Semarang, Indonesia
[3]Department of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia
[4]Department of Electronic Engineering, Ming Chi University of Technology, Taiwan

DOI: https://doi.org/10.52465/joiser.v2i2.344
Received 22 April 2024; Accepted 26 July 2024; Available online 27 July 2024

Article Info

Keywords:
Diabetes classification;
K-NN algorithm;
Performance evaluation;
K-value

Abstract

Diabetes mellitus is still an important health problem globally, so it requires an efficient classification model to help determine a patient's diagnosis. This study aims to determine the K-value on the accuracy performance of the diabetes classification model using the K-Nearest Neighbors (K-NN) algorithm. This research utilizes a simulated dataset generated through interaction with ChatGPT, we investigate various K-values in the K-NN model and assess its accuracy using a confusion matrix. Based on experiments, we found that the K-NN classification model with a K=6 obtained an optimal accuracy of 97.62%. Thus, our findings highlight the important role of selecting optimal K-values in improving the performance of diabetes classification models.

1. Introduction

The increase in blood glucose levels is a diagnostic sign of the chronic condition known as diabetes mellitus, which occurs when the body fails to produce or use enough insulin [1]. According to the official website of the World Health Organization (WHO), approximately 422 million people worldwide suffer from diabetes, with the majority residing in low- and middle-income countries, and 1.5 million deaths directly attributed to diabetes each year [2-4]. Meanwhile, according to the International Diabetes Federation (IDF), there are 537 million adults suffering from diabetes mellitus (DM), with 6.7 million deaths annually attributed to this disease [5]. Diabetes remains recorded as the most common cause of death worldwide, with mortality rates significantly increasing each year [6].

In an effort to address this issue, early detection and proper management become crucial. One approach that can be used is the development of classification models. Classification is a learning process aimed at determining the attributes of each object set. In some cases, classification is also referred to as the process of grouping data. Classification is also used to describe a dataset where each data type, whether nominal or binary, is included [7]. Classification tasks are well-known for their ability

[*] ***Corresponding Author:***

Nanda Putri Korina,
Department of Computer Science,
Universitas Negeri Semarang,
Semarang, Indonesia.
Email : nandaputri2202@students.unnes.ac.id

to handle missing attribute values and continuous and discrete data to predict the risk of diabetes in individuals based on relevant risk factors [8].

In this context, the K-NN (K-Nearest Neighbors) method has become one of the popular choices in developing diabetes classification models [9], [10]. K-NN, or K-Nearest Neighbors, is one of the popular non-parametric machine learning algorithms, meaning it does not assume a distribution. Its advantage lies in the highly flexible and non-linear decision boundary generated by the model [11]. The K-Nearest Neighbor (K-NN) algorithm is a lazy learning technique that belongs to the instance-based learning group. This algorithm is created by determining K groups of objects in the training data that are closest to the target in the new or test data. The algorithm requires a classification method to find expert information [12]. This algorithm assigns K-Nearest Neighbors based on the closest distance from the training data to the test data; after gathering the K-Nearest Neighbors, most of them are taken to make predictions for the test sample. The basic concept of K-NN is to find the nearest distance between the data to be evaluated and its K nearest neighbors in the training data [13], [14].

The K-Nearest Neighbor (K-NN) classification algorithm has been proposed by many researchers [12], [13], [14]. Some advantages of the K-NN method include simple training, fast, easy to understand, and effective when the size of the training data is large [15]. However, the success of the K-NN model depends heavily on selecting the optimal value of K and proper data division between training and testing data. With this background, this research aims to investigate the influence of the value of K and the data division ratio on the performance of the K-NN model in diabetes classification. Through this research, it is hoped that deeper insights can be obtained into how optimal K-values and proper data division can enhance the accuracy and effectiveness of diabetes classification models using the K-NN method.

2. Method

In this research, there are five stages to find the best K-value, starting with data collection, which then processes the raw data at the data preprocessing stage by removing missing values and balancing data classes using Synthetic Minority Oversampling (SMOTE) [16], after the data is ready it is processed into classification model by applying the K-NN classifier, then measuring the performance of the classification model using a confusion matrix, and the optimal K-value can be found at the best level of accuracy. In detail the research flow can be seen in Figure 1.

