International Journal of Electrical, Energy and Power System Engineering (IJEEPSE)
e-ISSN: 2654-4644
Vol. 6, No. 2, pp. 134-139, June 2023
Accredited by KEMENDIKBUDRISTEK, No. 230/E/KPT/2022
http://www.ijeepse.ejournal.unri.ac.id

Application Of Machine Learning K-Nearest Neighbour Algorithm To Predict Diabetes

Jack Billie Chandra*
Department of Information Technology
Institut Bisnis dan Teknologi Pelita Indonesia
Pekanbaru, Indonesia
jack.bc@student.pelitaindonesia.ac.id

Dewi Nasien
Department of Information Technology
Institut Bisnis dan Teknologi Pelita Indonesia
Pekanbaru, Indonesia
dewinasien@lecturer.pelitaindonesia.ac.id

**Corresponding author: jack.bc@student.pelitaindonesia.ac.id*

***Abstract*--- Diabetes is a chronic disease characterized by high blood sugar (glucose) levels or above abnormal values. This can occur when the body is no longer able to absorb glucose properly or when the intake of glucose is higher than needed. Glucose is the main energy source for the cells of the human body. Glucose that accumulates over the long term in the body can lead to complications and more serious and life-threatening diseases. As a result, patients with diabetes must be predicted prior to the onset of disease complications. Machine learning is one of the branches of artificial intelligence that can be used to provide predictive value to datasets of diabetic patients. The tested dataset has 390 observations with data on cholesterol levels, glucose, HDL cholesterol, cholesterol ratio, age, gender, blood pressure, BMI, waist and hip width with its ratio, and the patient's height and weight as variables. Predictions are applied using the K-Nearest Neighbor method, which shows an accuracy of 93.58% with a k value of 3, using 20% of all data as test data.**

Keywords—Diabetes, K-Nearest Neighbor, Prediction, Machine Learning

I. INTRODUCTION

Nowadays, technology is developing more rapidly and providing more and more benefits to human life. One of the benefits provided is computer technology, which has the ability to implement a human's way of thinking into a system on a computer. One of them is a machine-learning system that is used to detect or predict. Diabetes is a chronic disease that can be characterized by abnormally high levels of glucose (blood sugar). The people suffering from diabetes, their body is unable to properly process food for use as energy. The pancreas make a hormone called 'Insulin' helps glucose to penetrate into the cells of the Body, at times, the body doesn't make enough or any insulin. As a result, the glucose (or sugar) stays in the blood and an over a time period it causes health problems [1]. Diabetes is one of the most dangerous and deadly diseases in Indonesia, after stroke and coronary heart disease. Early prediction of diabetes risk is needed for early treatment of this disease. According to Sidartawan Soegondo, Indonesia is the fourth country in the world with the highest number of diabetics, which has increased to 14 million people. This is based on a report from the World Health Organization (WHO), where the number of people with diabetes in Indonesia in 2000 was 8.4 million, after India (31.7 million), China (20.8 million), and the United States (17.7 million). For people with diabetes worldwide, the WHO reports that there are more than 143 million sufferers, and this number is projected to double in prevalence by 2030 [2], and 77% of them occur in developing countries [3].

The increase in diabetes cases is due to the delay in establishing a diagnosis of the disease. The patient had died from complications before the diagnosis was made. The cause of the delay in establishing the diagnosis is the variety of factors that influence the existing choices. Therefore, we need a prediction that can be a tool in determining whether a person has diabetes mellitus or not. Disease is caused by people who combine excessive physical activity with a diet high in calories and fat that lacks fiber. Identification of diabetes is needed as a prevention strategy. By utilizing a data mining approach, it is possible to extract previously unknown information [4]. It is a great challenge for the healthcare organizations to provide cost-effective and high-quality clinical care for patients. This can be done only with the analyses of large healthcare database to extract the knowledge of disease and to make decisions. This is an important application in case of major diseases such as heart disease, cancer and diabetes [5]. The diagnosis of diabetes is very important; there are so many techniques in Machine Learning that can be effectively used for the prediction and diagnosis of diabetes disease. These algorithms in Machine Learning prove to be cost-effective and time saving for diabetic patients [6]. Therefore, machine learning algorithms are now used to identify and diagnose diseases in order to minimize the death risk and improve a patient's health status, as machine learning contributes to specific decisions [7].

