COMPARISON OF DECISION TREE, NAÏVE BAYES, AND NEURAL NETWORK ALGORITHM FOR EARLY DETECTION OF DIABETES

Wisti Dwi Septiani[1*)] ; Marlina[2]

[1*), 2]Univesitas Bina Sarana Informatika
www.bsi.ac.id
[1*)] wisti.wst@bsi.ac.id, [2] marlina.mln@bsi.ac.id
(*) Corresponding Author

Abstract— *Diabetes mellitus is included in the top 3 most deadly diseases in Indonesia. Based on WHO data in 2013, diabetes contributed 6.5% to the death of the Indonesian population. Diabetes is a chronic disease characterized by high blood sugar (glucose) levels that exceed normal limits. In the health sector, historical medical data can be processed to extract new information and can be used for decision-making processes such as disease prediction. This study aims to classify predictions for early detection of diabetes to obtain accurate results for decision-making. The data used are historical data on hospital disease patients in Sylhet, Bangladesh. The algorithms used are Decision Tree, Naive Bayes, and Neural Network. Then the three methods are compared using the Rapid miner tools. The measurement results are 95,96% accuracy with Decision Tree, 87,69% with Naive Bayes, and 61,54% with Neural Network. So that the best algorithm is obtained, namely the Decision Tree for predicting early detection of diabetes. The rule in the form of a decision tree generated from the Decision Tree is used for input or ideas for decision making in the health sector for diabetes.*

Keywords: *Data Mining, Classification, Prediction, Decision Tree, Naive Bayes, Neural Network, Diabetes*

Intisari— Diabetes melitus termasuk ke dalam 3 besar penyakit yang paling mematikan di Indonesia. Berdasarkan data WHO pada tahun 2013, diabetes menyumbang sebesar 6,5% pada kematian penduduk Indonesia. Diabetes merupakan penyakit kronis yang ditandai dengan tingginya kadar gula (glukosa) dalam darah yang melebihi batas normal. Pada bidang kesehatan, histori data medis dapat diolah untuk mengekstrak informasi baru dan dapat dimanfaatkan untuk proses pengambilan keputusan seperti prediksi penyakit. Penelitian ini bertujuan melakukan klasifikasi prediksi untuk deteksi dini penyakit diabetes sehingga didapatkan hasil yang akurat untuk pengambilan keputusan. Data yang digunakan adalah data riwayat pasien penyakit rumah sakit di Sylhet, Bangladesh. Algoritma yang digunakan adalah Decision Tree, Naive Bayes, dan Neural Network kemudian dilakukan komparasi terhadap ketiga metode tersebut menggunakan tools Rapid miner. Hasil pengukuran yaitu tingkat akurasi 95,96% dengan Decision Tree, 87,69% dengan Naive Bayes, dan 61,54% dengan Neural Network. Sehingga didapatkan algoritma terbaik yaitu Decision Tree untuk prediksi deteksi dini penyakit diabetes. Rule berupa pohon keputusan yang dihasilkan dari Decision Tree digunakan untuk masukan atau ide untuk pengambilan keputusan di bidang kesehatan untuk penyakit diabetes.

Kata Kunci*:* Data Mining, Klasifikasi, Prediksi, Decision Tree, Naive Bayes, Neural Network, Diabetes

INTRODUCTION

Diabetes is a disease when the body is unable to use sugar (or glucose), so there is too much sugar in the blood (hyperglycemia). There are three types of diabetes: type 1 (insulin-dependent), type 2 (non-insulin-dependent diabetes mellitus (NIDDM) or "adult-onset), and gestational diabetes mellitus (GDM). (Handayanna, Rinawati, Arisawati, & Dewi, 2017).

Diabetes mellitus is also known as sugar disease or diabetes, where sufferers experience chronic metabolic disorders characterized by increased blood sugar levels above normal values. (Efendi & Wibawa, 2018) and many people do not realize that they are already in a condition of diabetes which leads to complications (Apriliah, Kurniawan, Baydhowi, & Haryati, 2021), if blood sugar is not managed properly, the complications are coronary heart disease, stroke, and obesity(Argina, 2020). The International Diabetes Federation (IDF) in 2013 estimated that the number of people with diabetes in the world reached 382 million people and in Indonesia, the number of people with diabetes was quite high, which was around 12 million people in 2013, where this number turned out to be an increase compared to previous years. (Efendi & Wibawa, 2018).