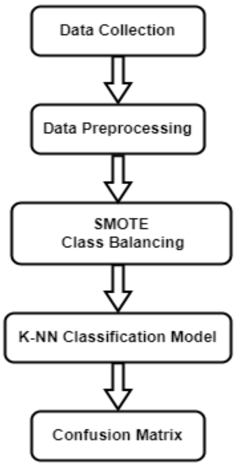


Figure 1. Research Framework

2.1 Data Collection

The data utilized in this study originated from meticulously crafted simulations aimed at mirroring the clinical characteristics commonly associated with diabetes mellitus. These simulations were meticulously devised based on the wealth of available medical knowledge, endeavoring to encapsulate the nuanced variations that may manifest within the broader patient population. Despite being synthetic in nature, this dataset was intricately designed to encompass vital features such as blood sugar levels, body weight, and blood pressure, all of which hold significant clinical relevance in both the diagnosis and management of diabetes. Additionally, the dataset incorporates supplementary variables such as blood pressure status and diabetes classification, thereby fostering a more comprehensive and nuanced analysis. It is noteworthy that the generation of this dataset was facilitated through an interactive session with the AI model, specifically prompted to create a dataset comprising 100-200 instances within the realm of healthcare, focusing on a classification problem with 2-3 distinct classes while incorporating elements of missing values and class imbalance.

As a result, the resulting dataset not only mirrors the inherent complexities and intricacies of real-world healthcare data but also encompasses deliberate instances of missing values and class imbalance, thereby enhancing its fidelity and applicability to real-world scenarios. Moreover, the careful consideration given to the number of samples within the dataset ensures the encapsulation of a diverse array of clinical parameters, thereby enriching the dataset with a breadth of variability. Furthermore, the deliberate calibration of the distribution of diabetes diagnosis classes reflects a pragmatic approach aimed at mirroring the realistic prevalence and distribution of normal, pre-diabetic, and diabetic cases within the broader patient population. Consequently, this meticulously curated dataset serves as a robust foundation for conducting comprehensive analyses pertaining to the classification and management of diabetes, offering valuable insights into the intricacies of machine learning models in healthcare applications.

2.2 Data Preprocessing

Preprocessing is conducted to avoid missing values to prevent noise in the K-Nearest Neighbor method. Additionally, the presence of data imbalance will affect the model's performance, thus requiring balancing with SMOTE. Based on the dataset that has been collected, there are three stages carried out including: (1) The "Replace missing value" operator is used to fill empty values with minimum, maximum, and mean values, (2) Class balancing utilize the SMOTE method, SMOTE has the advantage of not causing missing information, avoiding overfitting, building larger decision regions, and improving the accuracy of predicting minority classes. This makes this method suitable for use in this research [17]. (3) The split ratio to be used is 70:30 as recommended by most academic communities for data sizes between 100 and 1000000.

2.3 K-Nearest Neighbor

K-Nearest Neighbor (K-NN) is an instance-based learning method. This algorithm is also considered a lazy learning technique. The K-NN algorithm determines the distance values between test data and training data based on the smallest values of nearest neighbor distances. A classification system is required to be able to search for information [18].

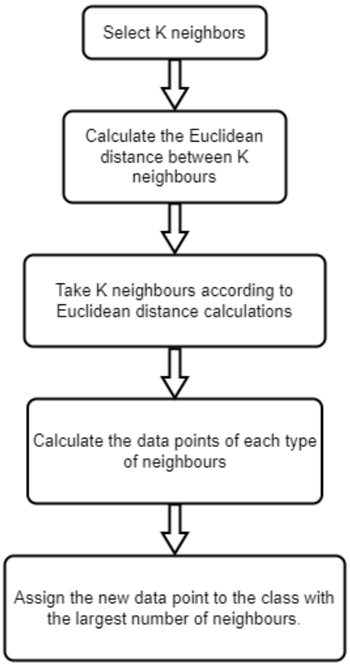

Select K neighbors
Calculate the Euclidean distance between K neighbours
Take K neighbours according to Euclidean distance calculations
Calculate the data points of each type of neighbours
Assign the new data point to the class with the largest number of neighbours.

