Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12998 p-ISSN : 2541-044X

Comparison Of The C.45 And Naive Bayes Algorithms To Predict Diabetes

Alam[1)*], Divi Adiffia Freza Alana.[2)], Christina Juliane[3)]
[1,2,3)] School of Business and Information Technology, STMIK LIKMI Bandung – Indonesia
[1)] mr.alam0116@gmail.com, [2)] divifrezafreza@gmail.com, [3)]christina.juliane@likmi.ac.id

Submitted : Sep 11, 2023 | **Accepted** : Sep 22, 2023 | **Published** : Oct 1, 2023

Abstract: Diabetes mellitus is an urgent global health problem and has a major impact on people around the world. This disease is characterized by high levels of glucose in the blood due to disturbances in the production or use of the hormone insulin by the body. This study aims to carry out accurate early detection of diabetics so that they can be treated as soon as possible to reduce the risk of death and to compare the two algorithms that have the best level of accuracy. The algorithms used in this study are the C4.5 and Naïve Bayes Decision Tree Algorithms. The results of the experiments carried out in this study the Decision Tree Algorithm C4.5 and Naïve Bayes can be used in modeling the early detection of diabetes. The highest average accuracy results were obtained at 90.835% using the Decision Tree C4.5 Algorithm. As for the Naïve Bayes Algorithm, an average accuracy rate of 90.745% is obtained. The pruning process was carried out using the Decision Tree Algorithm C4.5, the accuracy performance increased to 91.30%. There were 18 patterns or rules for the early detection of diabetics from the built model. The determination of attributes, the number of attribute dimensions, and the number of samples greatly affect the performance of the model built.
Keywords: Decision Tree Algorithm C4,5, Data Mining, Diabetes, Naïve Bayes

INTRODUCTION

Diabetes Mellitus is an urgent global health problem and has a major impact on people around the world. This disease is characterized by high levels of glucose in the blood due to disturbances in the production or use of the hormone insulin by the body. Diabetes can lead to major complications like heart disease, renal failure, eye issues, and amputations if it is not properly managed. Given the high prevalence of diabetes and the risk of serious complications, early detection is very important in efforts to prevent and manage this disease (Kopitar et al., 2020). However, identification of diabetes in its early stages is often a challenge, as some of the early symptoms may be atypical or not noticeable at all.
Rapid advances in technology and the development of data analysis methods in recent years have provided new opportunities to address the challenges of early detection of diabetes. One approach that stands out is the use of data mining techniques or data exploration. Data Mining is the process of extracting useful knowledge and insights from large and complex data sets using various statistical, mathematical and artificial intelligence methods (Fajriati et al., 2023).

This study aims to utilize Data Mining in the early detection of diabetes by conducting a comparison between two popular algorithms, namely the C4.5 Decision Tree Algorithm (also known as C4.5) and Naïve Bayes. Both of these algorithms have been widely used in various data mining applications and have different characteristics. Algorithm C4.5 is a Decision Tree Algorithm that is very effective in forming decision trees based on the rules found in data (Anggraini et al., 2018). This decision tree can be used for data classification and can provide insights that are easy for users to interpret.

The Naïve Bayes algorithm is based on the Bayes theorem and is suitable for classifying data with naive features (Enriko et al., 2021). Despite simple assumptions, these algorithms often provide