II. METHODOLOGY

A. Machine Learning

Machine learning is a branch of computer science that examines how a machine can solve problems without being explicitly programmed [8]. Peter Harington (2012) describes several machine learning performance flows, namely:

- Collect data, in the form of Excel, Ms Access, Text Files and so on.

Received: April 9, 2023 | Revised: May 10, 2023 | Accepted: June 1, 2023

- Prepare the data, by determining the quality of the data and then taking steps to correct problems such as data loss.
- Train a model with data prepared into two parts, namely training data used for model development and test data used as a reference.
- Evaluating the model, by determining the provisions in the selection of algorithms based on the test results.
- Improving performance, involves choosing a different model or introducing more variables to increase efficiency.

B. Data Mining

Data mining is the process of looking for interesting patterns or information in selected data using certain techniques or methods. Techniques, methods, or algorithms in data mining vary widely [9]. According to Rerun at 2018, Data mining has several stages, with an explanation of each stage in the following:

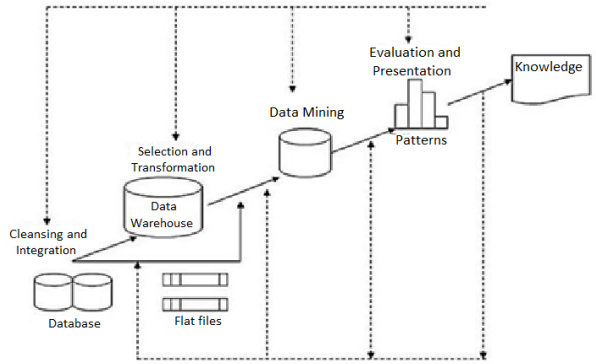


Fig. 1. Data Mining [10]

- Data Cleaning (to remove inconsistent data noise).
- Data Integration (split data sources can be unified).
- Data Selection (data relevant to the analysis task is returned to the database).
- Data Transformation (data changed or merged into the right form for mining with summary performance or operation aggression).
- Data Mining (essential process where intelligent methods are used to extract data patterns).
- Pattern Evolution (to identify really interesting patterns that represent knowledge based on some interesting action).
- Knowledge Presentation (where visualization techniques and knowledge images are used to provide the user with mined knowledge).

C. K-Nearest Neighbor

The K-Nearest Neighbor (KNN) method is a method of finding the shortest distance between the data to be evaluated and the closest K Neighbors in the training data. This technique belongs to the nonparametric classification group. This technique is very simple and easy to implement. Similar to the clustering technique, grouping new data based on the distance of the new data to some other data or its nearest neighbors [11].

D. Confusion Matrix

The confusion matrix is a method that is usually used to perform accuracy calculations on data mining concepts. The confusion matrix is illustrated by a table which states the amount of test data that is correctly classified and the amount of test data that is misclassified [12]. Accuracy is the comparison between the data that is classified correctly and the entire data. The accuracy value can be obtained from the following equation [13] :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \; x \; 100\% \qquad (1)$$

Precision is defined as the ratio of the selected relevant items to all selected items. Precision can be obtained by using the following equation [13] :

$$Precision = \frac{TP}{TP+FP} \; x \; 100\% \qquad (2)$$

Recall is defined as the ratio of the selected relevant items to the total number of available relevant items. Recall can be obtained from the following equation [13]:

$$Recall = \frac{TP}{TP+FN} \; x \; 100\% \qquad (3)$$

Errors are cases that are incorrectly identified in a number of data, in order to discover how much the error rate is in the system used. The percentage error can be calculated using the following equation [13]:

$$Error = \frac{FP}{TP} \; x \; 100\% \qquad (4)$$

E. Method

The method used is the K-Nearest Neighbor Algorithm method to classify new patient data to predict whether the patient had diabetes or not. The following are the stages in the research, as shown in Fig. 2. The first stage is to find the Euclidean Distance value of each of all the training data. Calculations are performed until all the training data have known Euclidean Distance values. The second stage is to determine the value of k that will be used for class classification. By determining the value of k, a number of training data can be taken into account as much as the number of k values. The training data taken are training data that have the closest Euclidean Distance value to the sample data being tested.

