BAREKENG: Journal of Mathematics and Its Applications
March 2024 Volume 18 Issue 1 Page 0193–0204
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss1pp0193-0204

IMPLEMENTATION OF MACHINE LEARNING ALGORITHM C4.5 IN CLASSIFICATION OF PATIENTS WITH TYPE 2 DIABETES MELLITUS

Dyah Ayu Sekar Kinasih Purwaningrum [1], Dina Agustina [2*]

[1,2]Mathematics Department, Mathematics and Science Faculty, Padang State University
Prof. Dr. Hamka Street, Air Tawar Padang, Sumatera Barat, 25132, Indonesia

*Corresponding author's e-mail: * **dinagustina@fmipa.unp.ac.id***

Article History:

Received: 12th August 2023
Revised: 24th November 2023
Accepted: 25th December 2023

Keywords:

C4.5 Algorithm;
Classification;
Diabetes Mellitus;
Machine Learning.

ABSTRACT

The neglect of a healthy lifestyle among the Indonesian population has led to an increased risk of diabetes mellitus, which currently affects 643 million people worldwide. Early and accurate diagnosis is crucial for preventing the progression of the disease. This study utilized the C4.5 machine learning algorithm to develop a model to classify individuals as diabetic or non-diabetic based on diabetes-associated factors. The data used in this research consisted of medical records from patients with and without diabetes at Padang General Hospital. The model's performance evaluation resulted in a recall value of 91%. By promoting a healthy lifestyle and raising awareness about the importance of regular check-ups, the burden of diabetes can be reduced, and the overall health of the population can be improved.

How to cite this article:

D. A. S. K. Purwaningrum and D. Agustina., "IMPLEMENTATION OF MACHINE LEARNING ALGORITHM C4.5 IN CLASSIFICATION OF PATIENTS WITH TYPE 2 DIABETES MELLITUS," *BAREKENG: J. Math. & App.*, vol. 18, iss. 1, pp. 0193-0204, March, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

The adoption of a healthy lifestyle can significantly reduce the risk of various diseases [1]. However, many people in Indonesia still tend to overlook this aspect; this fact is supported by a survey conducted by AIA Group across 15 Asia-Pacific countries, which ranked Indonesia 11th in terms of the adoption of a healthy lifestyle [2]. Consequently, the Indonesian population remains vulnerable to various diseases, one of which is diabetes mellitus.

Diabetes mellitus, commonly referred to as diabetes, is a disease primarily caused by hyperglycemia, which is the accumulation of glucose in the bloodstream. Nevertheless, it is believed that the cause of type 2 diabetes is also influenced by factors, such as excessive body mass index, aging, and family history [3]. If the accumulation of glucose is due to the immune system attacking the pancreas's insulin-producing cells, it is classified as type 1 diabetes. On the other hand, when insufficient insulin is produced, or the body's cells do not effectively utilize the insulin hormone, it is categorized as type 2 diabetes [4]. The number of cases for both types of diabetes continues to rise, but type 2 diabetes constitutes a larger ratio, accounting for 90% of all diabetes cases [3].

As of 2021, approximately 643 million individuals worldwide were affected by diabetes. Indonesia ranked fifth among countries with the highest number of diabetes patients, reaching 19.5 million cases, and it is predicted to increase further to 28.6 million by 2045 [3]. The alarming statistics necessitate various efforts to prevent the escalating number of diabetes patients, and one of these approaches involves utilizing medical record data [5]. In the realm of healthcare, a wealth of medical record data exists that could become invaluable if properly leveraged [6]**.** Thus, the effective utilization of medical record data is vital to gather information for curbing the prevalence of diabetes. Machine learning is a method that can facilitate this process.

Previous research has demonstrated that machine learning approaches improve the accuracy of predicting disease risk factors compared to conventional methods [7]. Mercaldo employed a machine-learning decision tree to predict diabetes based on glucose levels, body mass index, age, and other attributes[8]. Similarly, Azrar implemented the C4.5 algorithm to classify women experiencing diabetes [9]. Using the Pima Indian Diabetes Dataset, the highest accuracy was achieved using the C4.5 algorithm, reaching 75.65%. Further confirmation of C4.5's superiority in classifying diabetes patients over other machine-learning methods was provided by Ente's work [6].

Based on the description above, this research aims to build and determine the performance of the C4.5 algorithm in classifying type 2 diabetes based on its risk factors. As in the previous research, the C4.5 algorithm can classify diabetic patients with relatively good accuracy scores. The selection of the C4.5 algorithm is also due to its advantage: it is easy to interpret and interesting because it can be visualized in images [10]. The difference between this research and previous studies lies in the research feature. This research includes gender as a research feature. Therefore, unlike the other research, this research is applicable to both men and women patients.