Figure 2. Stage of K-NN Algorithm

2.4 Confusion Matrix

Confusion Matrix is a calculation method used to analyze the quality of a classification model in identifying tuples from existing classes [19]. The value in the confusion matrix can be calculated how the model performs using several metrics such as Accuracy, Precision, and Recall. The accuracy metric is the measure of how close the predicted values are to the actual values, which is calculated based on Equation (1). Meanwhile, the precision refers to the ratio of relevant items selected to all selected items. It represents the alignment between information requests and their corresponding responses, can be seen in Equation (2). Furthermore, the recall represents the ratio of relevant items selected to the total number of relevant items present, shown in Equation (3).

Table 1. Confusion matrix

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

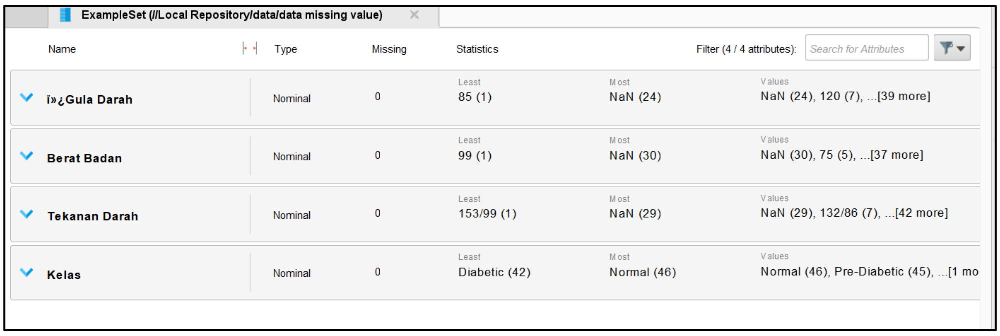
$$\mathbf{Accuracy} = \frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP} + \mathbf{FP} + \mathbf{FN} + \mathbf{TN}} \quad (1)$$

$$\mathbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}} \quad (2)$$

$$\mathbf{Recall} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} \quad (1)$$

3. Results and discussion

This study, collected diabetes record data with a total of 133 rows, with 3 attributes, and 1 class column. Each of these attributes is used to predict whether the patient has diabetes or not. The results of the classification model are assessed from how successful the model is in predicting into three classes stating the conditions "Normal", "Pre-diabetic", and "Diabetic".



ExampleSet (//Local Repository/data/data missing value)

Filter (4 / 4 attributes): Search for Attributes

Name	Type	Missing	Statistics: Least	Most	Values
ï»¿Gula Darah	Nominal	0	85 (1)	NaN (24)	NaN (24), 120 (7), ...[39 more]
Berat Badan	Nominal	0	99 (1)	NaN (30)	NaN (30), 75 (5), ...[37 more]
Tekanan Darah	Nominal	0	153/99 (1)	NaN (29)	NaN (29), 132/86 (7), ...[42 more]
Kelas	Nominal	0	Diabetic (42)	Normal (46)	Normal (46), Pre-Diabetic (45), ...[1 mo

Figure 3. Data Description

Because the data to be tested still contains missing values, it is necessary to perform cleaning first. This is because missing values can introduce bias and degrade the model's performance. With RapidMiner, treatment can be performed using the Replace Missing Value operator. The creation of a cleaning model is required to execute this program.

