Jurnal Teknik Informatika C.I.T Medicom, 16 (1) (2024) 01-12

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Comparison of data mining algorithms (random forest, C4.5, catboost) based on adaptive boosting in predicting diabetes mellitus

Yennimar[1], William Leonardi[2], Harris Weide[3], Devin Cantona[4], Gani Mores Hutagalung[5]
[1,2,3,4,5]Program Sarjana Teknik Informatika, Universitas Prima Indonesia, Medan, Indonesia

Article Info

Article history

Received : Feb 05, 2024
Revised : Feb 14, 2024
Accepted : Mar 28, 2024

Keywords:

Comparative Analysis;
Data Mining;
Diabetes Mellitus.

Abstract

This research aims to evaluate the performance of three algorithms data mining, namely C4.5, Random Forest, and Catboost Classifier, which are strengthened by Adaptive Boosting in predicting diabetes mellitus in humans. Through analysis, it was found that the C4.5 algorithm is based on Adaptive Boosting obtained an average accuracy of 73.74%, precision of 61.39%, and recall amounting to 69.00%. Random Forest algorithm based on Adaptive Boosting shows an average accuracy of 73.52%, precision of 65.79%, and recall amounting to 65.06%. Meanwhile, the Catboost Classifier algorithm is Adaptive based Boosting has an average accuracy of 73.67%, precision of 61.19%, and recall was 69.18%. Thus, although all three algorithms shows similar performance, the C4.5 algorithm based on Adaptive Boosting stands out with better performance in terms of accuracy, precision and recall. The implication of this research is that the use of the C4.5 algorithm is based Adaptive Boosting can be a more effective approach to support early detection of diabetes mellitus in humans.

Corresponding Author:

Yennimar,
Faculty of Science and Technology, Program Studi Informatics Engineering,
Prima Indonesia University,
Jl. Sampul No.3, Sei Putih Bar., Kec. Medan Petisah, Medan, Sumatera Utara 20118, Indonesia
Email: Yennimar@unprimdn.ac.id

1. Introduction

Diabetes mellitus is a chronic metabolic disease or disorder with multi etiology which is characterized by high blood sugar levels accompanied by with disorders of carbohydrate[1][2], lipid and protein metabolism as a consequence insufficiency of insulin function[3][4][5]. In Indonesia, diabetes mellitus is the cause The sixth most common death, after conditions related to childbirth[6][1]. In 2021, around 19.5 million individuals will be diagnosed with diabetes mellitus, making Indonesia ranked fifth globally in terms of highest number of diabetes patients. Untreated and untreated diabetes identified can cause serious complications in sufferers[7][8]. By Therefore, it is very important to carry out early detection of diabetes mellitus because if this disease is left for too long without treatment[9], it can resulting in dangerous complications such as kidney failure, damage to organ function others up to a heart attack[10].

From a computer science perspective, the process of early detection of diabetes mellitus in humans can do this through information technology by applying techniques data mining[11][12][13][14]. Data

mining involves a series of actions or processes to discover meaningful relationships through patterns and trends in large data sets using various methods and algorithms[15]. Advantages of using information systems and data mining techniques is its ability to perform early detection of diabetes mellitus in humans quickly[14][16][16][13]. Additionally, with data mining, the detection process can involve several parameters so not only relying on just one parameter[17][18]. Parameters such as age, weight, pressure blood, and other relevant factors can be included, resulting in more accurate detection results[19][20][21].

There are several previous studies that implemented the technique data mining in predicting diabetes mellitus[22][12][23][24]. Research conducted in 2023 implementing the Random Forest algorithm in detection early onset of diabetes mellitus in humans[25][26][9]. Algorithm performance evaluation results Random Forest for predicting diabetes gets an accuracy of 77.06%, precision of 71.43%, recall of 47.30%, and classification error amounted to 22.94% [27]. The next research will be carried out in 2023 implementing the Decision Tree C4.5 algorithm in carrying out classification diabetes mellitus[28][29]. Performance evaluation results of the Decision Tree C4.5 algorithm for the classification of diabetes, the accuracy was 79%, precision of 78%, recall of 45%, and F1-score of 57% [30]. Contribute to This research is to conduct a comparative study using a data algorithm mining that has been implemented in previous research is Random Forest and C4.5 in predicting diabetes mellitus in humans. As a distinction from previous comparative studies, 1 was added Another algorithm, namely the Catboost Classifier, has not yet been discovered by research a type that compares these algorithms. The three algorithms This is also combined with Adaptive Boosting so that it can improve the accuracy performance of the three algorithms.

2. Methods

This type of research is using quantitative types of research namely systematic scientific research on parts and phenomena as well the quality of the relationships[31][32]. The aim of quantitative research is develop and use mathematical models, theories or hypotheses relating to natural phenomena[33]. In this case, research The research carried out focuses on comparing the C4.5, Random Forest, and algorithm Catboost Classifier based on Adaptive Boosting to predict diseases diabetes mellitus in humans[34][35][36].

2.1 Work Procedures

In this research, the work procedures for the research to be carried out can be seen in Figure 1.