2. RESEARCH METHODS

2.1 Diabetes

Diabetes is a disease caused by hyperglycemia, characterized by the accumulation of glucose in the bloodstream. However, several other factors are believed to contribute to type 2 diabetes, including:

a. Glucose Level

Elevated glucose levels are a primary indicator of diabetes. Glucose levels are categorized as high when they exceed 200 mg/dL [3].

b. Body Mass Index (BMI)

BMI is a method of categorizing body mass or body fat. Excessive body fat accumulation can lead to type 2 diabetes, as excess body fat can disrupt cellular function, resulting in insulin resistance[11]. An individual is classified as having excess body fat if their BMI is greater than 25.0.

c. Gender

Men tend to have a higher amount of visceral fat (fat around the abdominal area) compared to women. In contrast, women have more subcutaneous fat (fat around the thigh area) relative to visceral fat. Visceral fat is more metabolically active than subcutaneous fat, making men more susceptible to obesity, which is closely linked to BMI and diabetes.

d. Age

Increasing age can lead to a decreased sensitivity of body cells to insulin, resulting in elevated glucose levels in the body [12]. Additionally, Song [13] states that the prevalence of diabetes is not limited to the elderly population; it continues to rise among young adults (18-27 years) and middle-aged adults (28-40 years). This is believed to be due to an unhealthy lifestyle.

e. Family History

Diabetes involves a genetic component [14]. Genetic factors contributing to diabetes can be passed down to children if their parents have diabetes [15].

2.2 C4.5 Algorithm

The C4.5 algorithm is a decision tree method used to analyze the relationship between input and output attributes. To construct a decision tree using the machine learning C4.5 algorithm, data is randomly divided into training and testing datasets. The training data is used to build the model, while the testing data is used to evaluate the performance of the established model. After data splitting, the decision tree is formed by following the flowchart shown in Figure 1.

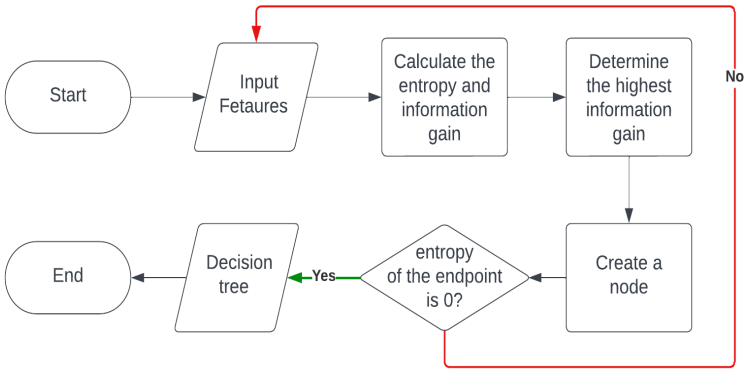


Figure 1. C4.5 Algorithm Flowchart

As defined in Figure 1, the decision tree is constructed by calculating entropy and information gain. The selection of which attribute becomes a node is determined by the highest information gain value. This process continues iteratively until all entropies at the decision tree's endpoints reach zero. Entropy and information gain are defined as follows:

Definition 1. Entropy is a parameter used to quantify the level of uncertainty associated with an attribute [16] [17].

$$Ent(D) = -\Sigma_{k=1}^{|y|} p_k \, log_2 \, p_k \tag{1}$$

Definition 2. Information gain is used to calculate the effectiveness of a feature in classifying data [17].

$$IG(D,a) = Ent(D) - \Sigma_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) \tag{2}$$

where:

$Ent(D)$: entropy of dataset D

k : class of dataset D

$|y|$: total of class

p_k : k-th class probability

$IG(D, a)$: information gain of feature a

V : possible value of feature a

Given the dataset D and the continuous feature a, suppose we observe n values of a in D and arrange them in ascending order $\{a^1, a^2, \ldots, a^n\}$. With the split point t, D is partitioned into the subsets D_t^- and D_t^+, where D_t^- includes samples with values t and less than t, and D_t^+ includes samples with values greater than t.

$$T_a = \left\{ \frac{a^i + a^{i+1}}{2}, 1 \leq i \leq n-1 \right\} \quad (3)$$

2.3 Performance Measure

This classification of diabetes disease is classified into two classes: diabetes and non-diabetes, making it suitable for binary classification, and enabling the utilization of a confusion matrix to measure the model's performance [18]. Table 1 illustrates the structure of a confusion matrix.