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12998 p-ISSN : 2541-044X

competitive classification results with good performance in most cases. The accuracy of the C4.5 Decision Tree Algorithm and the Nave Bayes Algorithm is below 80%, according to a previous study by Khanam and Foo (2021) comparing machine learning algorithms for predicting diabetes using the Pima Indian Diabetes Data Set. Only 5 factors—sugar level, body mass index, insulin, pregnancy, and age—were considered in his study. Moreover, according to research by Kumari et al. (2021) on the classification and prediction of Diabetes Mellitus using 9 attributes, the C4.5 Decision Tree Algorithm and Nave Bayes only have an accuracy of 71.42% and 74.12%, respectively. In addition, research by Choubey et al. (2020) using the Pima Indian Diabetes Dataset in evaluating the performance of the classification algorithm also revealed that the dataset and 8 attributes were used. In this paper, we will discuss the experimental design used to compare the performance of the two algorithms in the early detection of diabetes. Use relevant data sets and run C.45 and Naïve Bayes Algorithms on the same dataset to compare accuracy, sensitivity, specificity, and other evaluation metrics. This research can provide valuable information about comparing the accuracy of the C.45 and Naïve Bayes algorithms in detecting diabetes. And guides health professionals in choosing the most appropriate approach in practical application. It is hoped that the research findings will help the growth of scientific applications and data mining in the healthcare industry.

METHOD

This research will focus on the application of the Decision Tree Algorithm C4.5 and Naïve Bayes in classifying the early detection of diabetes. The flow of research conducted can be seen in Figure 1.

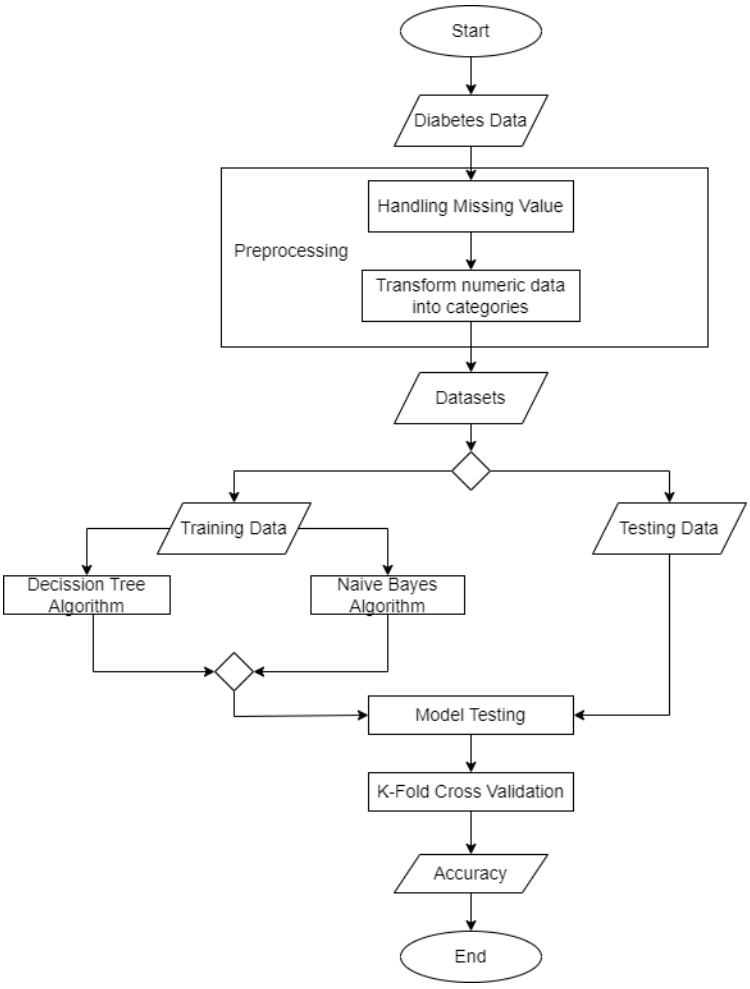

Figure 1. Research Flow
(Source: Primary)

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12998 p-ISSN : 2541-044X

Data Collection

The data used for current research activities is public data obtained from Kaggle. With 100,000 data samples, there are 9 attributes and 1 label. The dataset on Diabetes is described in Table 1.

Table 1. Diabetes Patients Data

No	Description	Type Data
1	Sex	Category
2	Age	Numeric
3	Hypertension	Numeric
4	Heart disease	Numeric
5	Smoking History	Category
6	BMI	Numeric
7	HBA1C levels	Numeric
8	Blood Glucose Levels	Numeric
9	Diabetes	Numeric

Missing Value

There are data on diabetics obtained from Kaggle that contain a missing value. To overcome this problem, researchers conducted data cleaning to clean up the missing value data (Tigga & Garg, 2020). After the data cleaning process was carried out, which totaled 100,000 data, only the remaining 42,064 data were obtained.