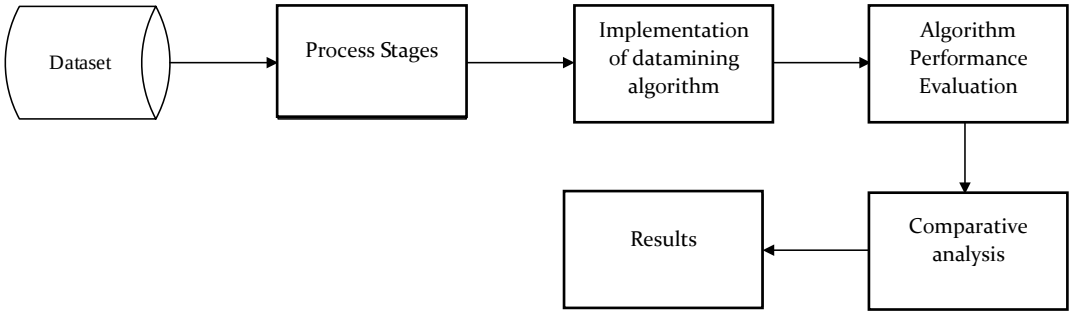

Figure 1. Work Procedure [36]

2.2 Dataset

The dataset used in this research was taken from the Kaggle website, namely the Pima Indians Diabetes Database with a total of 768 records. Table 1 shows the dataset used in this research.

Table 1. Research Datase

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DPF	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DPF	Age	Outcome
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
...	...	...	...	...	...	...	...	...	...
766	5	121	72	23	112	26.2	0.245	30	0
767	1	126	60	0	0	30.1	0.349	47	1
768	1	93	70	31	0	30.4	0.315	23	0

2.3 Pre-processing stages

The data pre-processing stage in this research was carried out using the average imputation method, which is a method for dealing with missing data in the dataset. In this process, missing values in a feature (column) are replaced with the average value of that feature. This method helps maintain consistency and integrity of data in the dataset. The pre-processing process was carried out directly using the Python programming language using Google Colab.

2.4 Implementation of Data Mining Algorithms

At this stage, the three data mining algorithms that will be analyzed in this research are implemented, namely the C4.5 algorithm, Random Forest, Catboost Classifier based on Adaptive Boosting in predicting diabetes mellitus in humans. The algorithm implementation was written using the Python programming language using Google Colab.

2.5 Algorithm Performance Evaluation

Evaluation of algorithm performance in this research was carried out using a Confusion Matrix, namely a cross-tabulation of positive and negative class data grouped into predicted and actual classes.

2.6 Comparative Analysis

At this stage, a comparative analysis of the three data mining algorithms implemented in this research was carried out before and after implementing Adaptive Boosting. The results of the comparative analysis show[36].

2.7 Result

The results of the research are in the form of a comprehensive discussion of the analysis that has been carried out, described in detail and linked to previous research.

3. Result and Discussion

The test results obtained in this research were processed using Google Colab where the aim of this research was to determine the performance of the C4.5 algorithm, Random Forest, and Catboost Classifier based on Adaptive Boosting in predicting diabetes mellitus in humans.

3.1 Preparing Datasets

The total data available is 768 data. This figure, Figure 2, shows the data set used in this research.

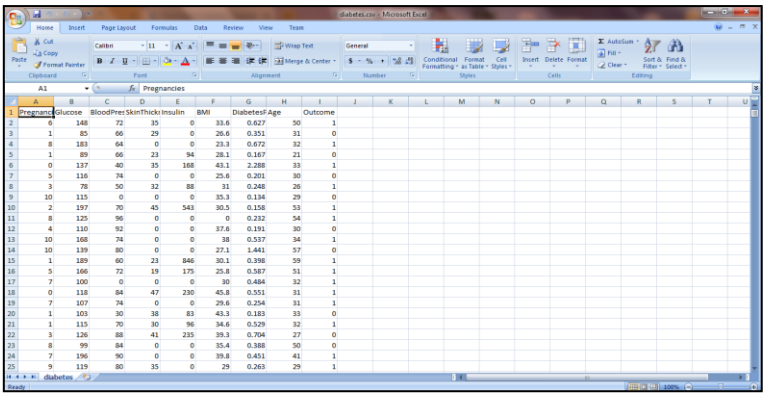

Figure 2. Research Dataset

There are 9 variable attributes in the data set used in this research, including: number of pregnancies (pregnancies), blood sugar levels (glucose), blood pressure (blood pressure), skin thickness, insulin, Body Mass Indeix (BMI), indicators of family history of diabetes (diabetes pedigree function), age and diagnosis results (outcome). In the dataset used in this research, there was still some data that was found to have no value or was often called missing value, so preprocessing of the dataset needed to be carried out.

3.2 Data Preprocessing Results

The data preprocessing stages in this research were carried out using Google Collab based on Python programming. Before implementing the data mining algorithm, a preprocessing stage was carried out using the average imputation method where the value was 0 for the attributes of blood glucose levels, blood pressure, skin thickness, insulin, and body mass. Index (BMI). Below, Figure 3 shows examples of results before and after data preprocessing is carried out.

```
Dataset Sebelum Preprocessing:
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1

Dataset setelah imputasi:
   Pregnancies  Glucose  BloodPressure  SkinThickness     Insulin   BMI  \
0            6    148.0           72.0      35.000000   79.799479  33.6
1            1     85.0           66.0      29.000000   79.799479  26.6
2            8    183.0           64.0      20.536458   79.799479  23.3
3            1     89.0           66.0      23.000000   94.000000  28.1
4            0    137.0           40.0      35.000000  168.000000  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1
```

Figure 3. Example of Dataset Preprocessing Results

In Figure 3, it can be seen especially in the data in the red box where in the insulin and skin thickness attributes there is data with a value of 0 so that the average imputation process is carried out to have an average value based on each attribute.