Table 1. Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

As shown in Table 1, TP (True Positive) signifies the correct identification of a diabetic patient as diabetic, while TN (True Negative) signifies the accurate identification of a non-diabetic patient as non-diabetic. FN (False Negative) represents the misclassification of a diabetic patient as non-diabetic, and FP (False Positive) denotes the incorrect classification of a non-diabetic patient as diabetic.

The information of confusion matrix can be used to calculate accuracy, precision, recall, and f-1 score [18], as noted in Table 2.

Table 2. Performance Measure Metrics

Metric	Description	Formula
Accuracy	The model's correct prediction rate	(TP+TN)/ (TP+TN+FP+FN)
Precision	The ability of the model to not identify negative data as positive.	TP/ (TP+FP)
Recall	The ability of the model to not identify positive data as negative.	TP/ (TP+FN)
f-1 score	Shows an excellent level of precision and Recall	2 x [(Precision x Recall)/ (Precision + Recall)]

The metrics in Table 2 can be used as the model performance evaluation. The higher the metric value, the better the model is built. Other than in metric form, model performance can also be graphically represented as a Receiver Operating Characteristic (ROC) Curve. By modifying the threshold, an ROC curve is generated based on the True Positive Rate (TPR) and False Positive Rate (FPR).

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN}$$
$$False\ Positive\ Rate\ (FPR) = \frac{FP}{TN + FP} \quad (4)$$

The ROC curve basically represents the combination of multiple confusion matrices, which can then provide insight into the model's ability to distinguish between two classes. Figure 2 portrays an example of the ROC curve:

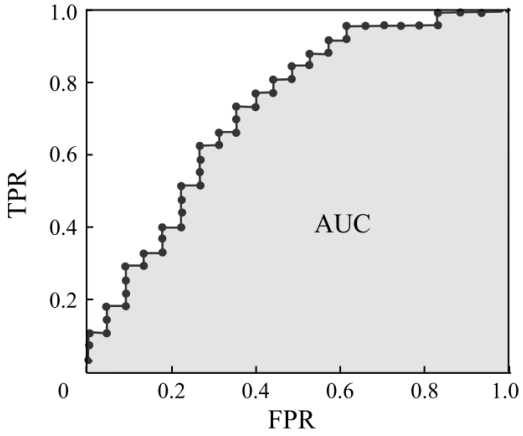


Figure 2. ROC curve and AUC

Figure 2 shows that the ROC curve lies in the range [0,1] of the FPR and TPR areas. The ideal point of ROC is at (0,1). At this point, the model correctly classifies each class**.** Based on **Figure 2,** there is also AUC, which stands for Area Under Curve. Besides ROC, the AUC can also be used to evaluate the quality of the model. The higher the AUC value, the better the model at distinguishing between two classes.

2.4 Data Description

Data was collected from the medical records of diabetes and non-diabetes patients at Padang Central Hospital from 2018 to 2022. **Figure 3** depicts a snippet of the dataset used.

Table 3. Dataset

No	Glucose	BMI	Gender	Age	Family History	Class
1	180	35.2	Male	54	No History	Diabetes
2	139	32.0	Male	58	No History	Non-Diabetes
3	159	26.4	Male	28	History	Diabetes
4	188	48.8	Male	37	History	Diabetes
5	131	23.8	Female	31	No History	Non-Diabetes
…	…	…	…	…	…	…
354	193	40.9	Female	47	History	Diabetes
355	110	34.2	Male	27	History	Diabetes
356	172	33.3	Male	52	History	Diabetes
357	197	37.1	Male	43	History	Diabetes
358	171	42.4	Female	29	History	Diabetes

Based on **Table 3,** the dataset comprises 6 columns and 358 rows. The six columns are: glucose, BMI, gender, age, family history, and class. Furthermore, the description of the dataset's features is presented in **Table 4**.

Table 4. Dataset Feature Description

Feature	Description	Data Type
Glucose (X1)	Blood glucose level (mg/dL)	Numerical
BMI (X2)	Body Mass Index (kg/(m)^2)	Numerical
Gender (X3)	Gender of the patient (Male or Female)	Categorical
Age (X4)	Age of the patient (in years)	Numerical
Family History (X5)	Presence or absence of diabetes in the family	Categorical
Class (Y)	Class label indicating diabetes or non-diabetes	Categorical

As shown in **Table 4,** the utilized input features are contributing to diabetes, namely glucose level (X1), BMI (X2), gender (X3), age (X4), and family history (X5). Among these five input features, three are numerical features: glucose, BMI, and age, while the other two are categorical features: gender and family history. The statistical descriptions of the numerical features are presented in **Table 5.**