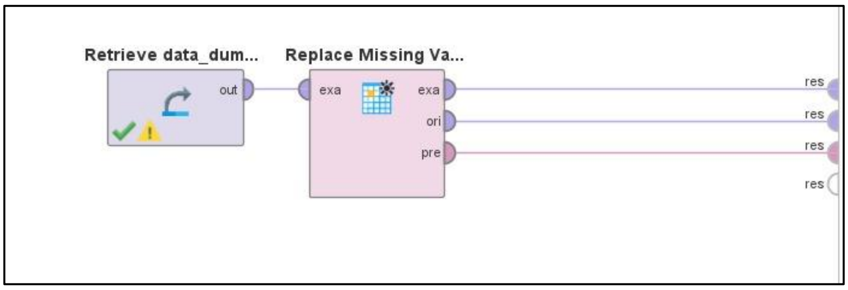
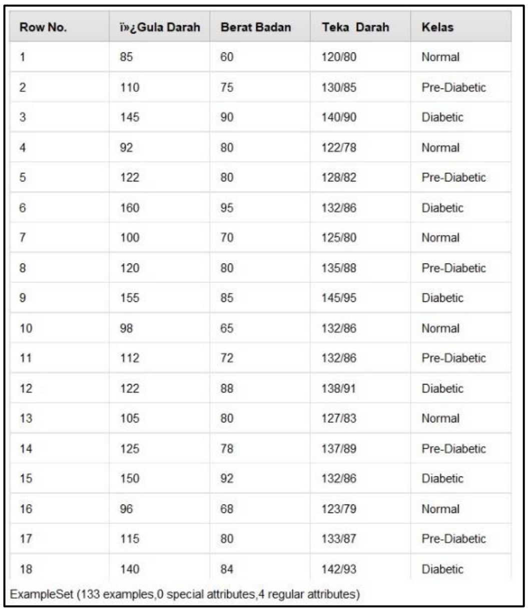


Figure 4. Design of Cleaning Model

The parameter to be used for cleaning missing values treatment is average. RapidMiner will fill empty data with the average value.



Row No.	ï»¿Gula Darah	Berat Badan	Teka Darah	Kelas
1	85	60	120/80	Normal
2	110	75	130/85	Pre-Diabetic
3	145	90	140/90	Diabetic
4	92	80	122/78	Normal
5	122	80	128/82	Pre-Diabetic
6	160	95	132/86	Diabetic
7	100	70	125/80	Normal
8	120	80	135/88	Pre-Diabetic
9	155	85	145/95	Diabetic
10	98	65	132/86	Normal
11	112	72	132/86	Pre-Diabetic
12	122	88	138/91	Diabetic
13	105	80	127/83	Normal
14	125	78	137/89	Pre-Diabetic
15	150	92	132/86	Diabetic
16	96	68	123/79	Normal
17	115	80	133/87	Pre-Diabetic
18	140	84	142/93	Diabetic

ExampleSet (133 examples,0 special attributes,4 regular attributes)

Figure 5. Data without missing value

After the data cleaning process from missing values, it turns out that the dataset to be tested has data imbalance. This will decrease the accuracy of the model's performance and complicate the process of finding the right K-value, thus balancing is needed. To handle the imbalanced data, this study applies the Synthetic Minority Oversampling Technique (SMOTE) method. Where the initial dataset consists of 133 rows, with the "normal" class consisting of 46 data, the "Pre-Diabetic" class consisting of 45 data, and the "Diabetic" class consisting of 42 data. In detail, the graph of the number of classes is presented in Figure 6.

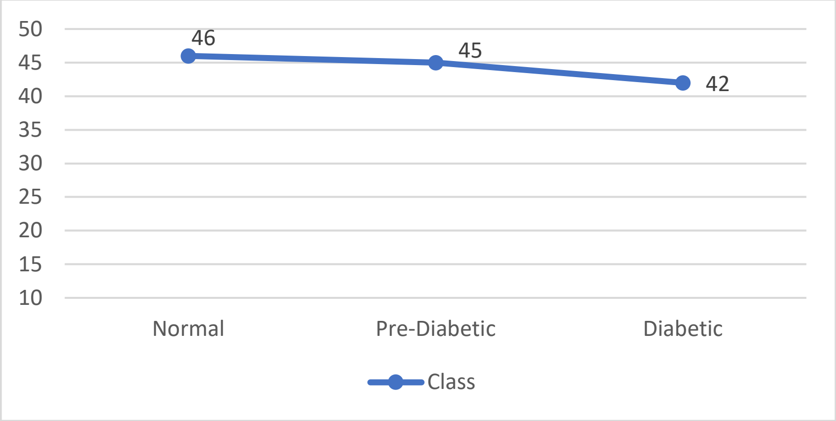
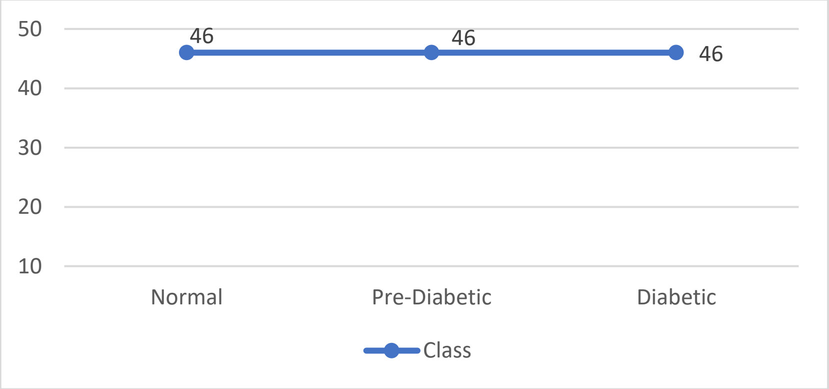