Diabetes is one of the fastest-growing life-threatening chronic diseases that has affected 422

million people worldwide according to a report by the World Health Organization (WHO), in 2018 (Apriliah et al., 2021). Delay in the diagnosis of diabetes has led to an increasing number of diabetics (Putri, Irawan, & Rizky, 2021). Also, the increasing number of diabetics due to diabetes is known as a silent killer. This refers to many who do not realize that they have diabetes. Patients are usually known to have this disease when complications occur without any initial treatment. (Efendi & Wibawa, 2018). Therefore, awareness and efforts are needed to make early detection of diabetes by recognizing the symptoms that occur. Medical data records in diagnosing a disease can be used as material for extracting data and generating information to predict disease symptoms.

The branch of data mining science which is often also called knowledge in database (KDD) is often used interchangeably in explaining the process of extracting information in large databases but still related to each other. (Sunge, 2018). The purpose of this study is to design a model that can estimate the likelihood of developing diabetes in patients with maximum accuracy. Classification is a data mining technique that assigns categories to data sets for more accurate prediction and analysis (Apriliah et al., 2021).

The literature on data mining in terms of diabetes prediction has been carried out with several methods, namely diabetes prediction using the K-Nearest Neighbor Classification Method with the highest accuracy of 39% and the highest precision of 65% (Argina, 2020), using the comparison of ID3 and Naive Bayes algorithms and the best algorithm is Naive Bayes with the result 76% (Nurdiana & Algifari, 2020). Then, predicting diabetes using Naive Bayes with an accuracy of 72% and getting an increase to 74.74% with the addition of genetic algorithm feature selection (Handayanna et al., 2017), using the ID3 algorithm with the best attribute selection with an accuracy of 84.77% (Efendi & Wibawa, 2018), and predicting diabetes in the early stages using the Random Forest classification algorithm (Apriliah et al., 2021). There have not been many studies on the comparison of classification algorithms.

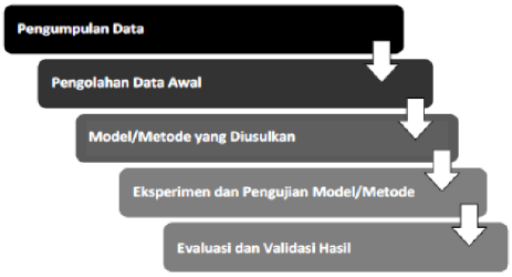
Data mining, which is called the mining process or data mining, will later produce valuable knowledge and can be implemented in decision support systems (Amalia, 2018), one of which is in the health sector for disease prediction.

According to Gorunescu, among the most popular data mining classification models are *Decision/Classification Trees, Bayesian Classifiers/Naïve Bayes Classifiers, Neural Networks, Statistical Analysis, Genetic Algorithms, Rough Sets, K-Nearest Neighbor Classifier, Rule-based Methods, Memory Based Reasoning, Support Vector Machines* (Septiani, 2020).

A decision tree is part of a classification algorithm that is widely used because it is easy to understand where tree branches are concluded in the form of classification (Septiani, 2020). Bayes classification which is also known as Naive Bayes is used to calculate the probability of a class, has the ability comparable to decision trees and Neural Networks. (Buani, 2018). Neural Networks have become an important standard tool for data mining and are used for pattern classification, prediction, and clustering tasks (Handayani, Nurlelah, Raharjo, & Ramdani, 2019). Therefore, in this experiment, a comparison of the data mining classification method with the proposed model is the Decision Tree, Naive Bayes, and Neural Network algorithms for predicting the early detection of diabetes.

MATERIALS AND METHODS

This study conducted an experiment in the form of a decision support system for predicting early detection of diabetes. To complete the research, a research design is made which is designed as a reference or research guideline and can be described as follows:



Source: (Septiani & Marlina, 2021)

Figure 1. Research Design

Figure 1 shows the design in the study which is a step from conducting experiments with an explanation of each step as follows:

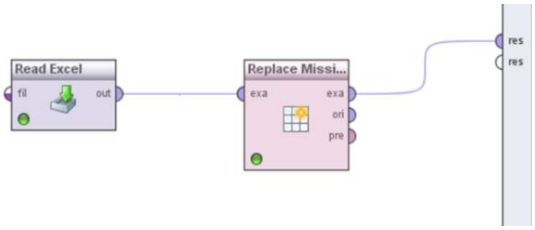
1. Data Collection

This study uses patient history data from a hospital in Sylhet, Bangladesh in the form of a diabetes dataset from the UCI Machine Learning Repository (University of California Irvine) with the web address: http://archives.uci.edu/ml/. The data collected was 520 data with attributes: *age, gender, polyuria, polydipsia, sudden_weight_loss, weakness, polyphagia, genital_thrush, visual_blurring, itching, irritability, delayed_healing, partial_paresis, muscle_stiffness, alopecia, obesity,* dan *class* (attribute prediction result).

