Variance
Journal of Statistics and Its Applications
P-ISSN: 2685-8738 E-ISSN: 2685-872X

Variance: Journal of Statistics and Its Applications
Volume 6 Issue 2 | October 2024
10.30598/variancevol6iss2page165-174

APPLICATION OF C4.5 ALGORITHM WITH FEATURE SELECTION IN CLASSIFICATION OF DISCHARGE STATUS OF HEAD INJURY PATIENTS

Putri[1*], Evy Sulistianingsih[2], Nurfitri Imro'ah[3], Naomi Nessyana Debataraja[4]
[1,2,3,4] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Tanjungpura
Prof. Dr. H. Hadari Nawawi Street, Pontianak, 78124, West Kalimantan, Indonesia

Corresponding Author's E-mail: *h1091201055@student.untan.ac.id*

Abstract: *Head trauma is a medical emergency that can cause brain damage and disability, leading to death. The discharge status of injured patients is classified into two: alive and dead. The purpose of this study is to apply the C4.5 algorithm without feature selection and by using Chi-Square and Mutual Information feature selection to show independent variables that significantly influence the discharge status of head injury patients. This research data is secondary data of patients who suffered head injuries at Dr. Abdul Aziz Hospital, Singkawang City, in 2019-2021. The independent variables used were age, gender, length of hospitalization, etiology of head injury, Suprasellar Cistern, and Glasscow Coma Scale, with the dependent variable being discharge status. Based on the study results, the Chi-Square feature selection results identified two variables that had a significant effect. In contrast, for the Mutual Information feature selection results, five variables had a significant impact on the dependent variable. The C4.5 Algorithm classification model without feature selection produces an accuracy of 88.57%, the C4.5 Algorithm classification model with Chi-Square feature selection produces an accuracy of 88.57%, and the C4.5 Algorithm classification model with Mutual Information feature selection produces an accuracy value of 91.42% with the highest accuracy obtained from the results of the C4.5 Algorithm model formation with Mutual Information feature selection.*

Keywords: *C4.5 Algorithm, Chi-Square, Mutual Information.*

1. INTRODUCTION

Traumatic brain injury is an emergency that can cause damage to the human brain and skull so that a head injury patient will experience several symptoms such as nausea, vomiting, unconsciousness, bleeding in the Ear, Nose, and Throat, headache, to severe cases that can cause memory loss, convulsions, and death [1]. Based on the results of the basic health research report (RISKESDAS) of West Kalimantan Province in 2019, it was found that the prevalence of head injuries in West Kalimantan Province reached 11.3%. Singkawang City is ranked second with the highest head injuries in West Kalimantan Province [2]. In general, head injury cases can occur due to falling, being hit by an object, traffic accidents, being hit, assault, and self-harm [3].

According to the neurosurgery department, there are several variables used in assessing the medical record data of head injury patients, namely gender, age, etiology of head injury, length of hospitalization, and Glasscow Coma Scale [4]. n addition, in 2014, an emphasis on predictive values was proposed by introducing Suprasellar Cistern (SSC) status. Thus, the variables used in the study were age, gender, length of hospitalization, etiology of head injury, Suprasellar Cistern, and Glasscow Coma Scale [5]. Six independent variables are used to show the variables that have the most significant influence on the dependent variable. In head injury cases, one of the efforts that can be made to overcome head injury cases is to provide effective and efficient treatment measures to minimize the exit status of head injury patients with the worst conditions. The discharge status of head injury patients is divided into two, namely, alive and dead. The classification of head injury patients' discharge status can be analyzed using one of the data mining algorithms. Data mining is a data analysis stage that uses statistical, mathematical, artificial intelligence, and machine learning methods. The aim is to explore helpful information for explaining a decision [6]. Classification is one of the main functions of data mining. The C4.5 algorithm is an example of a classification algorithm in data mining that combines

https://ojs3.unpatti.ac.id/index.php/variance/ *Submitted: 2024-03-21* *Accepted: 2024-10-30*

data analysis and data pattern search applied to decision trees to produce results in the form of a rule that is used to make specified decisions[7].

2. METHODOLOGY

2.1. Research Variables

The data used in this study are data on head injury patients at Dr. Abdul Azis Hospital, Singkawang City, from January 2019 - December 2021, with the inclusion and exclusion criteria of the sample set. The number of data samples used was 70 data, with independent variables, namely age (X_1), gender (X_2), length of hospitalization (X_3), etiology of head injury (X_4), Suprasellar Cistern (SSC) (X_5), and Glasscow Coma Scale (GCS) (X_6), as well as the dependent variable, namely discharge status (Y) which is classified into living status and dead status.

2.2. Stages of Research

This research uses two C4.5 Algorithm approaches: Algorithm C4.5 without and applying feature selection. The data analysis process in this study differs in the selection stage of independent variables that will be used in the C4.5 Algorithm decision tree formation process. Feature selection is a data mining technique that aims to reduce the complexity of variables in the data mining process. Feature selection is intended to identify variables that significantly influence the target variable to make the computational process faster and more efficient and have a better model interpretation [8]. In forming a decision tree without feature selection, all independent variables that have been previously selected are used as a whole, while in the C4.5 Algorithm using feature selection, each independent variable to be used is tested using the Chi-Square and Mutual Information methods. The Chi-Square value can be calculated using Equation (1) [9]:

$$\chi^2 = \sum_{r=1}^{k}\sum_{c=1}^{k}\frac{(O_{rc} - E_{rc})^2}{E_{rc}} \tag{1}$$

where k is the number of categories in the independent feature and dependent feature, O_{rc} is the frequency in the row-r and column-c, E_{rc} is the expected frequency in the row-r and column-c which can be calculated using Equation (2).

