Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 6, No.2, April 2024, pp. 0240207-01 ~ 0240207-08
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v6i2.18342

Implementation of the Adaboost Method to Increase the Accuracy of Early Diabetes Predictions to Prevent Death Decision Tree-Based

Laskar Alam*, Ahmad Zainul Fanani, Affandy

Faculty of Computer Science, University of Dian Nuswantoro, Imam Bonjol No. 207 Semarang, Central Java, 50131, Indonesia

*p31202102446@mhs.dinus.ac.id

Abstract. This research discusses the importance of early diabetes prediction and efforts to increase prediction accuracy using a Decision Tree Learning Algorithm and integration of the Adaboost Method. This study uses a data set from Kaggle with 520 records, 16 attributes, and one positive or negative diabetes class. The evaluation method used is the Confusion Matrix. The research results showed that the Decision Tree algorithm achieved an accuracy of 94.23%, but after integrating the Adaboost Method, the accuracy increased to 97.31%. The implications of these findings emphasize the importance of predictive approaches in early disease detection and highlight the potential of the Adaboost method in improving the accuracy of diabetes prediction.

Keywords: Decision Tree, Diabetes, Machine Learning, Prediction, Adaboost. Confusion Matrix

(Received 2024-02-03, Accepted 2024-03-01, Available Online by 2024-03-08)

1. Introduction

Diabetes is a disease that threatens global public health but is challenging to detect early because of the lack of apparent symptoms. According to the World Health Organization (WHO), diabetes is ranked 9th as the deadliest disease in the world. In Indonesia itself, diabetes is a severe problem, ranking 7th as the country with the highest number of people with diabetes [1]. The impact of diabetes is comprehensive, starting from damaging vital organs such as the kidneys, eyes, and nerves to increasing the risk of heart disease and even death in mothers during childbirth. Delayed diagnosis often causes complications that lead to death before the patient realizes that they have diabetes [2].

The importance of early detection in preventing the destructive effects of diabetes drives the need for an effective prediction system [3]. By analyzing the supporting attributes of diabetes, prediction systems can be a more affordable and efficient option compared to visits to specialist doctors and laboratory tests. For this reason, machine learning methods, such as the Decision Tree algorithm, are often used in efforts to predict diabetes. In the field of prediction, several algorithms are often used in research related to machine learning, namely Decision Tree [4], Neural Network [5], Support Vector Machine, and Naive Bayes [6]. One of the machine learning algorithms for making predictions with the

highest level of accuracy, the most popular, which is easy for humans to understand and is often used, is the Decision Tree algorithm [7], [8].

Although the Decision Tree algorithm is effective, it has limitations in prediction accuracy, especially in the case of disease prediction [8]. Therefore, this research aims to increase prediction accuracy by integrating the Adaboost method into the Decision Tree algorithm. The Adaboost method is used to improve the performance of a single algorithm by forming several prediction models from training data [9].

This research aims to overcome the limited accuracy of diabetes prediction by integrating the Adaboost method into the Decision Tree algorithm. Through data analysis from the Kaggle dataset consisting of 520 records with 16 attributes, including age, gender, and other symptoms, this research evaluates the prediction performance using the Confusion Matrix. So, this research will provide an overview of the effectiveness of using the Decision Tree algorithm enhanced with the Adaboost method in predicting diabetes. The implications of the findings from this study will help improve understanding of the importance of predictive approaches in managing chronic diseases such as diabetes.

2. Methods

This research methodology outlines the steps that will be taken in the assessment process to achieve the stated research objectives. The Cross Industry Standard Process for Data Mining (CRISP-DM) data mining standardization model was chosen for this research methodology. CRISP-DM was selected because it is one of the most frequently used data mining methods. This study uses the Decision Tree classification model with the C4.5 algorithm, and the boosting is carried out using the Adaboost algorithm. Below is a picture of the flow of the stages of the research.