2. Data Processing

The initial data processing is carried out: (1) Data validation, to identify and delete odd data, inconsistent data, and incomplete data. (2) Data

integration and transformation, to improve algorithm accuracy and efficiency. The data used in this study are categorical. (3) Data size reduction and discretization to obtain a dataset with fewer attributes and records but informative.



Source: (Septiani & Marlina, 2021)

Figure 2. *Replace Missing* Model

In Figure 2, the initial data processing is carried out in the form of a replace missing model which aims to eliminate duplication and anomalies or data inconsistencies. The results of the missing attributes performed and inconsistent data will be eliminated. From the data obtained as many as 520 records of the initial data processing process, the results of the attributes used are *age, gender, polyuria, polydipsia, sudden_weight_loss, weakness, polyphagia, genital_thrush, visual_blurring, itching, irritability, delayed_healing, partial_paresis, muscle_stiffness, alopecia, obesity,* dan *class* (attribute prediction result). Of the 520 records containing the classification, 320 data class "Positive" and 200 records class "Negative".

3. Proposed Model/Method

The proposed model is a data mining classification that has one of the roles of prediction, which is similar to classification and estimation (Prahartiwi & Dari, 2021). Pada In this study using the Diabetes dataset, the proposed model is the Decision Tree Algorithm, Naïve Bayes and Neural Network.

4. Experiment and Testing

Experiments were carried out by processing diabetes datasets using Decision Tree Algorithms, Naïve Bayes and Neural Networks. Then a comparison is made for the three modes. The tools used are Rapidminer. Algorithm comparisons were carried out to know the algorithm with the best performance and produce a better method for the dataset (Annisa, 2019), in this case, the diabetic dataset.

5. Evaluation and Valdiation

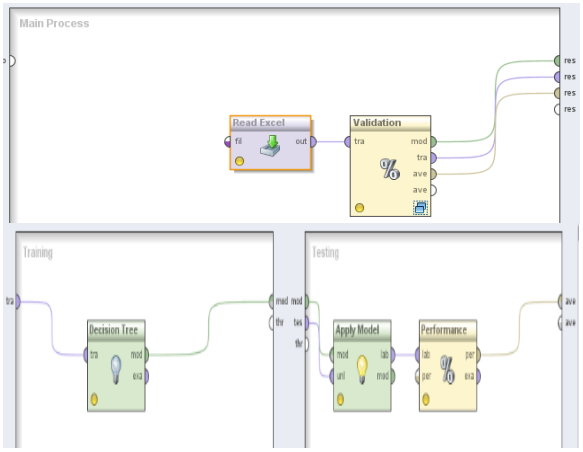
In the evaluation, observations and analysis of the results of model performance were carried out as measured by the Rapidminer tools (Nugraha, Shidiq, & Rahayu, 2019). Evaluation and validation were tested for accuracy with the Confusion Matrix.

RESULT AND DISCUSSION

Experiments and model testing are carried out to calculate and generate rules from the proposed model.

1. Decision Tree

The first experiment was carried out by testing the K-Fold Cross Validation model for the Decision Tree Algorithm as follows:



Source: (Septiani & Marlina, 2021)

Figure 3. Decision Tree Testing

In Figure 3, validation of the applied model, namely the Decision Tree, is carried out to test how the algorithm performs and produces the accuracy values in Table 1 below:

Table 1. *Decision Tree Algorithm Accuracy*

	True Positive	True Negative	Class Precision
Pred Positive	306	7	97.76%
Pred Negative	14	193	93.24%
Class Recall	97.56%	59.38%	
Accuracy		**95.96%**	