3.3 Data Mining Algorithm Implementation Results

In this research, the implementation of the data mining algorithm aims to determine the performance of the C4.5 algorithm, Random Forest, and Catboost Classifier based on Adaptive Boosting in predicting diabetes mellitus in humans. The process of implementing data mining algorithms is used with Google Colab tools using the Python programming language as shown in Figure 4.

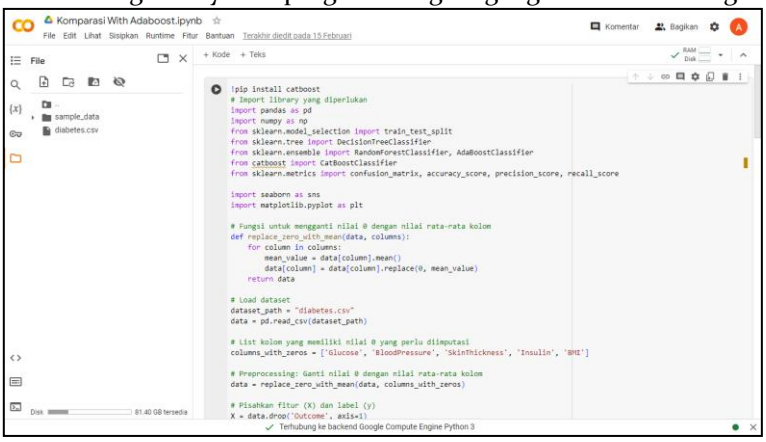

Figure 4. Data Mining Algorithm Implementation Results

3.4 Algorithm Performance Evaluation Results

After the data mining algorithm is implemented and produces a model, the resulting model is then tested using a Confusion Matrix with a comparison between testing data and training data, namely the first test is 10:90, the second test is 20:80, and the third test is 30:70. The following are the results of the performance evaluation of the data mining algorithms tested in this research, including:

a) Performance evaluation results of the C4.5 algorithm based on Adaptive Boosting.
The first test was with 10% testing data and 90% training data.

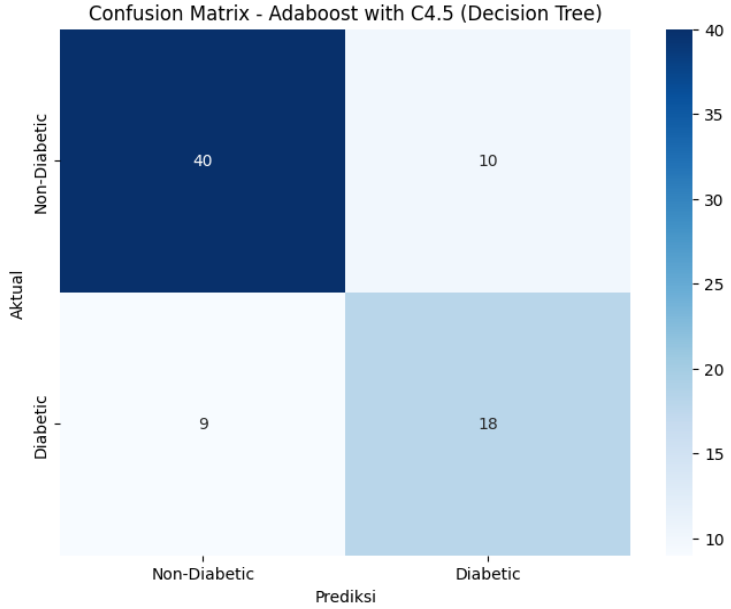

Figure 5. Confusion Matrix Plot of C4.5 Algorithm Based on Adaptive Boosting with 10% Testing Data and 90% Training Data

Based on Figure 3.4, it can be seen that the performance of the C4.5 algorithm based on Adaptive Boosting in the first test produced an accuracy of 75.32%, precision of 64.29%, recall of 66.67%, and misclassification error of 24.68%.

The second test was with 20% testing data and 80% training data.