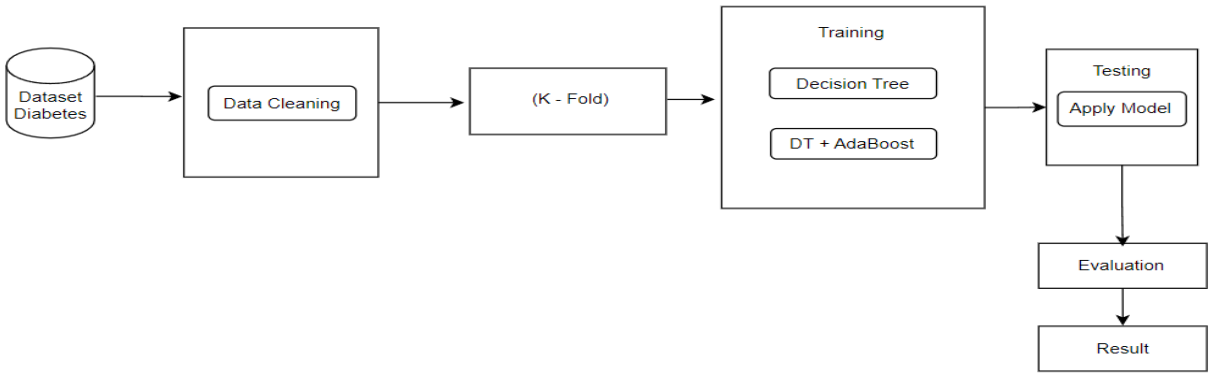

Figure 1. Research Methodology

2.1. Business Understanding

The first stage in CRISP-DM is understanding the business goals and needs from a business perspective. Based on the results of a literature study, it was found that diabetes is the deadliest disease, ranked 9th in the world. People who have diabetes have an increased risk of more severe and life-threatening health problems that can result in medical care costs, reduced quality of life, and increased mortality. With an estimated global prevalence of 9.3% in 2019, diabetes is a significant global public health problem, so computational analysis and disease prediction can help in diagnosis.

2.2. Data Understanding

This study uses public data from the Kaggle website with 520 data records. This dataset comprises 17 attributes, including age, gender, and symptoms that can influence a person's diabetes risk. This dataset's target field or label is diabetes disease status, with a negative value indicating not having the disease and a positive showing having the disease. The following dataset table is presented.

Table 1. Datasets

Age	Gender	Polydipsia	Sudden Weight Loss	…	Alopecia	Obesity	class
40	Male	Yes	No	…	Yes	Yes	Positive
58	Male	No	No	…	Yes	No	Positive
41	Male	No	No	…	Yes	No	Positive
45	Male	No	Yes	…	No	No	Positive
60	Male	Yes	Yes	…	Yes	Yes	Positive
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2.3. Data Preparation

After the data is collected, the next stage is the data preparation stage, including data cleaning. At the data cleaning stage, the initial data obtained is checked for missing/blank, noisy, and inconsistent data. If the data has missing values above 50%, the attributes can be ignored or deleted at the data cleaning stage. This can be done to overcome missing data values using the replace missing value method using the average contained in the return missing value. When processed, it produces a data pattern according to the dataset table, which explains that there is no empty/missing value or noisy or inconsistent data. With a total of 520 data.

2.4. Modelling

This stage directly involves Machine Learning to determine data mining techniques and algorithms. This research uses the Decision Tree classification model with the C4.5 algorithm, and the boosting is carried out using the Adaboost algorithm. This model will be implemented to see an immediate increase in accuracy.

1. Decision Tree C4.5

   Find tree roots. The selected attribute will be the basis for taking root, and the method for calculating the gain value for each attribute will be used; the highest gain value will be the first root. Before calculating the gain value for each attribute, calculate the entropy value first. The entropy value is calculated as follows:

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \quad (1)$$

   To calculate the Gain value, use the Equation:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (2)$$

2. Adaboost

   Then the Adaboost calculation is as follows:

   Initiation of weight values on training samples

$$D^1(i) = \frac{1}{m} \; Untuk \; i = 1, \dots, m \quad (3)$$

   Calculate the training sample error

$$€t = \sum Dt\,(i) \quad (4)$$

   Calculate the training sample weight values

$$\alpha t = \frac{1}{2} \ln \left(\frac{1 - €t}{€t}\right) \quad (5)$$

After that, update the sample weight value for correct predictions

$$Dt + 1(i) = Dt(i)x\ \{^{-\alpha t} \quad (6)$$

After that, update the sample weight value for incorrect predictions

$$Dt + 1(i) = Dt(i)x\ \{^{\alpha t} \quad (7)$$

Output of final prediction

$$H(X) = sign\ (\sum_{t=1}^{T} \alpha t\ ht(X)) \quad (8)$$

2.5. Evaluation

Looking at for the performance level of the patterns created by the algorithm is a way to carry out this stage. The evaluation algorithm uses Matrix Confusion with arrangement for accuracy, precision, and recall values. It is possible to calculate this value:

$$\text{Accuracy} = \frac{\text{Number of correct classifications}}{\text{Total testing samples tested}}\ x\ 100\% \quad (9)$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}\ x\ 100\% \quad (10)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}\ x\ 100\% \quad (11)$$

2.6. Deployment

Then, the evaluation stage is completed, where a specific and detailed assessment is made based on the results of a model. Therefore, implementation of all the models that have been created is carried out. Apart from what has been explained, adjustments are also made to the model to produce results following the initial objectives of the proposed CRISP-DM stage.

3. Results and Discussion

3.1. Modelling

At this stage, 520 records are used with data with 17 attributes used. The last attribute (Class) is the target class, so there are 16 data attributes. There are two treatments for modeling the dataset: data using the decision tree algorithm and dataset using the Decision Tree algorithm, which is optimized with Adaboost.

1. Decision Tree Modeling C4.5

   Modeling the C4.5 Decision Tree Algorithm begins by calculating the entropy value. After that, estimate the gain values of the ten attributes used to build the classification tree. The parent node is determined from the attribute with the highest gain value. From the parent node, a branch is created from the parent node category. Then, check whether there are any remaining attributes. If the condition is still there, repeat the process of calculating the entropy value. If the condition is not, continue building the C4.5 Decision Tree Algorithm, interpreting the results of the Decision Tree created, and calculating the accuracy value based on the confusion matrix. The results of calculating the entropy value using equation (1), the gain using the following equation (2):

Table 2. Calculation of Entropy and Gain Values

		Total (S)	Yes(Si)	No(Si)	Entropy	Gain
Age	23-45	203	100	103	0.9998	0.0385
	46-90	317	195	132	0.9612	

Gender	Male	328	148	180	0.9930	0.1452
	Female	192	170	22	0.5135	
Polyuria	Yes	258	230	28	0.4954	**0.3269**
	No	262	75	187	0.8638	
Polydipsia	Yes	233	220	13	0.3105	0.2802
	No	287	90	197	0.8972	
Sudden WL	Yes	217	187	30	0.5796	0.1452
	No	303	130	173	0.9853	
Weakness	Yes	305	225	80	0.8301	0.0614
	No	215	103	112	0.9986	
Polyphagia	Yes	237	191	46	0.7098	0.0948
	No	283	133	150	0.9974	
Genital	Yes	116	80	36	0.8935	0.0049
	No	404	240	164	0.9742	
Visual Blur	Yes	233	176	57	0.8026	0.0497
	No	287	145	142	0.9998	
Itching	Yes	253	153	100	0.9680	0.0004
	No	267	167	100	0.9540	
Irratability	Yes	124	110	14	0.5085	0.0864
	No	394	214	180	0.9945	
Delay heal	Yes	239	152	87	0.9460	0.0019
	No	281	165	116	0.9779	
Partial pares	Yes	224	180	44	0.7146	0.0917
	No	296	128	168	0.9867	
Muscle stif	Yes	195	130	65	0.9183	0.0303
	No	325	180	145	0.9916	
Alopecia	Yes	341	243	98	0.8653	0.0530
	No	179	79	100	0.99	
Obesity	Yes	88	53	35	0.9695	0.0181
Total(Class)		520	320	200	0.9612	

Based on the table above, it can be seen that of the 16 attributes used in this research, the Polyuria attribute has the highest gain value, namely 0.3269. This means that the polyuria attribute has the most significant influence in predicting diabetes. Then, the Polyuria attribute will be the root node.