Data Transformation

The data is then transformed from numeric type to categorical. The reason is that the Decision Tree and Naïve Bayes Algorithms are algorithms for processing categorical data (Kumari et al., 2021). The categories used are taken from several references, namely:

a. Age category is taken from Permenkes No. 25 of 2016
b. The BMI category is taken from the information from the Kaggle dataset that we obtained.
c. The HBA1C category and blood sugar levels are taken from categories according to the Ministry of Health.

Meanwhile, for attribute categories with values of 1 and 0, we only initiate numbers into category criteria (Kopitar et al., 2020). The following changes in value to category criteria can be seen in Table 2.

Table 2. Category Criteria Data

No	Component Name	Type	Function
1	Age	>= 60 >= 45 >= 19 >= 10 >= 6 >= 5 >= 1	Elderly Pre Elderly Mature Teenager Children Preschool Children Toddler
2	Hypertension	1 0	Yes No
3	Heart disease	1 0	Yes No
4	BMI	>= 30 >= 25 >= 18,5 < 18,5	Obesity Overweight Normal Underweight
5	HBA1C Levels	>= 6,5 >= 5,7 >= 3,5	Diabetic indications Pre Diabetes Normal

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12998 p-ISSN : 2541-044X

6	Blood Glucose Levels	>= 126 >= 100 < 100	Diabetic indications Pre Diabetes Normal
7	Diabetes	1 0	Yes No

Determining the Amount of Training Data

We conducted several experiments by determining the amount of training data processed, to obtain the best accuracy results. The stages of the amount of data processed start from 4 data training scenarios, namely: 25%, 50%, 75%, and 100% of the entire dataset (Singh & Yassine, 2018). The distribution of the number of datasets can be seen in Table 3.

Table 3. Determining the Amount of Training Data

No	Total Data	Data retrieved	Amount of data
1.	42.064	25%	10.516
2.	42.064	50%	21.032
3.	42.064	75%	31.548
4.	42.064	100%	42.064

Algorithm Decision Tree

The Decision Tree algorithm is a tree structure that has root nodes, decision nodes, and leaf nodes (W. Chen et al., 2020)(Fersellia et al., 2023). Root nodes represent dataset features and are used to make decisions, while leaf nodes are the output of these decisions and do not have additional branches (Rianto et al., 2020).

Entropy Concept:

Entropy(S) = $\sum_{i-0}^{n} - Pi * log_2 Pi$ ……..……....….(1)

Information:

S : Set of Cases.
A : Features.
n : Number of partition S.
pi : Proportion of Si to S

Gain Concept:

Gain(S,A) = Entropy(S) - $\sum_{i-0}^{n} \frac{|Si|}{|S|} * Entropy(Si)$……..…….....(2)

Information:

S : Set of Cases.
A : Attribute.
n : Number of partitions attribute A.
|Si| : Number of cases on partition to index i
|S| : Number of cases in S

Algorithm Naïve Bayes

The Naive Bayes algorithm is a group of algorithms based on Bayesian theory (Fajriati et al., 2023). Bayes' theorem is a mathematical model based on statistics and probability (S. Chen et al., 2020). This algorithm is also used for probabilistic machine learning to solve various classification tasks. The advantages of the Naive Bayes algorithm are that it is fast in building models and making predictions, easy to use, and has a low error rate (W. Chen et al., 2020).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12998 p-ISSN : 2541-044X

Naïve Bayes Method Equation:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Information:

X : Data that has an unknown class
H : Hypothetical data is a specific category.
P(H\X) : Hypothesis H is based on the condition of situation X, which is known as the posteriori probability.
P(H) : Hypothesis H, or prior probability.
P(X\H) : The hypothetical situation H determines the probability of X.
P(X) : Probability X.