The method to calculate the Euclidean Distance can be represented as follows:

$$dis\ (x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2} \qquad (5)$$

X_1: training data, X_2: test data, I: data variable, dis: distance, n: data dimension

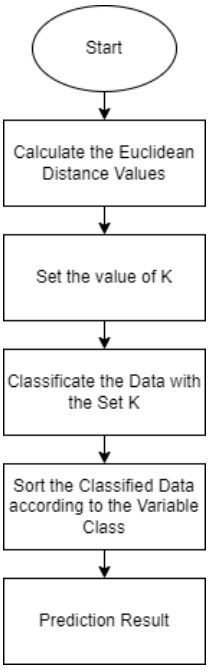


Fig. 2. Research Methodology

The third stage is classifying the training data based on the value of k. After obtaining training data samples that are included in the k value, the training data can be separated according to their classification class, namely diabetes or no diabetes. The fourth stage is to calculate the results of the number of class variable classifications from all training data that are included in the k value. At this stage, it will be calculated how much training data is included in the Diabetes classification and how much training data is included in No diabetes. Each class of classification will be counted in order for the next stage to draw conclusions.

The final stage is drawing conclusions. The test data will be compared with the training data. If the number of diabetes classifications in the training data is greater than the number of no diabetes classifications, it can be concluded that the test data is included in the Diabetes classification. If the number of no diabetes classification is more dominant, then the test data is classified into the classification no diabetes.

III. Result And Discussion

A. Dataset

TABLE I. Raw Dataset

patient_number	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	gender	height	weight	bmi	systolic_bp	diastolic_bp	waist	hip	waist_hip_ratio	diabetes
1	193	77	49	3,9	19	female	61	119	22,5	118	70	32	38	0,84	No diabetes
2	146	79	41	3,6	19	female	60	135	26,4	108	58	33	40	0,83	No diabetes
3	217	75	54	4	20	female	67	187	29,3	110	72	40	45	0,89	No diabetes
4	226	97	70	3,2	20	female	64	114	19,6	122	64	31	39	0,79	No diabetes
5	164	91	67	2,4	20	female	70	141	20,2	122	86	32	39	0,82	No diabetes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
231	188	88	51	3,7	50	female	61	147	27,8	160	66	34	41	0,83	No diabetes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
388	301	90	118	2,6	89	female	61	115	21,7	218	90	31	41	0,76	No diabetes
389	232	184	114	2	91	female	61	127	24	170	82	35	38	0,92	Diabetes
390	165	94	69	2,4	92	female	62	217	39,7	160	82	51	51	1	No diabetes

14 variables

As shown at Table II above, there are 390 patients' data with 14 variables. The raw dataset must be processed first in order to enable the calculation. The 'gender' and 'diabetes' variables are converted into an integer value. 'female' is converted into value of '0', 'male' to '1', 'diabetes' to '1' while 'no diabetes' turns into '0'. The following Table III is the result after preprocessing.

Received: April 9, 2023 | Revised: May 10, 2023 | Accepted: June 1, 2023

TABLE II. POST PREPROCESSED DATASET

patient_number	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	gender	height	weight	bmi	systolic_bp	diastolic_bp	waist	hip	waist_hip_ratio	diabetes
1	193	77	49	3,9	19	0	61	119	22,5	118	70	32	38	0,84	0
2	146	79	41	3,6	19	0	60	135	26,4	108	58	33	40	0,83	0
3	217	75	54	4	20	0	67	187	29,3	110	72	40	45	0,89	0
4	226	97	70	3,2	20	0	64	114	19,6	122	64	31	39	0,79	0
5	164	91	67	2,4	20	0	70	141	20,2	122	86	32	39	0,82	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
231	188	88	51	3,7	50	0	61	147	27,8	160	66	34	41	0,83	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
388	301	90	118	2,6	89	0	61	115	21,7	218	90	31	41	0,76	0
389	232	184	114	2	91	0	61	127	24	170	82	35	38	0,92	1
390	165	94	69	2,4	92	0	62	217	39,7	160	82	51	51	1	0

B. *Calculating Euclidean Distance*

Calculating the Euclidean Distance value of each training data based on the test data. The data must be split into training data and test data first, with a fixed ratio. Preferably with the training data's ratio higher than the test data. In this example, the data is split on 20% test data and 80% training data. Test data is chosen randomly by the machine learning. The machine will calculate the Euclidean Distance values of every training data based on the test data. The following Table III, as in Table IV, shows the results of calculating the Euclidean Distance from all training data for test data.