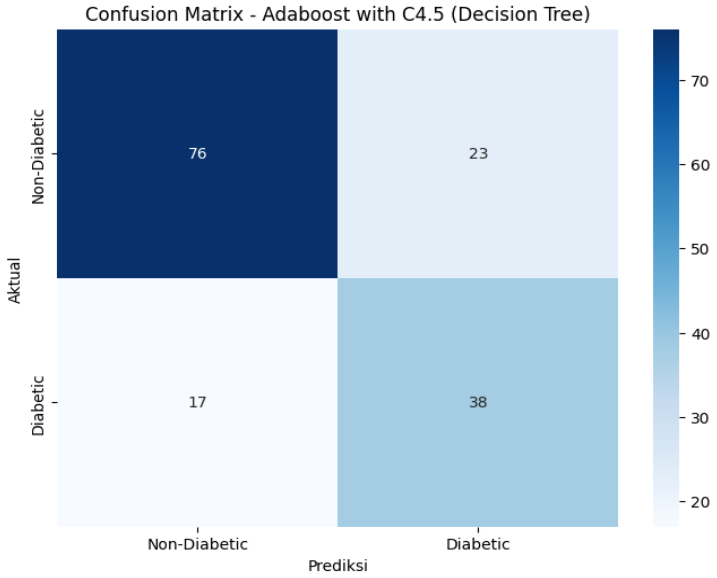

Figure 6. Confusion Matrix Plot of C4.5 Algorithm Based on Adaptive Boosting with 20% Testing Data and 80% Training Data

Based on Figure 3.5, it can be seen that the performance of the C4.5 algorithm based on Adaptive Boosting in the second test produced an accuracy of 74.03%, precision of 62.30%, recall of 69.09%, and misclassification error of 25.97%.

The third test is with 30% testing data and 70% training data.

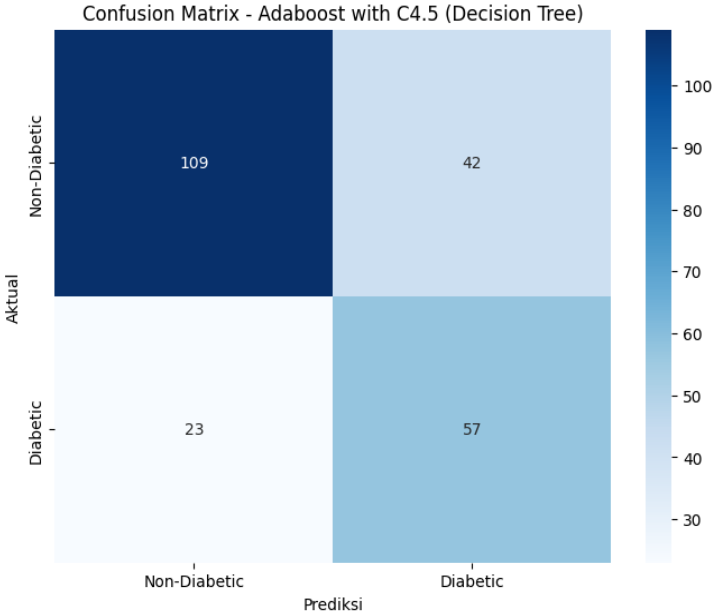

Figure 7. Confusion Matrix Plot of C4.5 Algorithm Based on Adaptive Boosting with 30% Testing Data and 70% Training Data

Based on Figure 7, it can be seen that the performance of the C4.5 algorithm based on Adaptive Boosting in the third test produced an accuracy of 71.86%, precision of 57.58%, recall of 71.25%, and misclassification error of 28.14%.

b) Performance evaluation results of the Random Forest algorithm based on Adaptive Boosting. The first test was with 10% testing data and 90% training data.

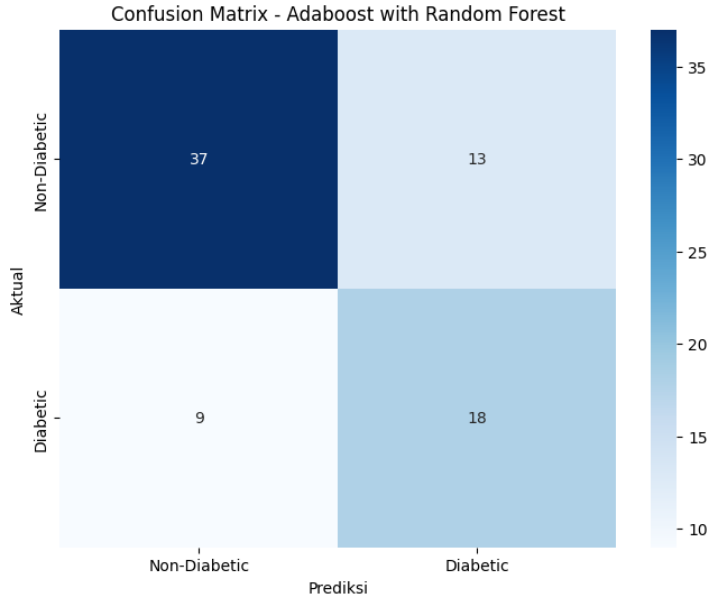

Figure 8. Confusion Matrix Plot of Random Forest Algorithm Based on Adaptive Boosting with 10% Testing Data and 90% Training Data

Based on Figure 8., it can be seen that the performance of the Random Forest algorithm based on Adaptive Boosting in the first test produced an accuracy of 71.43%, precision of 58.06%, recall of 66.67%, and misclassification error of 28.57%.

The second test was with 20% testing data and 80% training data.