Figure 6. Number of Data Classes

After the cleaning process is complete, it is known that the data class has an unbalanced number. This will reduce the accuracy of the model's performance and complicate the process of finding the right K-value, so balancing is needed. The data class balancing process has been carried out by applying the SMOTE method, and produces a balanced number of each class. Thus, the total data is 138 records. Thus, the total data is 138 records, as shown in Figure 7.



Gambar 7. Number of data classes after balancing using SMOTE

After balancing is performed, the initially imbalanced data becomes balanced. This is necessary when the data is extremely imbalanced. With this, the data is ready to be split and fed into the model using K-NN classifier, and the model design can be shown in Figure 8.

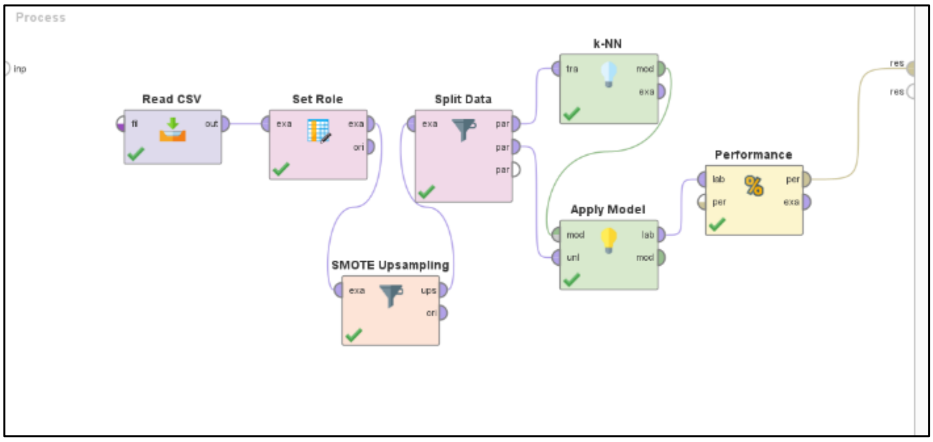


Figure 8. K-Nearest Neighbor Model

To proceed to the classification stage, the data is divided using the split data technique with a ratio of 70:30. As much as 70% of the data will go to the K-NN operator for training, while the remaining 30% will go directly to testing. The K-value is determined in the K-NN operator. In this test, the K-values that will be evaluated are 5, 6, and 7.

Table 2. Accuracy results of each K-value

K-Value	Accuracy	Recall	Precision
5	95.24%	95.24%	95.24%
6	**97.62%**	97.62%	97.62%
7	95.24%	95.24%	95.24%

Based on Table 2, it can be seen that from each tested value, the value of K = 6 has the highest accuracy, precision, and recall, which is 97.62%. Meanwhile, the values of K = 6 and K = 7 have the same accuracy, precision, and recall values, which are 95.24%.