Table 5. Statistical Description of Numerical Features

	Glucose	BMI	Age
Count	358	358	358
Mean	140.7737	33.1315	34.3045
Std	42.1652	7.0987	12.0149
Min	70	18.2000	21
25%	103	27.8000	25
50%	139	32.8500	30.5000
75%	178	37.5500	42
Max	300	59.4000	81

According to **Table 5**, all numerical features have the same data count, which is 358, indicating that there are no missing values for any feature. There is more statistical data information, such as glucose is within the range [70, 300], with a mean of 140.7737, a standard deviation of 42.1652, and the 25th, 50th, and 75th percentile, respectively, are 103, 139, and 178. The statistical description of categorical features can be seen in **Table 6.**

Table 6. Statistical Description of Categorical Features

	Gender	Family History
Total	358	358
Unique	2 (male, female)	2 (history, no history)
Modus	Female	No history
Total of Modus	197	184

Similar to the numerical features, according to **Table 6,** the categorical features also do not have any missing values, as each feature has a count of 358. Additionally, there is more statistical information. For instance, the "Gender" feature has two unique values: 'Female' and "Male". The mode of the "Gender" feature is "Female", with a count of 197 data.

The class feature (Y) acts as an output feature and is classified into two classes: diabetes and non-diabetes. The distribution of the dataset classes is shown in **Figure 3**:

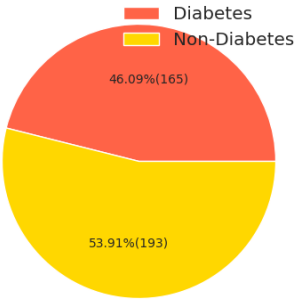
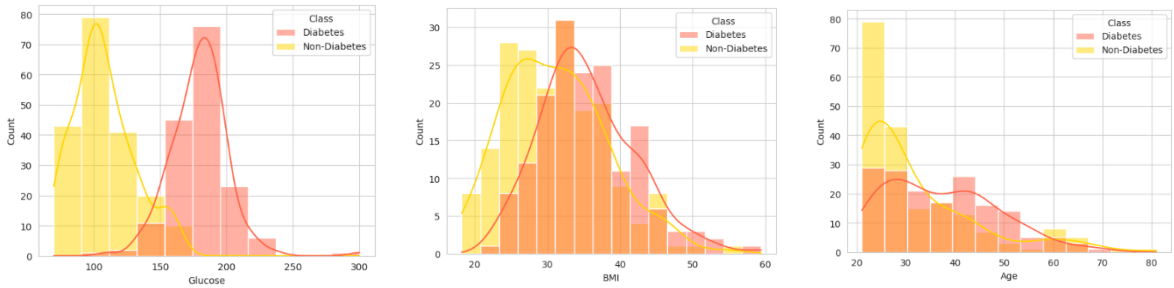


Figure 3. Dataset Classes Distribution

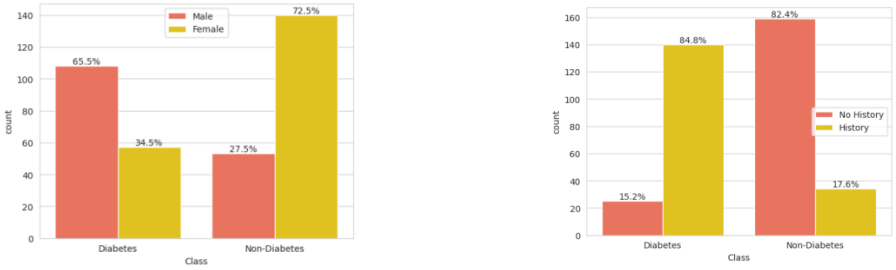
Based on **Figure 3**, there are 53.91% data for non-diabetes and 46.09% data for diabetes. It shows that the proportions of the dataset classes are almost the same. A further distribution of dataset classes of numerical features can be seen in **Figure 4.**



(a) **(b)** **(c)**

Figure 4. Dataset Classes Distribution of (a) Glucose Feature, (b) BMI Feature, and (c) Age Feature

In Figures 4 part (a) and (b), the histogram data show that diabetes tends to occur in people with high blood glucose and BMI levels. In terms of age, Figure 4(c) indicates that diabetes is more prevalent among individuals aged 20 to 40 years. The distribution of dataset classes with respect to gender and family history is shown in Figure 5 below:



(a) **(b)**

Figure 5. Dataset classes distribution of (a) gender feature and (b) family history feature

Figure 5 (a) shows that men have a higher percentage of experiencing diabetes compared to women, with a rate of 65.5%. Then, individuals with a history of diabetes are at higher risk, with an 84.8% chance of having diabetes, as shown in Figure 5 (b).