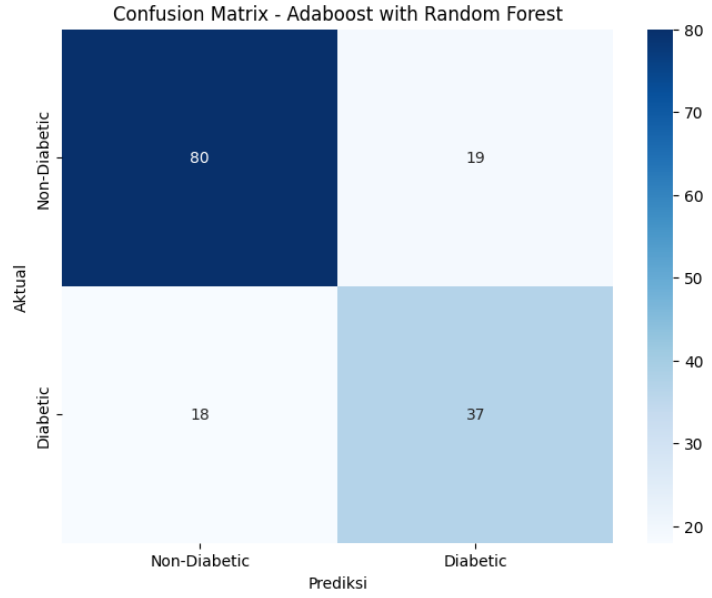

Figure 9. Confusion Matrix Plot of Random Forest Algorithm Based on Adaptive Boosting with 20% Testing Data and 80% Training Data

Based on Figure 9, it can be seen that the performance of the Random Forest algorithm based on Adaptive Boosting in the second test produced an accuracy of 75.97%, precision of 66.07%, recall of 67.27%, and misclassification error of 24.03%.

The third test is with 30% testing data and 70% training data.

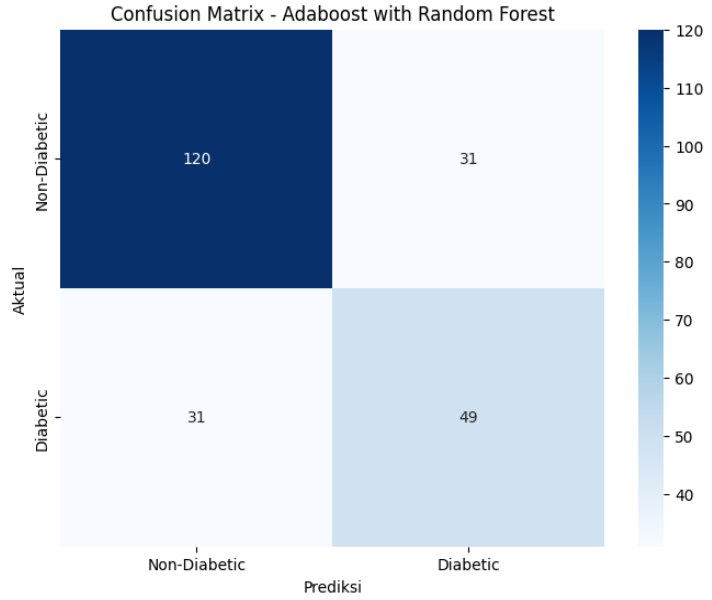

Figure 10. Confusion Matrix Plot of Random Forest Algorithm Based on Adaptive Boosting with 30% Testing Data and 70% Training Data

Based on Figure 10, it can be seen that the performance of the Random Forest algorithm based on Adaptive Boosting in the third test produced an accuracy of 73.16%, precision of 61.25%, recall of 61.25%, and misclassification error of 26.84%.

c) Performance evaluation results of the Catboost Classifier algorithm based on Adaptive Boosting.

The first test was with 10% testing data and 90% training data.

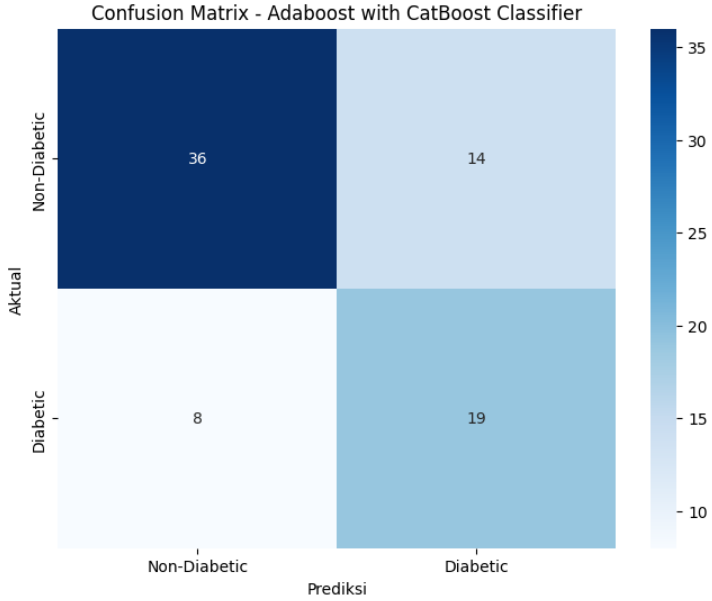

Figure 11 Confusion Matrix Plot of Catboost Classifier Algorithm Based on Adaptive Boosting with 10% Testing Data and 90% Training Data

Based on Figure 11, it can be seen that the performance of the Catboost Classifier algorithm based on Adaptive Boosting in the first test produced an accuracy of 71.43%, precision of 57.58%, recall of 70.37%, and misclassification error of 28.57%.