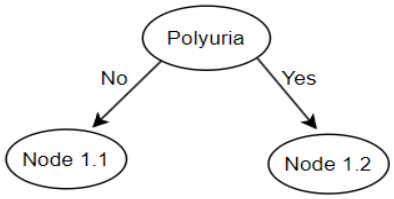

Figure 2. Node Root

Based on the root node above, the following nodes can be continued. Eliminate the previously selected attributes and repeat the calculation as at the beginning of the Entropy value, Information Gain, by choosing the largest Information Gain and making it the internal node of the tree. Repeat the calculation until all tree attributes have a class. Finally, a decision tree is produced as follows.

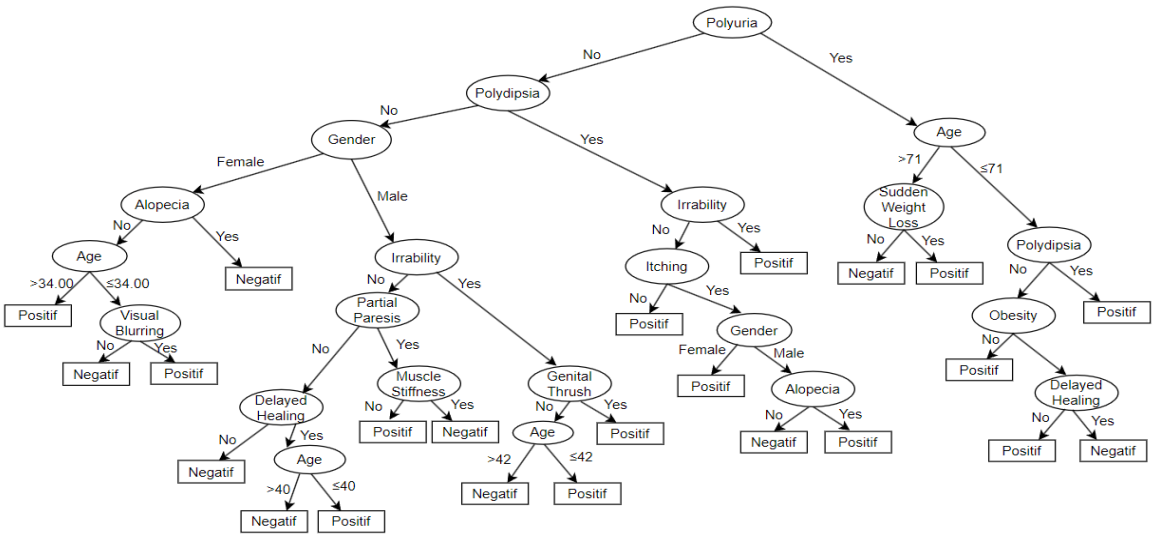

Figure 3. Decision Tree

Based on the decision tree image, a rule is formed as follows.
R1: **IF** Polyuria=No ^ Polydipsia=No ^ Gender=Female ^ Alopecia=Yes **THEN** Diabetes=Negatif
R2: **IF** Polyuria=No ^ Polydipsia=No ^ Gender=Female ^ Alopecia=No ^ Age= >34 **THEN** Diabetes=Positif
R3: **IF** Polyuria=No ^ Polydipsia=No ^ Gender=Female ^ Alopecia=No ^ Age= ≤34 ^ Visual Blurring=No **THEN** Diabetes=Negatif
R4: **IF** Polyuria=No ^ Polydipsia=No ^ Gender=Female ^ Alopecia=No ^ Age= ≤34 ^ Visual Blurring=Yes **THEN** Diabetes=Positif
R5: **IF** Polyuria=No ^ Polydipsia=No ^ Gender=Male ^ Irrability=No ^ Partial Paresis=No ^ Delayed Healing=No **THEN** Diabetes=Negatif

After the C4.5 Decision Tree Algorithm is built, the model is evaluated using a confusion matrix. The initial data is predicted based on the C4.5 Decision Tree Algorithm that has been created. It was found that 306 patients were declared to have diabetes, 184 patients were displayed not to have diabetes or were non-diabetic, and 16 patients with diabetes who were identified as non-diabetics were included in the "Type I Error." In contrast, non-diabetic patients who were identified as having diabetes, as many as 14 are included in the "Type II Error."