3. RESULTS AND DISCUSSION

The dataset is randomly split into training data and testing data at a ratio of 70:30, 80:20, and 90:10. Using Equation (3), the 80:20 dataset generates 285 thresholds for each of its numerical features: $\boldsymbol{T_{glucose}} = \{\mathbf{71}, \ldots, \mathbf{223.5}\}$, $\boldsymbol{T_{BMI}} = \{\mathbf{18.2}, \ldots, \mathbf{57.2}\}$, and $\boldsymbol{T_{age}}\ \{\mathbf{21}, \ldots, \mathbf{71}\}$. The entropy and information gain calculations to select the root node are presented in Table 7.

Table 7. Selecting Root Node

Attribute		Total Sample	Diabetes	Non-Diabetes	*Entropy*	*Information Gain*
		286	128	158	0.99205	
Glucose	≤ 71	3	0	3	0.00000	0.00904
	> 71	283	128	155	0.99342	
	...	...	...	...	...	...
	≤ 134.5	144	3	141	0.14609	0.65606
	> 134.5	142	125	17	0.52855	
	...	...	...	...	...	...
	≤ 223.5	285	127	158	0.99145	0. 00407
	> 223.5	1	1	0	0.00000	
BMI	≤ 18.2	2	0	2	0.00000	0.00602
	> 18.2	284	128	156	0.99298	
	...	...	...	...	...	...
	≤ 57.2	285	127	158	0.99145	0.00407
	> 57.2	1	1	0	0.00000	
Gender	Male	159	46	113	0.86783	0.09312
	Female	127	82	45	0.93788	
Age	≤ 21	285	128	157	0.66658	0.02102
	> 21	1	0	1	0.99765	
	...	...	...	...	...	...
	≤ 71	284	127	157	0.99252	0.00300
	> 71	2	1	1	0.00000	
Family History	History	142	112	30	0.74390	0.36931
	No-History	144	16	128	0.50326	

In **Table 7**, for each possible value of the attributes, the entropy of the subsets created by splitting on that value, the highest information gain is glucose with a threshold of 134.5. Therefore, the decision tree is formed as **Figure 6** follows:

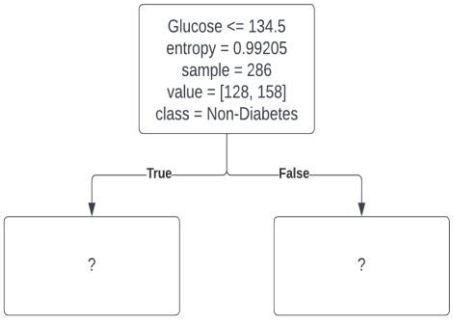


Figure 6. The temporary form of a decision tree

As stated in **Figure 6**, It is observed that the root node "glucose ≤ 134.5" contains 286 training data, comprising 128 diabetes and 158 non-diabetes cases. Since the majority class within the root node is "non-diabetes", this node is labeled as "non-diabetes". The combination of the two classes in the root node results in a relatively high entropy value of 0.99, indicating that the root node has not yet reached purity. Therefore, feature splitting is necessary until leaf nodes with entropy 0 are reached, indicating nodes that have achieved purity (containing only one class of dataset). The result of the decision tree with an 80:20 splitting dataset with the program implemented in Google Collaboratory is shown in **Figure 7**.