4. Conclusion

This study aimed to explore the influence of determining the value of K on improving a diabetes classification model using the K-NN algorithm. Diabetes mellitus presents a significant global health challenge, necessitating efficient classification models for diagnosis and management. Through the utilization of a simulated dataset generated with the assistance of ChatGPT, various K-values were evaluated within the K-NN model, and their accuracy was assessed using relevant evaluation metrics. The main findings of this research indicate that selecting an optimal K-value has a significant impact on the performance of diabetes classification models. Preprocessing steps, such as handling missing values and balancing data using SMOTE, also proved crucial in enhancing model accuracy. Furthermore, the results demonstrate the effectiveness of the K-NN algorithm in classifying diabetes, with certain K values yielding higher accuracy than others. In this context, the best-performing K-value was determined to be K = 6, achieving an accuracy of 97.62%. These findings provide valuable insights into optimizing the application of the K-NN algorithm in diabetes classification, thereby advancing the field of medical informatics.

References

[1] M. Dwivedi and A. R. Pandey, “Diabetes mellitus and its treatment: an overview,” *J Adv Pharmacol*, vol. 1, no. 1, pp. 48–58, 2020.

[2] O. F. Offu, “A Systematic Review of the Prevalence and Treatment of Type 2 Diabetes in Nigeria.”.

[3] R. A. Afaya, “Self-management of diabetes among type 2 diabetes mellitus patients attending diabetes clinics in selected hospitals in the Tamale Metropolis, Northern Region, Ghana.,” 2021.

[4] O. P. Olatidoye, “Diabetic Care Center and Nutrition/Dietetics in Nigeria,” in *Medical Entrepreneurship*, Springer, 2023, pp. 287–310.

[5] A. D. Susanto and N. A. Kusumastuti, “Pendidikan Kesehatan Diabetes Melitus Di Ruangan Mahoni Rumah Sakit Umum Daerah Pakuhaji,” *Gudang Jurnal Pengabdian Masyarakat*, vol. 2, no. 1, pp. 81–86, 2024.

[6] G. Sanhaji, A. Febrianti, and H. Hidayat, “Aplikasi DIATECT Untuk Prediksi Penyakit Diabetes Menggunakan SVM Berbasis Web,” *Jurnal Tekno Kompak*, vol. 18, no. 1, pp. 150–163, 2024.

[7] G. Obaido *et al.*, “An interpretable machine learning approach for hepatitis b diagnosis,” *Applied sciences*, vol. 12, no. 21, p. 11127, 2022.

[8] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, “Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus,” *Multimed Syst*, pp. 1–19, 2022.

[9] S. Peerbasha, Y. M. Iqbal, K. P. Praveen, M. M. Surputheen, and A. S. Raja, “Diabetes Prediction using Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Logistic

Regression Classifiers," *JOURNAL OF ADVANCED APPLIED SCIENTIFIC RESEARCH*, vol. 5, no. 4, pp. 42–54, 2023.

[10] S. C. Gupta and N. Goel, "Enhancement of performance of K-nearest neighbors classifiers for the prediction of diabetes using feature selection method," in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, IEEE, 2020, pp. 681–686.

[11] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, p. 100071, 2022.

[12] N. Gharaei, W. Ismail, C. Grosan, and R. Hendradi, "Optimizing the setting of medical interactive rehabilitation assistant platform to improve the performance of the patients: A case study," *Artif Intell Med*, vol. 120, p. 102151, 2021.

[13] N. Ali, D. Neagu, and P. Trundle, "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets," *SN Appl Sci*, vol. 1, pp. 1–15, 2019.

[14] A. A. Nababan, M. Khairi, and B. S. Harahap, "Implementation of K-Nearest Neighbors (KNN) algorithm in classification of data water quality," *Jurnal Mantik*, vol. 6, no. 1, pp. 30–35, 2022.

[15] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indonesian Journal of Data and Science*, vol. 1, no. 2, pp. 39–43, 2020.

[16] O. Heranova, "Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring," *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 1, no. 10, pp. 10–12, 2021.

[17] E. Sutoyo and M. A. Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 6, no. 3, p. 379, 2020, doi: 10.26418/jp.v6i3.42896.

[18] B. S. Zemi, "Penerapan Algoritma Dijsktra," no. November, 2016.

[19] C. S. R. I. Murtono *et al.*, "Model Klasifikasi Potensi Penyakit Diabetes Mellitus Menggunakan Metode K-Nearest Neighbor (K-Nn) Penyakit Diabetes Mellitus," *Skripsi*, 2022.