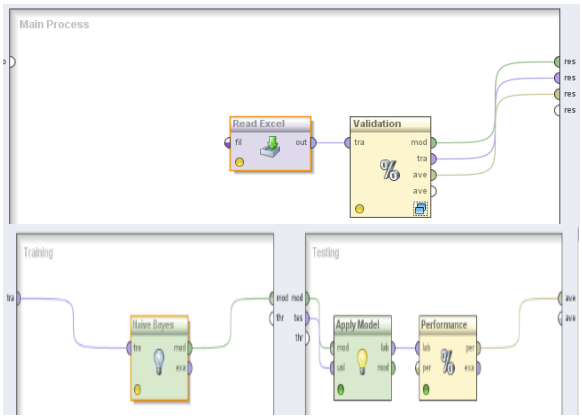
Source: (Septiani & Marlina, 2021)

Based on the accuracy results in Table 1, it can be concluded that of the 520 records tested with the Decision Tree, the results obtained were 313 records predicted to be "Positive" and 207 records "Negative" with the following criteria:

- From the 313 records predicted to be "Positive" the result is 306 records according to "Positive" and 7 data that are predicted to be "Positive" but turn out to be "Negative".
- From the 207 records predicted to be "Negative," the result is 193 records according to "Negative" and 14 data which are predicted to be "Negative" but turn out to be "Positive.

2. Naïve Bayes

The second experiment was carried out by testing the K-Fold Cross Validation model for the Naïve Bayes Algorithm as follows:



Source: (Septiani & Marlina, 2021)

Figure 4. Naïve Bayes Testing

In Figure 4, validation of the applied model, namely the Naïve Bayes, is carried out to test how the algorithm performs and produces the accuracy values in Table 2 below:

Table 2. Akurasi Algoritma *Naïve Bayes*

	True Positive	True Negative	Class Precision
Pred Positive	276	20	93.24%
Pred Negative	44	180	80.36%
Class Recall	97.56%	59.38%	
Accuracy		**87.69%**	

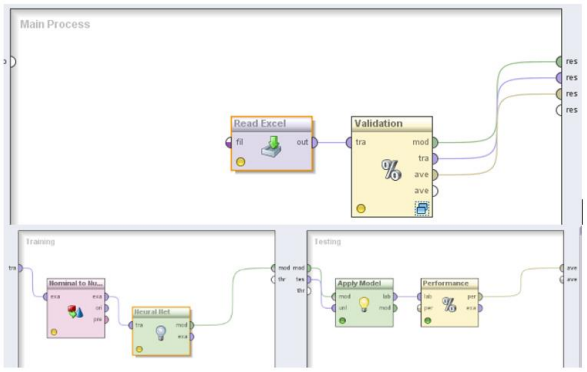
Source: (Septiani & Marlina, 2021)

Based on the accuracy results in Table 2, it can be concluded that of the 520 records tested with Naïve Bayes, 296 records were predicted to be "Positive" and 224 records "Negative" with the following criteria:

- From the 296 records predicted to be "Positive" the result is 276 records according to "Positive" and 20 data that are predicted to be "Positive" but turn out to be "Negative".
- From the 224 records predicted to be "Negative" the result is 180 records according to "Negative" and 40 data that are predicted to be "Negative" but turn out to be "Positive".

3. Neural Network

The second experiment was carried out by testing the K-Fold Cross Validation model for the Neural Network Algorithm as follows:



Source: (Septiani & Marlina, 2021)

Figure 5. Neural Network Testing

In Figure 5, validation of the applied model, namely the Neural Network, is carried out to test how the algorithm performs and produces the accuracy values in Table 3 below:

Table 3. Akurasi Algoritma *Neural Network*

	True Positive	True Negative	Class Precision
Pred Positive	320	200	61.54%
Pred Negative	0	0	0%
Class Recall	100.00%	0.00%	
Accuracy		**61.54%**	

Source: (Septiani & Marlina, 2021)

Based on the accuracy results in Table 3, it can be concluded that of the 520 records tested with Neural Network, 520 records were predicted to be "Positive" and 0 records "Negative" with the following criteria:

- From the 520 records predicted to be "Positive" the result is 320 records according to "Positive" and 200 data that are predicted to be "Positive" but turn out to be "Negative".
- From the 0 records predicted to be "Negative" the result is 0 records according to "Negative" and nothing is predicted "Positive".