The second test was with 20% testing data and 80% training data

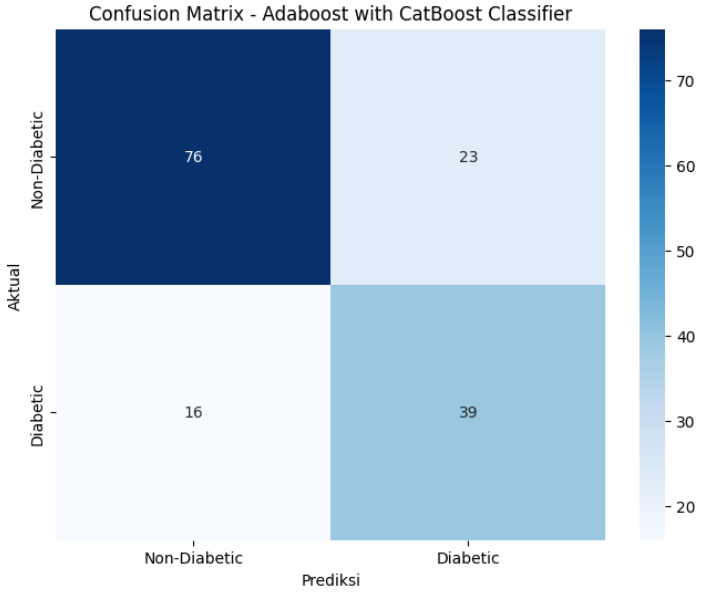

Figure 12 Confusion Matrix Plot of Catboost Classifier Algorithm Based on Adaptive Boosting with 20% Testing Data and 80% Training Data

Based on Figure 12, it can be seen that the performance of the Catboost Classifier algorithm based on Adaptive Boosting in the second test produced an accuracy of 74.68%, precision of 62.90%, recall of 70.91%, and misclassification error of 25.32%.
The third test is with 30% testing data and 70% training data.

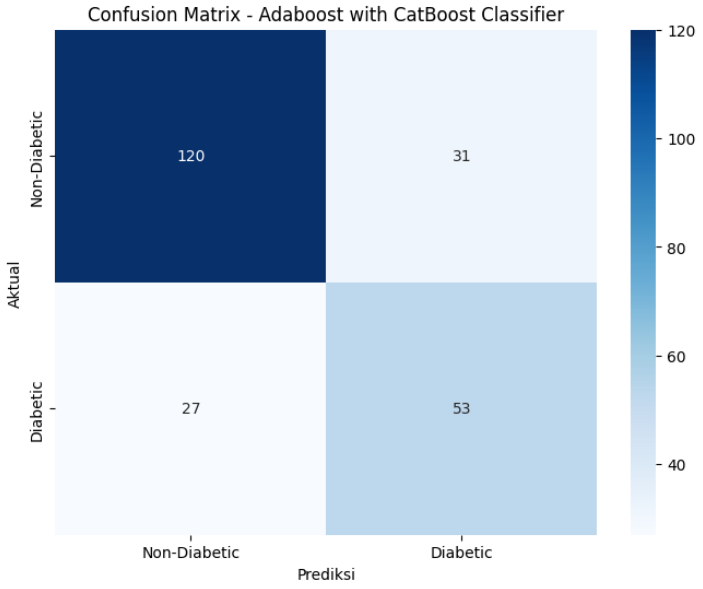

Figure 13. Confusion Matrix Plot of Catboost Classifier Algorithm Based on Adaptive Boosting with 30% Testing Data and 70% Training Data

Based on Figure 13, it can be seen that the performance of the Catboost Classifier algorithm based on Adaptive Boosting in the third test resulted in an accuracy of 74.89%, precision of 63.10%, recall of 66.25%, and misclassification error of 25.11%.

3.5 Comparative Analysis Results.

After the algorithm performance evaluation process has been carried out, a comparison of the three algorithms tested in this study is then carried out to find out which algorithm is superior in predicting diabetes mellitus in humans. The results of the comparative analysis are presented in Table 2 below.

Table 2. Algorithm Comparative Analysis Results

Algoritma	Data *Testing*:Data *Training*	Akurasi		Presisi		*Recall*		*Missclassification Error*	
		*	**	*	**	*	**	*	**
C4.5	10:90	74,03	75,32	62,07	64,29	66,67	66,67	25,97	24,68
	20:80	72,08	74,03	59,38	62,30	69,09	69,09	27,92	25,97
	30:70	68,83	71,86	54,17	57,58	65,00	71,25	31,17	28,14
	Rata-Rata	**71,65**	**73,74**	**58,54**	**61,39**	**66,92**	**69,00**	**28,35**	**26,26**
Random Forest	10:90	70,13	71,43	56,67	58,06	62,96	66,67	29,87	28,57
	20:80	76,62	75,97	66,10	66,07	70,91	67,27	23,38	24,03
	30:70	74,03	73,16	62,20	61,25	63,75	61,25	25,97	26,84
	Rata-Rata	**73,59**	**73,52**	**61.66**	**61.79**	**65,87**	**65,06**	**26,41**	**26,48**
Catboost Classifier	10:90	74,03	71,43	62,07	57,58	66,67	70,37	25,97	28,57
	20:80	74,03	74,68	64,15	62,90	61,82	70,91	25,97	25,32
	30:70	74,03	74,89	62,82	63,10	61,25	66,25	25,97	25,11
	Rata-Rata	**74,03**	**73,67**	**63,01**	**61,19**	**63,25**	**69,18**	**25,97**	**26,33**