TABLE III. EUCLIDEAN DISTANCES

Patient_number	*distance*
1	60.96007
2	76.30839
3	78.39224
4	73.56964
5	61.39585
...	...
...	...
389	132.60972
390	91.33635

Distances are then sorted in order from the closest to the test data to the furthest, with the data of 255 being the closest to the test data while the data of 257 being the furthest among the training data, as shown below:

TABLE IV. SORTED DISTANCES

Patient_number	*Distance*	*Closest Distance Order*
255	15.94302355	1
200	28.01607396	2
176	31.98282195	3
371	32.44895838	4
165	35.56193049	5
...	...	...
...	...	...
243	290,655	385
257	304,4431	386

C. *K Value and Prediction*

After calculating the Euclidean Distances, the value of k must be assigned a value. The value is no less than 3 and if possible, use odd numbers for better performance. In this example, the value of k is assigned as 3, therefore 3 training data with the closest distance such as those of 255, 200, and 176 will be used for classification prediction. The following table shows the prediction:

TABLE V. K-NEAREST NEIGHBOR CLASSIFICATION

Patient Number	*cholesterol*	*glucose*	*...*	*...*	*waist_hip_ratio*	*diabetes*
255	185	84	...	...	0.83	No diabetes
200	177	87	...	...	0.85	No diabetes
176	191	81	...	...	0.86	No diabetes

As shown at Table VI, according to the training data in the range of the given k value, the test data is classified as 'no diabetes', since 'no diabetes' classifications are more dominant than the 'diabetes' classifications. And since the actual data of 231 is classified as 'no diabetes', the prediction proves to be accurate. Prediction processes are repeated until all test data are predicted, and the prediction result will be calculated by the confusion matrix accordingly.

D. *Algorithm Result*

The research shows different results by implementing several tests runs with different k values and split data ratios, as shown in the Table VII.

TABLE VI. ALGORTIHM PERFORMANCE RESULT

Test Run	*Test Data*	*Training Data*	*Accuracy*	*Best K Value*	*Precision*	*Recall*	*Error*
1	20%	80%	93.58%	3	100%	78%	6.4%
2	30%	70%	88.88%	3	91%	45%	11.11%
3	40%	60%	91.02%	17	94%	54%	8.97%

As shown in Table VII, the conclusion is that the K-Nearest Neighbor has the best prediction result on 20% - 80% split data ratio with three as the k value, whose accuracy of 93.58% is the highest accuracy score and its error rate 6.4% the lowest error rate.

E. Interface Discussion

Load Dataset Form View, this view is used to load the raw dataset for the machine learning to use. As shown in Fig. 3, the file must be in csv. format in order to run.

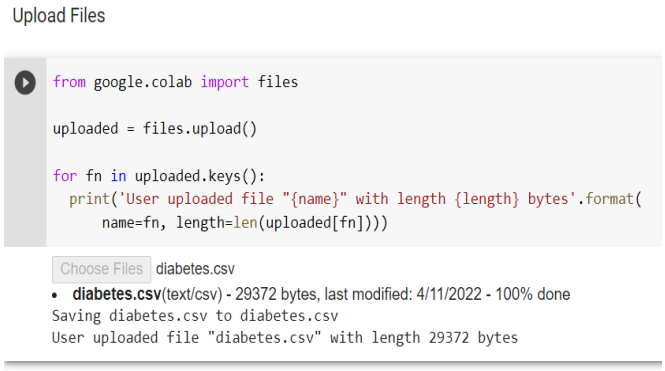
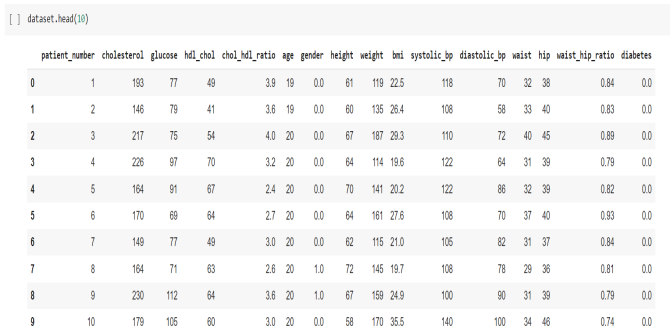


Fig. 3. Load Dataset Form View

Dataset View, this view shows the contents of the dataset as shown in Fig. 4. This view only serves to indicate that the dataset has been successfully uploaded.