K-Fold Cross Validation

K-Fold Cross Validation is a useful technique because it provides more accurate estimates of model performance than splitting training data or conventional testing data (Shrinivasan et al., 2023). In addition, because all data is used for training and testing, the use of data becomes more efficient (Thabtah et al., 2020). In this approach, the data is divided into k equal-sized multiples, where k is the user-selected value. The model is tested on the remaining folds after being tested on the first k-1 folds. Each fold was utilized exactly once as the test set in this operation, which was repeated k times. To produce one estimate, the results are then averaged. In our test, we used a value of K=10 because a value of K=10 is more stable in the testing process (Wong & Yeh, n.d.).

RESULT

The results of data processing through the Data Preparation stage obtained datasets with categorical types totaling 42,064 data. After obtaining the dataset, modeling is carried out according to the selected algorithm by determining 4 data training scenarios, namely: 25%, 50%, 75%, and 100% (Singh & Yassine, 2018). The modeling process uses Rapidminer tools version 10.1. Then the results of an evaluation of the model are carried out using the K-Fold Cross Validation method together with the value K = 10 (Wong & Yeh, n.d.) so that you can see the accuracy value. obtained from the built model.

Experimental Results Dataset 25 % (10.516)

Table 4. Experimental Results Dataset 25%

Algorithm	Precision	Recall	AUC	Accuracy
DT C4.5	65,45%	40,63%	0,913%	90,97%
Naïve Bayes	61.60%	40,08 %	0,927%	90,84%

Table 4 shows that, when compared with the Naive Bayes Algorithm with a total of 10,526 data sets, the C4.5 Decision Tree Algorithm performs with the accuracy performance mentioned above. However, there isn't much difference between the two algorithms—only 0.13%. We ran experiments with a 50% data set to get better results.

Experimental Results Dataset 50 % (21.032)

Table 5. Experimental Results Dataset 50%

Algorithm	Precision	Recall	AUC	Accuracy
DT C4.5	63,03%	39,81%	0,904%	90,61%
Naïve Bayes	60.71%	49,30 %	0,929%	90,72%

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12998
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In table 5 using a 50% dataset, the Naïve Bayes Algorithm has a better accuracy performance than the C4.5 Decision Tree Algorithm. However, the results of the accuracy of the two algorithms were not better with a dataset accuracy of 25%, so experiments were carried out with a dataset of 75%.

Experimental Results Dataset 75 % (31.548)

Table 6. Experimental Results Dataset 75%

Algorithm	Precision	Recall	AUC	Accuracy
DT C4.5	62,20%	45,75%	0,905%	90,85%
Naïve Bayes	60,36%	48,48 %	0,930%	90,70%

It can be seen in Table 6. The C4.5 Decision Tree Algorithm is again superior in accuracy performance to the Naïve Bayes Algorithm in a 75% dataset. However, as is the case with using a 50% dataset, the accuracy results are not better when compared to performance when using a 25% dataset.

Experimental Results Dataset 100 % (42.064)

Table 7. Experimental Results Dataset 100%

Algorithm	Precision	Recall	AUC	Accuracy
DT C4.5	64,99%	40,53%	0,892%	90,91%
Naïve Bayes	60,49%	48,00 %	0,929%	90,72%

The last experiment was to use 100% of the data processing result dataset, totaling 42,064 data in Table 7. As a result, the C4.5 Decision Tree Algorithm outperforms the Nave Bayes Algorithm in terms of accuracy performance. However, the accuracy results are still not better with a 25% dataset. The accuracy graph can be seen in Figure 2.