Description:
* = Non Adaptive Boosting
** = Adaptive Boosting

4. Conclusion

Based on the research that has been carried out, conclusions are drawn The performance of the C4.5 algorithm based on Adaptive Boosting obtained average accuracy of 73.74%, precision of 61.39%, recall of 69.00%, and misclassification error of 26.26%. Performance of the Random Forest algorithm Adaptive Boosting based obtained an average accuracy of 73.52%, precision of 65.79%, recall of 65.06%, and misclassification error of 26.48%. Performance of Catboost Classifier algorithm based on Adaptive Boosting obtained an average accuracy of 73.67%, precision of 61.19%, recall amounted to 69.18%, and misclassification error amounted to 26.33%. The research results show that the performance of the C4.5 based algorithm Adaptive Boosting is superior to Random Forest algorithms and Catboost Classifier is based on Adaptive Boosting. While this study provides valuable insights into the predictive performance of C4.5, Random Forest, and Catboost Classifier algorithms with Adaptive Boosting in predicting diabetes mellitus, several limitations should be acknowledged. Firstly, the findings are contingent upon the specific dataset used for analysis, potentially limiting the generalizability of the results. Moreover, the study focuses solely on these three algorithms, neglecting the exploration of other potentially effective models or ensemble methods. Additionally, the evaluation metrics employed, though common, may not fully capture the nuances of algorithm performance, especially in imbalanced datasets. Furthermore, the study may not have exhaustively explored all parameter settings for each algorithm, leaving room for further investigation into optimal parameter tuning strategies. Despite these limitations, the research presents promising avenues for future exploration. Firstly, researchers could delve into advanced feature engineering techniques to enhance the predictive power of the models, specifically tailored to diabetes mellitus prediction. Secondly, exploring novel ensemble strategies or hybrid approaches could lead to improved predictive accuracy and robustness. Additionally, incorporating temporal analysis techniques could capture dynamic patterns in health data, enabling early detection and intervention for individuals at risk of developing diabetes mellitus. Moreover, enhancing the interpretability and explainability of the models is crucial for their acceptance in clinical settings, warranting further research in this area. Finally, validating the findings on external datasets from diverse populations and healthcare settings could ascertain the generalizability and

applicability of the C4.5 algorithm with Adaptive Boosting in supporting early detection of diabetes mellitus.

References

[1] S. Alam, M. K. Hasan, S. Neaz, N. Hussain, M. F. Hossain, and T. Rahman, "Diabetes Mellitus: insights from epidemiology, biochemistry, risk factors, diagnosis, complications and comprehensive management," *Diabetology*, vol. 2, no. 2, pp. 36–50, 2021.

[2] O. O. Oguntibeju, "Type 2 diabetes mellitus, oxidative stress and inflammation: examining the links," *Int. J. Physiol. Pathophysiol. Pharmacol.*, vol. 11, no. 3, p. 45, 2019.

[3] R. C. R. Meex, E. E. Blaak, and L. J. C. van Loon, "Lipotoxicity plays a key role in the development of both insulin resistance and muscle atrophy in patients with type 2 diabetes," *Obes. Rev.*, vol. 20, no. 9, pp. 1205–1217, 2019.

[4] S. Hong and K. M. Choi, "Sarcopenic obesity, insulin resistance, and their implications in cardiovascular and metabolic consequences," *Int. J. Mol. Sci.*, vol. 21, no. 2, p. 494, 2020.

[5] P. Morigny, J. Boucher, P. Arner, and D. Langin, "Lipid and glucose metabolism in white adipocytes: pathways, dysfunction and therapeutics," *Nat. Rev. Endocrinol.*, vol. 17, no. 5, pp. 276–295, 2021.

[6] A. A. Choudhury and V. D. Rajeswari, "Gestational diabetes mellitus-A metabolic and reproductive disorder," *Biomed. Pharmacother.*, vol. 143, p. 112183, 2021.

[7] J. S. Varghese *et al.*, "Diabetes diagnosis, treatment, and control in India: results from a national survey of 1.65 million adults aged 18 years and older, 2019-2021," *medRxiv*, pp. 2002–2023, 2023.

[8] A. Sanyaolu *et al.*, "Diabetes mellitus: An overview of the types, prevalence, comorbidity, complication, genetics, economic implication, and treatment," *World J. Meta-Analysis*, vol. 11, no. 5, pp. 134–143, 2023.

[9] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *J. Big data*, vol. 6, no. 1, pp. 1–19, 2019.

[10] A. Chauin, "The main causes and mechanisms of increase in cardiac troponin concentrations other than acute myocardial infarction (Part 1): physical exertion, inflammatory heart disease, pulmonary embolism, renal failure, sepsis," *Vasc. Health Risk Manag.*, pp. 601–617, 2021.