Based on the analysis of algorithm testing that has been done, the results can be summarized in table 4 bellow:

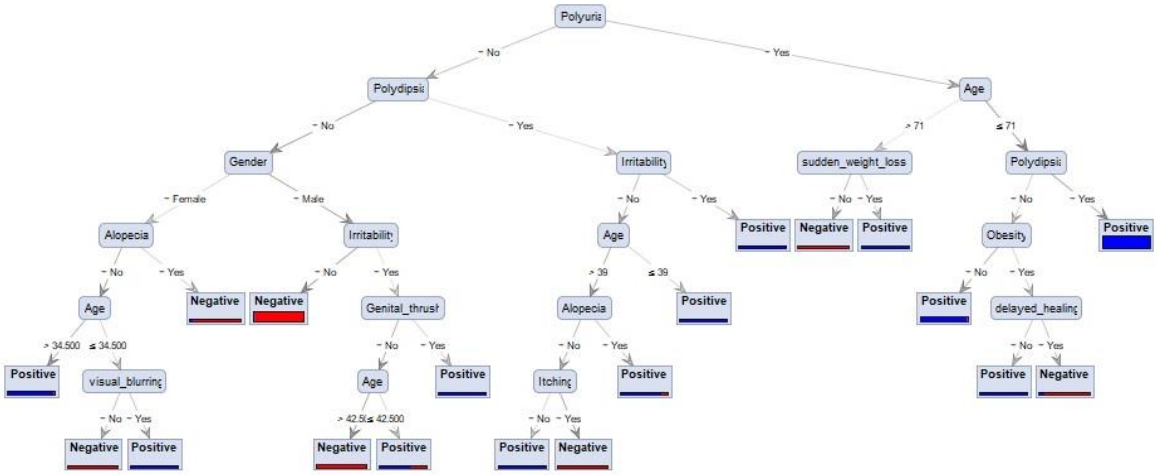
Table 4. Comparison of Accuracy

	Decision Tree	***Naïve Bayes***	***Neural Network***
Accuracy	95,96%	87,69%	61,54%

Source: (Septiani & Marlina, 2021)

Based on the data in Table 4 which is the result of the comparison of accuracy, it is concluded in this study that Decision Tree is the best algorithm of choice to be applied to this diabetes detection data.

In testing the Decision Tree model, a decision tree is obtained as shown in Figure 9 below:



Source: (Septiani & Marlina, 2021)

Figure 6. Decision Tree

Figure 6 shows the decision tree or rules generated by the Decision Tree algorithm. 19 rules were generated with the following conditions:

R1: IF Polyuria = NO and Polydipsia = NO and Gender = Female and Alopecia = NO and Age > 34,5 then "POSITIVE"
R2: IF Polyuria = NO and Polydipsia = NO and Gender = Female and Alopecia = NO and Age < 34,5 and Visual_blurring= NO then "NEGATIVE"
R3: IF Polyuria = NO and Polydipsia = NO and Gender = Female and Alopecia = NO and Age < 34,5 and Visual_blurring= YES then "POSITIVE"
R4: IF Polyuria = NO and Polydipsia = NO and Gender = Female and Alopecia = YES then "NEGATIVE"
R5: IF Polyuria = NO and Polydipsia = NO and Gender = Male and Irritability = NO then "NEGATIVE"
R6: IF Polyuria = NO and Polydipsia = NO and Gender = Male and Irritability = YES and Genital_trush = NO and Age > 42,5 then "NEGATIVE"
R7: IF Polyuria = NO and Polydipsia = NO and Gender = Male and Irritability = YES and Genital_trush = NO and Age < 42,5 then "POSITIVE"
R8: IF Polyuria = NO and Polydipsia = NO and Gender = Male and Irritability = YES and Genital_trush = YES then "POSITIVE"
R9: IF Polyuria = NO and Polydipsia = YES and Irritability = NO and Age > 39 and Alopecia = NO and Itching = NO then "POSITIVE"
R10: IF Polyuria = NO and Polydipsia = YES and Irritability = NO and Age > 39 and Alopecia = NO and Itching = YES then "NEGATIVE"
R11: IF Polyuria = NO and Polydipsia = YES and Irritability = NO and Age > 39 and Alopecia = YES then "POSITIVE"
R12: IF Polyuria = NO and Polydipsia = YES and Irritability = NO and Age < 39 then "POSITIVE"
R13: IF Polyuria = NO and Polydipsia = YES and Irritability = YES then "POSITIVE"
R14: IF Polyuria = YES and Age > 71 and Sudden_weight_loss = NO then "NEGATIVE"
R15: IF Polyuria = YES and Age > 71 and Sudden_weight_loss = YES then "POSITIVE"
R16: IF Polyuria = YES and Age < 71 and Polydispia = NO and Obesity = NO then "POSITIVE"
R17: IF Polyuria = YES and Age < 71 and Polydispia = NO and Obesity = YES and Delayed_healing = NO then "POSITIVE"
R18: IF Polyuria = YES and Age < 71 and Polydispia = NO and Obesity = YES and Delayed_healing = YES then "NEGATIVE"
R19: IF Polyuria = YES and Age < 71 and Polydispia = YES then "POSITIVE"