Table 3. Confusion Matrix Decision Tree

	true Positive	true Negative
pred. Positive	306	16
pred. Negative	14	184

The evaluation results of the C4.5 Algorithm model using the confusion matrix shown in Table 3 show an accuracy value of 94.23%. After that, modeling was carried out using the Adaboost method.

2. Adaboost Modeling
The initialization weight value of the data in the first iteration using Equation (3) with a maximum iteration of 10 is 0.00192. Data found that did not match the original class in the first iteration were 30 data. The next step is to calculate the research data error using Equation (4). The data error value

in the initial iteration is 0.0576. After calculating research data errors, the next step is to calculate the data weights using the Equation (5). The data weight value obtained was 1.3994. Using Equation (6), the data weights are updated in the first iteration where positive status is a positive result, while negative is not having diabetes. The weight of the data whose initial status was positive and was correctly predicted as positive and whose initial status was negative and correctly predicted as unfavorable was 0.000473. The weight of data whose initial status is positive and is expected to be hostile or vice versa is 0.007774.

Table 4. Hasil Update Bobot

No	Age	Gender	Polyuria	Polydipsia	Class	Class Prediction	Update Bobot
1	54.0	Female	Yes	Yes	Positive	Positive	0.000473
2	48.0	Female	Yes	Yes	Positive	Positive	0.000473
3	60.0	Male	Yes	Yes	Positive	Positive	0.000473
4	53.0	Male	Yes	Yes	Positive	Positive	0.000473
5	41.0	Male	Yes	Yes	Positive	Positive	0.000473
6	63.0	Male	Yes	Yes	Positive	Positive	0.000473
7	48.0	Female	No	No	Positive	Positive	0.000473
8	60.0	Female	Yes	Yes	Positive	Positive	0.000473
9	50.0	Female	No	Yes	Positive	Negative	0.007774
10	25.0	Female	No	No	Positive	Negative	0.007774
11	39.0	Female	Yes	Yes	Positive	Positive	0.000473
…	…	…	…	…	…	…	…

The calculation is repeated until the error value is at least 0.5 and the maximum iteration is reached. After that, the process can be stopped. Next, evaluate the model using the Adaboost method using a confusion matrix. Prediction data from the C4.5 Algorithm is predicted again based on the boosting results using the Adaboost method. It was found that 311 patients were declared to have diabetes, 195 patients were displayed not to have diabetes or were non-diabetic, and five patients with diabetes who were identified as non-diabetics were included in the "Type I Error." In contrast, non-diabetic patients who were identified as having diabetes as many as nine are included in the "Type II Error."

Table 5. Confusion Matrix Adaboost

	true Positive	true Negative
pred. Positive	311	5
pred. Negative	9	195

Based on the outcome of the evaluation of the C4.5, The algorithm model, after being boosted with a confusion matrix, uses the Adaboost method in Table 5; the accuracy value was 97.31%. This means there is an increase in the accuracy of the results obtained with the Decision Tree C4.5 algorithm. The accuracy value increased by 3% after boosting using the Adaboost method.
This is because the Adaboost method can handle samples that are difficult to predict by the decision tree model. Adaboost will improve its predictions on these samples by reducing the number of errors and increasing accuracy. The Decision Tree model that makes the best contribution to errors is given greater weight in Adaboost. Giving weight to the best model helps increase the strength of the ensemble and dominates the decisions of better decision tree models [10]. The Adaboost models produced at each iteration are combined into an ensemble model. The final prediction is made by taking the majority decision from all models. This can help overcome the weaknesses of the Decision Tree model.