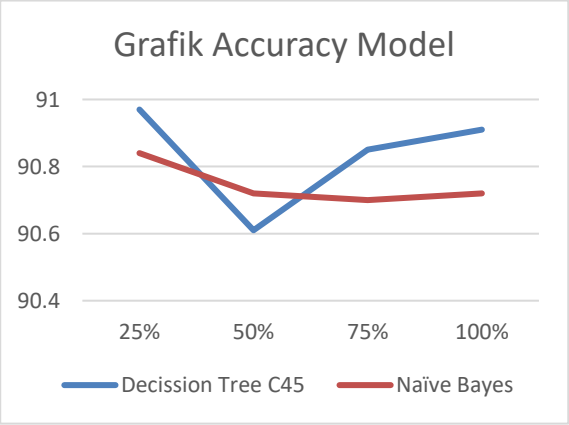

Figure 2. Graph of Experiment Results
(Source: Primary)

DISCUSSIONS

Based on the results of the modeling experiments carried out, using 4 data training scenarios, namely: 25%, 50%, 75%, and 100% (Singh & Yassine, 2018). The accuracy average is calculated. The results of calculating the average accuracy can be seen in Table 8.

Table 8. Average Accuracy

No.	Algorithm	Accuracy
1	DT C4.5	90,835 %
2	Naïve Bayes	90,745%

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12998
e-ISSN : 2541-2019
p-ISSN : 2541-044X

As observed in the table above, the average accuracy of the DT C4.5 algorithm is 90.835%, while the average accuracy of the Naive Bayes algorithm is 90.745%, which indicates that the DT C4.5 algorithm has the highest accuracy. Then we use a cropping step to increase the correctness of the created model. Decision Tree Algorithm C4.5 performs the pruning procedure with the best level of accuracy on 25% of the dataset. Decision trees are pruned to make them more efficient (Patra et al., 2021). After trimming, the accuracy performance increased to 91.30%. The results of the decision tree model using the C4.5 Decision Tree Algorithm are as follows.

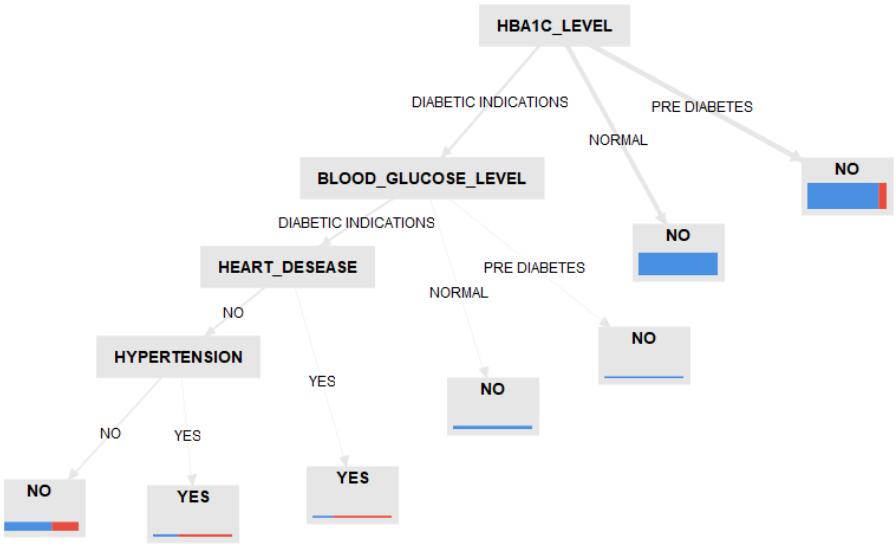

Figure 3. DT C45 Decision Tree Model
(Source: Primary)

From the description above, a pattern or rule is obtained which can be seen in Table 9.

Tabel 9. Pattern / Rules

No.	Algorithm	Accuracy
1	C1 = Normal	TD
2	C1 = Prediabetes	TD
3	C1 = Indication of Diabetes C2 = Pre Diabetes	TD
4	C1 = Indication of Diabetes C2 = Normal	TD
5	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = YES	D
6	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = YES	D
7	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = NO C5 = Normal	TD
8	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = NO C5 = Underweight	TD
9	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = NO C5 = Obesity C6 = Adolescent	TD
10	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = NO C5 = Obesity C6 = Pre-Elderly	D
11	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = NO C5 = Obesity C6 = Elderly	D
12	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = NO C5 = Obesity C6 = Adult	TD
13	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = No C5 = Overweight C6 = Adolescent	TD
14	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = No C5 = Overweight C6 = Pre-Elderly	TD