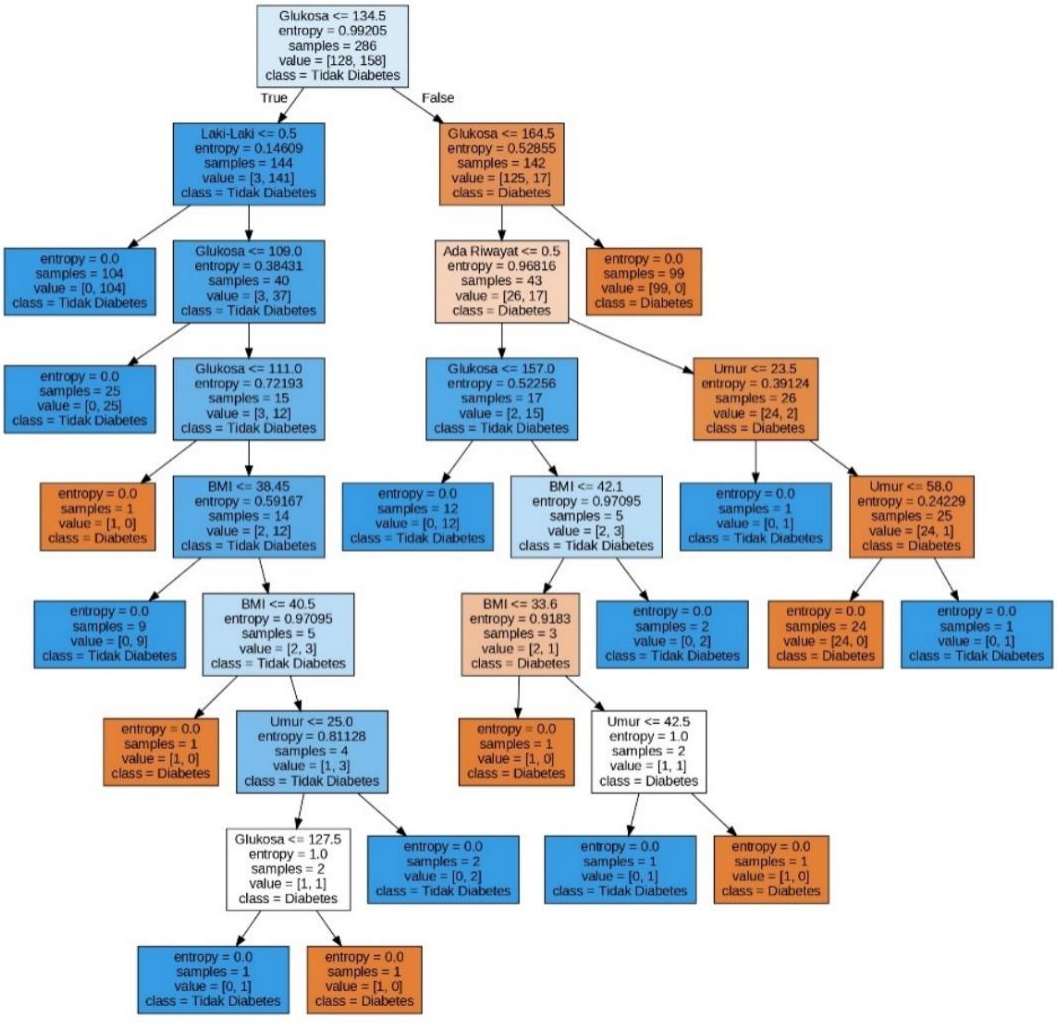
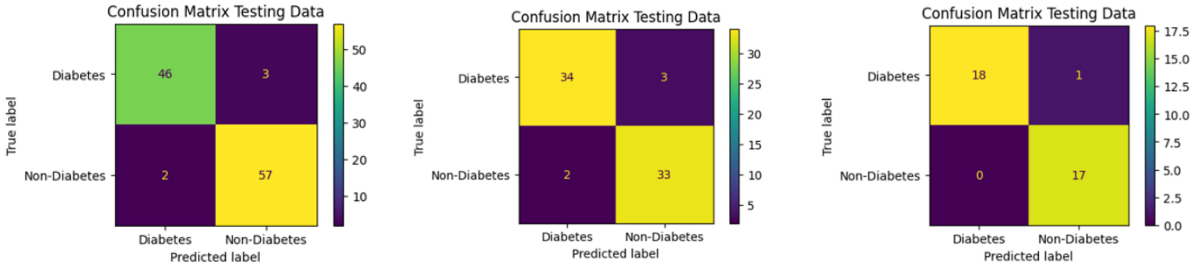


Figure 7. Final Decision Tree of 80:20 dataset ratio

As shown in **Figure 7** above, the generated decision tree comprises 33 nodes. The blue-colored nodes represent instances where the diabetes class is the majority, the orange-colored nodes represent instances where the non-diabetes class is the majority, and the white-colored nodes signify an equal ratio of both classes. The shade of the color reflects the purity of the node; darker shades indicate higher purity.

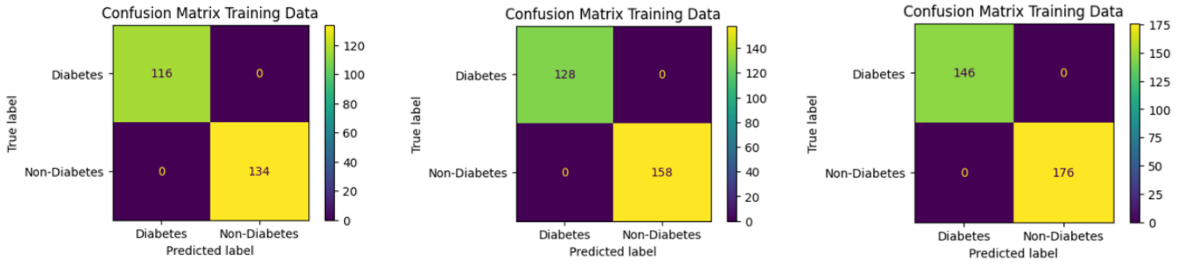
The decision tree traversal begins from the root node. If a patient's glucose level is equal to or lower than 134.5, the decision process follows the left branch; otherwise, it proceeds along the right branch. This process continues iteratively until a leaf node is reached, classifying the patient into a specific dataset class.