4. Conclusion

This research emphasizes the importance of predicting diabetes early as an initial step in preventing the negative impacts that may arise due to this disease. With global prevalence continuing to increase, early detection is crucial to reduce the risk of potentially fatal complications. The research results show that using the Adaboost method to improve the accuracy of diabetes predictions is very effective. The integration of Adaboost with the Decision Tree algorithm increased prediction accuracy from 94.23% to 97.31%. This shows that ensemble learning can be an effective solution to improve the performance of a single algorithm. The findings of this study effectively highlight the potential of the Adaboost method in improving the accuracy of diabetes prediction. By improving predictions on samples that are difficult for Decision Tree models to predict, Adaboost manages to reduce the number of errors and increase overall accuracy. This suggests that Adaboost could be helpful in early disease detection efforts. With increased prediction accuracy, healthcare can provide patients with more precise diagnoses and earlier treatment. This can help in reducing the risk of potentially fatal complications due to diabetes. This research still does not have any treatment for exploration that can further explore the influence of Adaboost parameters, such as the number of iterations (T) that can influence the results of Adaboost [11].

References

[1] Kementerian Kesehatan RI., "Infodatin tetap produktif, cegah, dan atasi Diabetes Melitus 2020," *Pusat Data dan Informasi Kementerian Kesehatan RI*. pp. 1–10, 2020.

[2] S. Ucha Putri, E. Irawan, F. Rizky, S. Tunas Bangsa, P. A. -Indonesia Jln Sudirman Blok No, and S. Utara, "Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5," *Januari*, vol. 2, no. 1, pp. 39–46, 2021.

[3] M. R. Firmansyah and Y. P. Astuti, "Stroke Classification Comparison with KNN through Standardization and Normalization Techniques," *ASSET*, vol. 6, no. 1, pp. 1–8, 2024.

[4] P. Songthung and K. Sripanidkulchai, "Improving type 2 diabetes mellitus risk prediction using classification," *2016 13th Int. Jt. Conf. Comput. Sci. Softw. Eng. JCSSE 2016*, 2016, doi: 10.1109/JCSSE.2016.7748866.

[5] M. Komi and X. Zhang, "Application of Data Mining Methods in Diabetes Prediction Messan," no. S Ix, pp. 1006–1010, 2017.

[6] R. S. Raj, D. S. Sanjay, M. Kusuma, and S. Sampath, "Comparison of Support Vector Machine and Naïve Bayes Classifiers for Predicting Diabetes," *1st Int. Conf. Adv. Technol. Intell. Control. Environ. Comput. Commun. Eng. ICATIECE 2019*, pp. 41–45, 2019, doi: 10.1109/ICATIECE45860.2019.9063792.

[7] L. Amini *et al.*, "Prediction and control of stroke by data mining," *Int. J. Prev. Med.*, vol. 4, pp. S245–S249, 2013.

[8] T. Kansadub, S. Thammaboosadee, S. Kiattisin, and C. Jalayondeja, "Stroke risk prediction model based on demographic data," *BMEiCON 2015 - 8th Biomed. Eng. Int. Conf.*, pp. 3–5, 2016, doi: 10.1109/BMEiCON.2015.7399556.

[9] T. Asra, A. Setiadi, M. Safudin, E. W. Lestari, N. Hardi, and D. P. Alamsyah, "Implementation of AdaBoost Algorithm in Prediction of Chronic Kidney Disease," *2021 7th Int. Conf. Eng. Appl. Sci. Technol. ICEAST 2021 - Proc.*, pp. 264–268, 2021, doi: 10.1109/ICEAST52143.2021.9426291.

[10] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. 2011. doi: 10.1016/C2009-0-61819-5.

[11] A. H. Yunial, "Optimization Analysis of Support Vector Machine Classification Algorithms, Decision Trees, and Neural Networks Using Adaboost and Bagging," *J. Inform. Univ. Pamulang*, vol. 5, no. 3, p. 247, 2020.