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12998 p-ISSN : 2541-044X

15	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = No C5 = Overweight C6 = Elderly	D
16	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = No C5 = Overweight C6 = Adult	TD
17	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = No C5 = Overweight C6 = Children	TD
18	C1 = Indication of Diabetes C2 = Indication of Diabetes C3 = No C4 = No C5 = Overweight C6 = Preschool Children	TD

Information:
C1 : HBA1C Levels
C2 : Blood Glucose Levels
C3 : Heart disease
C4 : Hypertension
C5 : BMI
C6 : Age
TD : No Diabetes
D : Diabetes

The accuracy of prior research on diabetes categorization using the C4.5 Decision Tree Algorithm and Nave Bayes is displayed in Table 10 as results, which shows that the performance of this algorithm is better in this study when compared to previous studies.

Table 10. Previous Research Results

No	Author	Algorithm	Accuracy
1	(J. J. Khanam and S. Y. Foo, 2021)	Naïve Bayes Decision Tree C45	75,53% 74,24%
2	(Kumari et al., 2021)	Naïve Bayes Decision Tree C45	74,12% 71,42%
3	(N. P. Tigga and S. Garg, 2020)	Naïve Bayes Decision Tree C45	80,6% 84%

Table 11. Previous Research Results (Continue)

No	Author	Algorithm	Accuracy
4	(Hasan et al., 2020)	Naïve Bayes Decision Tree C45	87,9% 91,2%
5	(Choubey et al., 2020)	Naïve Bayes Decision Tree C45	77,32% 79,92%
6	(N. Nnamoko and I. Korkontzelos, 2020)	Naïve Bayes Decision Tree C45	76% 74,7%
7	(Nagaraj et al., 2021)	Naïve Bayes	74,9%
8	(Gadekallu et al., 2020)	Naïve Bayes Decision Tree C45	57% 88,2%

The accuracy performance of the Nave Bayes Algorithm in the previous experiment was less than 90%. However, Hasan et al. (2020) found that the C4.5 Decision Tree Algorithm performed better than our accuracy results in their investigation. The Pima India Diabetes Dataset, which has 8 variables including gestational age, blood sugar level, blood pressure, triceps, insulin, body mass index, family history of diabetes, and age, was used in the study by Hasan et al. There are variations in the number of attributes used. In addition, the attributes used also have differences, as in our study we did not use the gestational age attribute because this study was used for the detection of diabetes in general. Then the attributes of triceps, insulin, and hereditary history of diabetes are not present in the dataset that we use. The higher accuracy results are due to the different attributes used. The determination of attributes greatly influences accuracy performance (Hasan et al., 2020). Apart from attributes, the number of attributes and sample dimensions also greatly influence the performance of the model built (Choubey et al., 2020).

*name of corresponding author

CONCLUSION

Alternative methods for early diagnosis of diabetes are the Decision Tree C4.5 and Naive Bayes methods. By using the Decision Tree Algorithm C4.5, the highest average accuracy is obtained with a value of 90.835%. Pruning can help increase the accuracy of the Decision Tree C4.5 algorithm from its greatest accuracy performance of 90.97% to 91.30%. 18 patterns or guidelines for early identification of diabetics were derived from this model. The performance of the developed model is significantly influenced by the training attributes, the number of attribute dimensions, and the number of samples.

ACKNOWLEDMENT

The author thanks Dr. Christina Juliane, who never tired of helping and providing guidance to complete this research.