CONCLUSION

The results of testing the diabetes dataset in the form of accuracy values, namely the Decision algorithm of 96.96%, Naive Bayes of 87.69%, and Neural Network of 61.54%. This study produces an algorithm with the best accuracy value, namely Decision Tree, and also obtained a decision tree that produces 19 rules and can be used as decision making for early detection of diabetes. This research can be used as material for further research by using optimization or feature selection methods such as Genetic Algorithm, Adabbost, or

PSO.

REFERENCE

Amalia, H. (2018). Perbandingan Metode Data Mining SVM Dan NN Untuk Klasifikasi Penyakit Ginjal Kronis. *Jurnal PILAR Nusa Mandiri*, *14*(1), 1–6. Retrieved from http://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/80

Annisa, R. (2019). Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung. *Jurnal Teknik Informatika Kaputama (JTIK)*, *3*(1), 22–28. Retrieved from https://jurnal.kaputama.ac.id/index.php/JTIK/article/view/141/156

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Sistemasi: Jurnal Sistem Informasi*, *10*(1), 163–171. https://doi.org/10.32520/stmsi.v10i1.1129

Argina, A. M. (2020). Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes. *Indonesian Journal of Data and Science*, *1*(2), 29–33. https://doi.org/10.33096/ijodas.v1i2.11

Buani, D. C. P. (2018). Prediksi Penyakit Hepatitis Menggunakan Algoritma Naive Bayes Dengan Seleksi Fitur Algoritma Genetika. *Jurnal Evolusi*, *6*(2), 1–5. Retrieved from ejournal.bsi.ac.id

Efendi, M. S., & Wibawa, H. A. (2018). Prediksi Penyakit Diabetes Menggunakan Algoritma ID3 dengan Pemilihan Atribut Terbaik (Diabetes Prediction using ID3 Algorithm with Best Attribute Selection). *JUITA*, *VI*(1), 29–35.

Handayani, P., Nurlelah, E., Raharjo, M., & Ramdani, P. M. (2019). Prediksi Penyakit Liver Dengan Menggunakan Metode Decision Tree dan Neural Network. *CESS (Journal of Computer Engineering System and Science)*, *4*(1), 55–59. https://doi.org/10.24114/cess.v4i1.11528

Handayanna, F., Rinawati, Arisawati, E., & Dewi, L. S. (2017). Prediksi Penyakit Diabetes Menggunakan Naive Bayes dengan Optimasi Parameter Menggunakan Algoritma Genetika. *KNiST (Konferensi Nasional Ilmu Sosial & Teknologi)*, 71–76.

Nugraha, F. S., Shidiq, M. J., & Rahayu, S. (2019). Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara. *Jurnal Pilar Nusa Mandiri*, *15*(2), 149–156. https://doi.org/10.33480/pilar.v15i2.601

Nurdiana, N., & Algifari, A. (2020). *Studi Komparasi Algoritma ID3 dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus*. *6*(2), 18–23.

Prahartiwi, L. I., & Dari, W. (2021). *Komparasi Algoritma Naive Bayes, Decision Tree dan Support Vector Machine Untuk prediksi Penyakit Kanker Payudara*. *7*(1), 51–54. https://doi.org/10.31294/jtk.v4i2

Putri, S. U., Irawan, E., & Rizky, F. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5. *KESATRIA Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, *2*(1), 39–46.

Septiani, W. D. (2020). Optimasi Algoritma C4.5 Menggunakan Algoritma Genetika Untuk Prediksi Penyakit Hepatitis. *Inti Nusa Mandiri*, *15*(1), 59–64.

Septiani, W. D., & Marlina. (2021). *Laporan Akhir Penelitian*. Jakarta.

Sunge, A. S. (2018). Prediksi Kompetensi Karyawan Menggunakan Algoritma C4.5 (Studi Kasus : PT Hankook Tire Indonesia). *Seminar Nasional Teknologi Informasi Dan Komunikasi 2018 (SENTIKA 2018)*, 15–22.