The same approach was applied to split the dataset at ratios of 70:30, 80:20 and 90:10, produce the matrices in **Figure 8** and **Figure 9** below:



(a) **(b)** **(c)**

Figure 8. Confusion matrix of testing data with (a) 70:30 dataset, (b) 80:20 dataset, and (c) 90:10 dataset

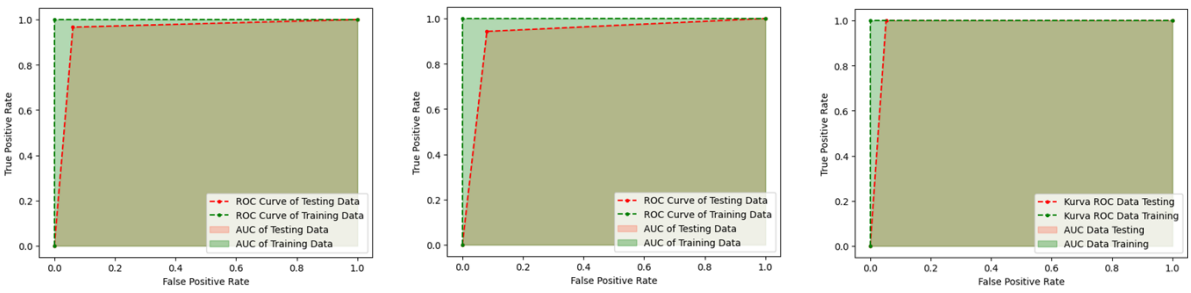


(a) **(b)** **(c)**

Figure 9. Confusion matrix of training data with (a) 70:30 dataset, (b) 80:20 dataset, and (c) 90:10 dataset

Figure 8 and **Figure 9** represent the results of a confusion matrix for determining the values of true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) from Table 1. For instance, in Figure 9 (a), out of 108 testing data points, there were 46 TP values, 3 FN values, 2 FP values, and 57 TN values. This data is used to calculate accuracy, precision, recall, and F1-score, as presented in **Table 2**. The TP and TN values are larger than the FP and FN values. This pattern holds true for their respective training dataset as well, with some splits resulting in all zero FP and FN values. They suggest that the model is classifying more data correctly than incorrectly.

Based on the generated confusion matrices, the ROC curves are formed as **Figure 10** follows:



(a) **(b)** **(c)**

Figure 10. ROC curve of (a) 70:30 dataset, (b) 80:20 dataset, and (c) 90:10 dataset

As shown in **Figure 10**, the ROC curves obtained for the testing data are closely situated near the point (0,1), indicating that the model performs well for classification. Furthermore, due to the absence of false positives and false negatives, the ROC curves for the training data display a perfect shape. These observations imply that the model excels at distinguishing between the diabetes and non-diabetes classes. In addition to the graphical representation of the ROC curves, the model's performance can also be summarized in **Table 8**.

Table 8. Metrics of Performance Evaluation

	Splitting Dataset	Accuracy	Precision	Recall	f-1 score	AUC
Testing Data						
	70:30	0.9500	0.9500	0.9300	0.9400	0.9520
	80:20	0.9300	0.9400	0.9100	0.9300	0.9310
	90:10	0.9700	0.1000	0.9400	0.9700	0.9740
Training Data						
	70:30	0.1000	0.1000	0.1000	0.1000	0.1000
	80:20	0.1000	0.1000	0.1000	0.1000	0.1000
	90:10	0.1000	0.1000	0.1000	0.1000	0.1000

Based on **Table 8**, the highest recall value for testing data is achieved in the dataset with a ratio of 90:10. A recall value of 0.940, or 94% in percentage terms, indicates that out of the total number of individuals who are diabetic, 94% of them are correctly predicted as diabetic by the model. This reflects the model's ability to effectively identify a significant portion of true positive cases among all actual positive cases. For training data, the accuracy for each splitting dataset is 100%.