	patient_number	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	gender	height	weight	bmi	systolic_bp	diastolic_bp	waist	hip	waist_hip_ratio	diabetes
0	1	193	77	49	3.9	19	0.0	61	119	22.5	118	70	32	38	0.84	0.0
1	2	146	79	41	3.6	19	0.0	60	135	26.4	108	58	33	40	0.83	0.0
2	3	217	75	54	4.0	20	0.0	67	187	29.3	110	72	40	45	0.89	0.0
3	4	226	97	70	3.2	20	0.0	64	114	19.6	122	64	31	39	0.79	0.0
4	5	164	91	67	2.4	20	0.0	70	141	20.2	122	86	32	39	0.82	0.0
5	6	170	69	64	2.7	20	0.0	64	161	27.6	108	70	37	40	0.93	0.0
6	7	149	77	49	3.0	20	0.0	62	115	21.0	105	82	31	37	0.84	0.0
7	8	164	71	63	2.6	20	1.0	72	145	19.7	108	78	29	36	0.81	0.0
8	9	230	112	64	3.6	20	1.0	67	159	24.9	100	90	31	39	0.79	0.0
9	10	179	105	60	3.0	20	0.0	58	170	35.5	140	100	34	46	0.74	0.0

Fig. 4. Dataset View

Training Test Split View, in this view, the user can set the training test split ratio by changing the value of the 'test-size'. The amounts of training and test data are displayed below the code box after running the codes, as shown in Fig. 5.

Fig. 5. Training Test Split View

K-Nearest Neighbor Model View, in this view the value of k can be changed for different prediction results as shown in Fig. 6.

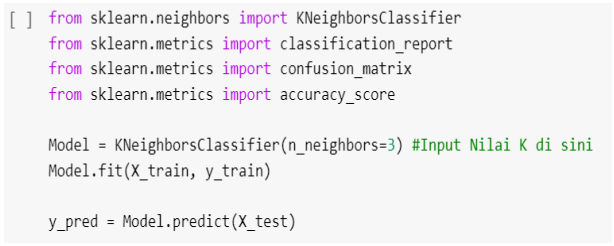


Fig. 6. K-Nearest Neighbor Model View

Confusion Matrix View, this view displays the confusion matrix table results. Confusion matrix and accuracy are displayed, as shown in Fig. 7.

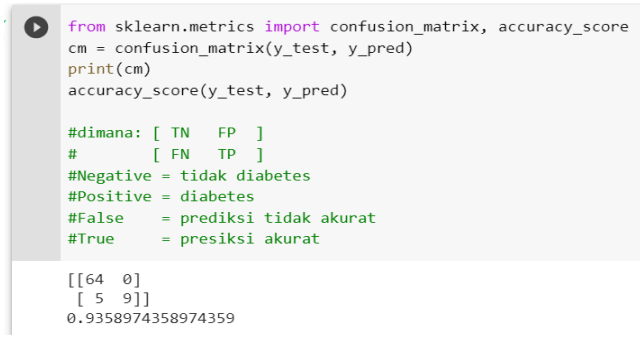


Fig. 7. Confusion Matrix View

Classification Report View, this view shows the performance result of the K-Nearest Neighbor in diabetes prediction. Precision, Recall, and Accuracy are displayed as shown in Fig. 8. Prediction View, in this view, a sample data can be entered to determine its prediction. The input can only be numbers, either an integer or float, but not a string. In case the value is a float type, the value must use a period (.), instead of a comma (,).

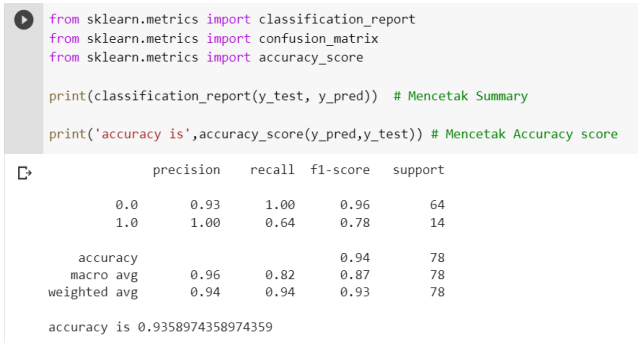


Fig. 8. Classification Report View

Error Rate View, this view displays graph of error rate and the value of k. Its purpose is to check which k value has the lowest error rate. Lower error rates provide better accuracy.