[11] L. Chaves and G. Marques, "Data mining techniques for early diagnosis of diabetes: a comparative study," *Appl. Sci.*, vol. 11, no. 5, p. 2218, 2021.

[12] F. A. Khan, K. Zeb, M. Al-Rakhami, A. Derhab, and S. A. C. Bukhari, "Detection and prediction of diabetes using data mining: a comprehensive review," *IEEE Access*, vol. 9, pp. 43711–43735, 2021.

[13] Y. Liu, Z. Yu, and Y. Yang, "Diabetes risk data mining method based on electronic medical record analysis," *J. Healthc. Eng.*, vol. 2021, 2021.

[14] H. Thakkar, V. Shah, H. Yagnik, and M. Shah, "Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis," *Clin. eHealth*, vol. 4, pp. 12–23, 2021.

[15] H. Yan, N. Yang, Y. Peng, and Y. Ren, "Data mining in the construction industry: Present status, opportunities, and future trends," *Autom. Constr.*, vol. 119, p. 103331, 2020.

[16] W.-T. Wu *et al.*, "Data mining in clinical big data: the frequently used databases, steps, and methodological models," *Mil. Med. Res.*, vol. 8, no. 44, pp. 1–12, 2021, doi: https://doi.org/10.1186/s40779-021-00338-z.

[17] S. Deng, N. Zhang, J. Kang, Y. Zhang, W. Zhang, and H. Chen, "Meta-learning with dynamic-memory-based prototypical network for few-shot event detection," in *Proceedings of the 13th international conference on web search and data mining*, 2020, pp. 151–159.

[18] F. E. Bock, R. C. Aydin, C. J. Cyron, N. Huber, S. R. Kalidindi, and B. Klusemann, "A review of the application of machine learning and data mining approaches in continuum materials mechanics," *Front. Mater.*, vol. 6, p. 110, 2019.

[19] E. Martinez-Ríos, L. Montesinos, M. Alfaro-Ponce, and L. Pecchia, "A review of machine learning in hypertension detection and blood pressure estimation based on clinical and physiological data," *Biomed. Signal Process. Control*, vol. 68, p. 102813, 2021.

[20] A. Shrivastava, M. Chakkaravarthy, and M. A. Shah, "A new machine learning method for predicting systolic and diastolic blood pressure using clinical characteristics," *Healthc. Anal.*, vol. 4, p. 100219, 2023.

[21] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–15, 2019.

[22] F. Kazerouni, A. Bayani, F. Asadi, L. Saeidi, N. Parvizi, and Z. Mansoori, "Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: a comparison of four data mining approaches," *BMC Bioinformatics*, vol. 21, pp. 1–13, 2020.

[23] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and prediction of diabetes complication disease using data mining algorithm," *Procedia Comput. Sci.*, vol. 161, pp. 449–457, 2019.
[24] M. M. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Computer vision and machine intelligence in medical image analysis*, Springer, 2020, pp. 113–125.
[25] X. Wang *et al.*, "Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier," *BMC Med. Inform. Decis. Mak.*, vol. 21, pp. 1–14, 2021.
[26] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Sci. Rep.*, vol. 10, no. 1, p. 11981, 2020.
[27] A. Andi, T. Thamrin, A. Susanto, E. Wijaya, and D. Djohan, "Analysis of the random forest and grid search algorithms in early detection of diabetes mellitus disease," *J. Mantik*, vol. 7, no. 2, pp. 1117–1124, 2023.
[28] B. A. C. Permana, R. Ahmad, H. Bahtiar, A. Sudianto, and I. Gunawan, "Classification of diabetes disease using decision tree algorithm (C4. 5)," in *Journal of Physics: Conference Series*, IOP Publishing, 2021, p. 12082.
[29] P. Purbandini, E. Purwanti, E. Hariyanti, and F. Y. Ramadhan, "Application of the decision tree C4. 5 method on the classification of diet types of people with diabetes mellitus," in *AIP Conference Proceedings*, AIP Publishing, 2023.
[30] A. U. Haq *et al.*, "Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data," *Sensors*, vol. 20, no. 9, p. 2649, 2020.
[31] B. Kreshpaj *et al.*, "What is precarious employment? A systematic review of definitions and operationalizations from quantitative and qualitative studies," *Scand. J. Work. Environ. Health*, vol. 46, no. 3, pp. 235–247, 2020.
[32] T. Hascher and J. Waber, "Teacher well-being: A systematic review of the research literature from the year 2000–2019," *Educ. Res. Rev.*, vol. 34, p. 100411, 2021.
[33] H. K. Mohajan, "Quantitative research: A successful investigation in natural and social sciences," *J. Econ. Dev. Environ. People*, vol. 9, no. 4, pp. 50–79, 2020.
[34] S. Acharya, "Comparative analysis of classification accuracy for XGBoost, LightGBM, CatBoost, H2O, and Classifium." 2022.
[35] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cogn. Comput. Eng.*, vol. 2, pp. 40–46, 2021.
[36] M. R. Islam, S. Banik, K. N. Rahman, and M. M. Rahman, "A comparative approach to alleviating the prevalence of diabetes mellitus using machine learning," *Comput. Methods Programs Biomed. Updat.*, vol. 4, p. 100113, 2023.