4. CONCLUSIONS

Based on the data obtained from the medical records of diabetic and non-diabetic patients at Padang Central Hospital, diabetes was categorized based on its risk factors, including blood glucose level, BMI, gender, age, and family history, utilizing the C4.5 algorithm. A thorough analysis of the data indicated that diabetes tends to manifest in individuals with elevated blood glucose levels and higher BMIs, particularly among males aged 20 to 40 who have a family history of diabetes. Utilizing this information, a decision tree model was constructed using the Google Colaboratory platform. The dataset was divided into training and testing sets at various ratios: 70:30, 80:20, and 90:10. The model with a 90:10 training and testing data ratio achieved the highest recall value of 94%. This signifies that the model effectively identifies 94% of the actual diabetic cases among the entire set of individuals who have diabetes.

REFERENCES

[1] World Health Organization, "Health and Development Through Physical Activity and Sport," Geneva, 2003.
[2] AIA Group, "The AIA Healthy Living Index 2018," 2018. Accessed: Jun. 17, 2023. [Online]. Available: www.aia-financial.co.id
[3] International Diabetes Federation, "IDF Diabetes Atlas 10th edition," 2021. [Online]. Available: www.diabetesatlas.org
[4] M. Maniruzzaman *et al.*, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Comput Methods Programs Biomed*, vol. 152, pp. 23–34, Dec. 2017, doi: 10.1016/j.cmpb.2017.09.004.
[5] R. FS, B. A. Tama, and M. Mulya, "Pengembangan Perangkat Lunak Diagnosa Penyakit Diabetes Mellitus Tipe II Berbasis Teknik Klasifikasi Data," 2010, Accessed: Jun. 30, 2023. [Online]. Available: https://repository.unsri.ac.id/8352/1/Pengembangan_Perangkat_Lunak_Diagnosa_Penyakit_Diabetes_Mellitus_Tipe_Ii_Berbasis_Teknik_Klasifikasi_Data.pdf
[6] D. R. Ente, S. Arifin, Andreza, and S. A. Thamrin, "Comparison of C4.5 algorithm with naive Bayesian method in classification of Diabetes Mellitus (A case study at Hasanuddin University hospital Makassar)," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Nov. 2019. doi: 10.1088/1742-6596/1341/9/092009.
[7] A. S. Selya and D. Anshutz, "Machine learning for the classification of obesity from dietary and physical activity patterns," in *Smart Innovation, Systems and Technologies*, Springer Science and Business Media Deutschland GmbH, 2018, pp. 77–97. doi: 10.1007/978-3-319-77911-9_5.
[8] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques," *Procedia Comput Sci*, vol. 112, pp. 2519–2528, 2017, doi: 10.1016/j.procs.2017.08.193.
[9] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data Mining Models Comparison for Diabetes Prediction," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 320–323, 2018, [Online]. Available: www.ijacsa.thesai.org
[10] F. Gorunescu, *Computational intelligence: collaboration, fusion and emergence*. Berlin: Springer, 2009.

[11] S. Klein, A. Gastaldelli, H. Yki-Järvinen, and P. E. Scherer, "Why does obesity cause diabetes?" *Cell Metabolism*, vol. 34, no. 1. Cell Press, pp. 11–20, Jan. 04, 2022. doi: 10.1016/j.cmet.2021.12.012.
[12] K. Mordarska and M. Godziejewska-Zawada, "Diabetes in the elderly," *Przeglad Menopauzalny*, vol. 16, no. 2, pp. 38–43, 2017, doi: 10.5114/pm.2017.68589.
[13] S. H. Song, "Emerging Type 2 Diabetes In Young Adults," Sheffield, 2012.
[14] T. A. Harrison *et al.*, "Family history of diabetes as a potential public health tool," *American Journal of Preventive Medicine*, vol. 24, no. 2. Elsevier Inc., pp. 152–159, 2003. doi: 10.1016/S0749-3797(02)00588-3.
[15] R. Watta, G. Masi, M. E. Katuuk, M. Program Studi Ilmu Keperawatan Fakultas Kedokteran Universitas Sam Ratulangi, and P. Studi Ilmu Keperawatan Fakultas Kedokteran, "Screening Faktor Resiko Diabetes Melitus Pada Individu Dengan Riwayat Keluarga Diabetes Melitus Di Rsud Jailolo," *Jurnal Keperawatan (JKp)*, vol. 8, pp. 44–50.
[16] E. Alpaydin, *Introduction to Machine Learning*, 3rd ed. London: The MIT Press, 2014.
[17] Z.-H. Zhou, *Machine Learning*. Nanjing: Tsinghua University Press, 2021. doi: 10.1007/978-981-15-1967-3.
[18] B. Rajoub, "Supervised and unsupervised learning," in *Biomedical Signal Processing and Artificial Intelligence in Healthcare*, Elsevier, 2020, pp. 51–89. doi: 10.1016/B978-0-12-818946-7.00003-2.