IV. CONCLUSION

The conclusion obtained based on the research conducted is that the K-Nearest Neighbor algorithm has a good performance result in predicting diabetes, with a fairly high accuracy of 93.58% and a fairly low probability of prediction error of 6.4%.

Received: April 9, 2023 | Revised: May 10, 2023 | Accepted: June 1, 2023

ACKNOWLEDGMENT

The authors would like to thank the Faculty of Science Computer, Institut Bisnis dan Teknologi Pelita Indonesia for the facilities that has been provided and for its support.

REFERENCES

[1] S. Kumar, "Detailed Analysis Of Classifiers For Prediction Of Diabetes," Vol. 11, No. 09, pp. 209–212, 2022.

[2] R. Saxena And S. Kumar Sharma Manali Gupta, "Role Of K-Nearest Neighbour In Detection Of Diabetes Mellitus," Turkish J. Comput. Math. Educ., Vol. 12, No. 10, Pp. 373–376, 2021.

[3] J. J. Pangaribuan, "Mendiagnosis Penyakit Diabetes Melitus Dengan Menggunakan Metode Extreme Learning Machine," J. Isd, Vol. 2, No. 2, Pp. 69–76, 2016.

[4] M. S. Mustafa And I. W. Simpen, "Implementasi Algoritma K-Nearest Neighbor (Knn) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba," Semin. Ilm. Sist. Inf. Dan Teknol. Inf., Vol. Viii, No. 1, pp. 1–10, 2019.

[5] P. C. Thirumal And N. Nagarajan, "Applying Average K Nearest Neighbour Algorithm To Detect Type-2 Diabetes," Aust. J. Basic Appl. Sci., Vol. 8, No. 7, Pp. 128–134, 2014.

[6] S. V. M And U. K, "Type 2 Diabetic Prediction Using Machine Learning Algorithm," Am. Sci. Res. J. Eng. Technol. Sci., Vol. 45, No. 1, pp. 299–307, 2018.

[7] M. Panda, D. P. Mishra, S. M. Patro, And S. R. Salkuti, "Prediction Of Diabetes Disease Using Machine Learning Algorithms," Iaes Int. J. Artif. Intell., Vol. 11, No. 1, pp. 284–290, 2022.

[8] M. Ula And A. Faridhatul Ulva, "Implementasi Machine Learning Dengan Model Case Based Reasoning Dalam Mendagnosa Gizi Buruk Pada Anak," J. Inform. Kaputama, Vol. 5, No. 2, pp. 333–339, 2021.

[9] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," Edik Inform., Vol. 2, No. 2, Pp. 213–219, 2017.

[10] R. R. Rerung, "Penerapan Data Mining Dengan Memanfaatkan Metode Association Rule Untuk Promosi Produk," J. Teknol. Rekayasa, Vol. 3, No. 1, P. 89, 2018.

[11] Yeni Kustiyahningsih And N. Syafa'ah, "Sistem Pendukung Keputusan Untuk Menentukan Jurusan Pada Siswa Sma Menggunakan Metode Knn Dan Smart," J. Istek, Vol. Vi, No. 1, pp. 40–42, 2013.

[12] M. F. Rahman, D. Alamsah, M. I. Darmawidjadja, And I. Nurma, "Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (Rbnn)," J. Inform., Vol. 11, no. 1, p. 36, 2017.

[13] B. P. Pratiwi And A. Silvia, "Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi Wsn Menggunakan Confusion Matrix", Vol. 6, No. 2, pp. 66–75, 2020.

BIOGRAPHIES OF AUTHORS

JACK BILLIE CHANDRA was born in Pekanbaru, Indonesia and he is a student from Faculty of Computer Science, Institut Bisnis dan Teknologi Pelita Indonesia Pekanbaru. He graduated in 2022. He also received his A.P in 2019 from the same institute.

DEWI NASIEN received her Ph.D. in 2012 and has worked at Universiti Teknologi Malaysia, Johor Bahru, Malaysia, from 2012 to 2016. She is currently a lecturer at a private university at Pelita Indonesia Institute of Business and Technology. Moreover, she is also an adjunct lecturer at several universities. Her areas of expertise include image processing, pattern recognition, machine learning, and soft computing.