REFERENCES

Anggraini, S., Defit, S., & Nurcahyo, G. W. (2018). Analisis Data Mining Penjualan Ban Menggunakan Algoritma C4. 5. *Jurnal Ilmu Teknik Elektro …*, *5*(2), 0–7. https://core.ac.uk/download/pdf/295348196.pdf

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, *192*(xxxx), 105361. https://doi.org/10.1016/j.knosys.2019.105361

Chen, W., Li, Y., Xue, W., Shahabi, H., Li, S., Hong, H., Wang, X., Bian, H., Zhang, S., Pradhan, B., & Ahmad, B. Bin. (2020). Modeling flood susceptibility using data-driven approaches of naïve Bayes tree, alternating decision tree, and random forest methods. *Science of the Total Environment*, *701*, 134979. https://doi.org/10.1016/j.scitotenv.2019.134979

Choubey, D. K., Kumar, P., Tripathi, S., & Kumar, S. (2020). Performance evaluation of classification methods with PCA and PSO for diabetes. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *9*(1). https://doi.org/10.1007/s13721-019-0210-8

Enriko, I. K. A., Melinda, M., Sulyani, A. C., & Astawa, I. G. B. (2021). Breast cancer recurrence prediction system using k-nearest neighbor, naïve-bayes, and support vector machine algorithm. *Jurnal Infotel*, *13*(4), 185–188. https://doi.org/10.20895/infotel.v13i4.692

Fajriati, N., Prasetiyo, B., Semarang, U. N., & Korespondensi, P. (2023). *Optimasi algoritma naïve bayes dengan diskritisasi k-means optimization of naïve bayes algorithm using k-means discretization in heart disease diagnosis*. *10*(3), 503–512. https://doi.org/10.25126/jtiik.2023106510

Fersellia, F., Utami, E., & Yaqin, A. (2023). Sentiment Analysis of Shopee Food Application User Satisfaction Using the C4.5 Decision Tree Method. *Sinkron*, *8*(3), 1554–1563. https://doi.org/10.33395/sinkron.v8i3.12531

Gadekallu, T. R., Khare, N., Bhattacharya, S., Singh, S., Maddikunta, P. K. R., Ra, I. H., & Alazab, M. (2020). Early detection of diabetic retinopathy using pca-firefly based deep learning model. *Electronics (Switzerland)*, *9*(2), 1–16. https://doi.org/10.3390/electronics9020274

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516–76531. https://doi.org/10.1109/ACCESS.2020.2989857

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-68771-z

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*(November 2020), 40–46. https://doi.org/10.1016/j.ijcce.2021.01.001

Nagaraj, P., Deepalakshmi, P., Mansour, R. F., & Almazroa, A. (2021). Artificial flora algorithm-based feature selection with gradient boosted tree model for diabetes classification. *Diabetes, Metabolic Syndrome and Obesity*, *14*, 2789–2806. https://doi.org/10.2147/DMSO.S312787

Patra, K. C., Sethi, R. N., & Behera, D. K. (2021). Benchmark of Unsupervised Machine Learning Algorithms for Condition Monitoring. In *Lecture Notes in Networks and Systems: Vol. 185 LNNS*.

*name of corresponding author

https://doi.org/10.1007/978-981-33-6081-5_17

Rianto, H., Amrin, Rudianto, Pahlevi, O., Kusumawardhani, P., & Hadi, S. S. (2020). Determining the Eligibility of Providing Motorized Vehicle Loans by Using the Logistic Regression, Naive Bayes and Decission Tree (C4.5). *Journal of Physics: Conference Series*, *1641*(1). https://doi.org/10.1088/1742-6596/1641/1/012061

Shrinivasan, L., Verma, R., & Nandeesh, M. D. (2023). Early prediction of diabetes diagnosis using hybrid classification techniques. *IAES International Journal of Artificial Intelligence*, *12*(3), 1139–1148. https://doi.org/10.11591/ijai.v12.i3.pp1139-1148

Singh, S., & Yassine, A. (2018). Big data mining of energy time series for behavioral analytics and energy consumption forecasting. *Energies*, *11*(2). https://doi.org/10.3390/en11020452

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, *513*, 429–441. https://doi.org/10.1016/j.ins.2019.11.004

Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Computer Science*, *167*(2019), 706–716. https://doi.org/10.1016/j.procs.2020.03.336

Wong, T., & Yeh, P. (n.d.). *10.1109@Tkde.2019.2912815*. *1*, 1.

*name of corresponding author