$$E_{rc} = \frac{n_r \times n_c}{S} \tag{2}$$

where S is the total number of samples used, n_r is the number of samples in the row-r, dan n_c is the number of samples in the column-c. In addition to using the Chi-Square method, a commonly used feature selection method is Mutual Information feature selection. Mutual Information value calculation can be obtained using Equation (3) [10]:

$$MI\ (X,Y) = \ Entropy(X) + Entropy(Y) - Join\ Entropy(X,Y) \tag{3}$$

where *entropy* (X) is the *entropy* value of feature X (independent), *entropy* (Y) is the *entropy* value of feature Y (dependent), and *join entropy* (X,Y) is the *join entropy* value between feature X and feature Y. The *entropy* value of feature X can be calculated using Equation (4).

$$Entropy\ (X) = -\sum_{x\in X} p(x)\log_2 p(x) \tag{4}$$

The *entropy* value of feature Y can be calculated using equation (5).

$$Entropy\ (Y) = -\sum_{y\in Y} p(y)\log_2 p(y) \tag{5}$$

The *join entropy* value between feature X and feature Y can be calculated using Equation (6).

$$Join\ Entropy\ (X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(x,y) \tag{6}$$

with $p(x)$ is the probability of feature X, $p(y)$ is the probability of feature Y, and $p(x, y)$ is the probability of featured X and feature Y. In the Mutual Information, if the value obtained is greater, the relationship between the two variables is higher [11]. The next stage is to apply the C4.5 algorithm with the aim of forming a decision tree structure.

The C4.5 algorithm functions as a classification algorithm that shows a rule. The C4.5 algorithm has several advantages, namely handling noise data, handling discrete and continuous type variables, handling missing values, and producing decisions from each node by selecting the optimal branch [12]. The decision tree modeling stage of the C4.5 algorithm is described as follows [13]:

a. Preparing the data to be used

b. Calculating the entropy value, the entropy value is a measure of uncertainty in data or a difference in decisions about values in certain variables. The entropy value can be calculated using Equation (7).

$$Entropy\ (S) = -\sum_{j=1}^{m} P_j \log_2 P_j \tag{7}$$

where S is the sum of all samples, m is the number of classes in the dependent variabele, dan P_j is the proportion of data in the -j class in the data.

c. Calculating the *gain* value, the *gain* is a level of influence or a measure of the effectiveness of a variable in the data classification process. To calculate the *gain* value can be calculated using equation (8).

$$Gain\ (S, X) = Entropy(S) - \sum_{j=1}^{k} \frac{|S_j|}{|S|} \times Entropy\ (S_j) \tag{8}$$

where S is the total number of samples, X is the independent variable, S_j is the total number of samples for the jth category-j, dan k is the number of categories in the independent variable X. In the C4.5 Algorithm, the highest gain value will be used to determine which variable will become the node of a decision tree.

d. The next step is to calculate the splitinfo value using Equation (9)

$$SplitInfo\ (S, X_j) = -\sum_{j=1}^{k} \frac{|S_j|}{|S|} \log_2 \frac{|S_j|}{|S|} \tag{9}$$

where S is the total number of samples, X is the j independent variable, S_j is the total number of samples for the category-j, and k is the number of categories in the independent variable X.

e. After calculating the splitinfo value, the next step is to calculate the gain ratio value. The gain ratio value can be calculated using Equation (9).

$$Gain\ Ratio\ (S, X_j) = \frac{Gain\ (S, X)}{SplitInfo\ (S, X_j)} \tag{10}$$

with $Gain\ (S, X)$ is a measure of the effectiveness of the independent variable X and $SplitInfo\ (S, X_j)$ shows the potential information on the independent variable X of the j category. The highest Gain ratio value will be used as the root node in the decision tree, namely node 1.

f. After obtaining node 1 as the root node, the following process is to repeat the second and third processes by selecting the highest gain value to determine the variable that will become the next node until all nodes and branches have classes.

The C4.5 Algorithm decision tree formation process is carried out until all nodes in each decision tree have been formed. The branch formation process will stop if all cases in the node have obtained the same class, no independent variables can be partitioned again, and no cases in the empty branch. If all decision trees have been formed, the next step is to measure model evaluation using a confusion matrix. The model evaluation measurements calculate the accuracy, sensitivity, and specificity values. The form of the confusion matrix table commonly used in model evaluation is shown in Table 1.

Table 1. Confusion Matrix

Classification	**Actual Class**	
	True	**False**
Predicted: True	True Positive (TP)	False Negative (FP)
Predicted: False	False Positive (FN)	True Negative (TN)

Calculation of accuracy, specificity, and sensitivity values can be calculated using the following equation [14]:

a. Accuracy is used to calculate the level of classification accuracy of a resulting model. The accuracy value can be calculated using Equation (10).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (11)$$

Accuracy values can be classified into five groups, namely:

Table 2. Classified Value Accuracy

Value	**Classification**
90,01% - 100%	Very Good
80,01% - 90,00%	Good
70,01% - 80,00%	Fair
60,01% - 70,00%	Bad
≤ 60,00%	Failed

Source: [15]

b. Sensitivity is intended to calculate the ability of a model to find related information. The sensitivity value can be obtained using Equation (12).

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (12)$$

c. Specificity shows how many percent of positive category data is correctly classified. The specificity value can be obtained using Equation (13).

$$Specitivity = \frac{TN}{TN + FP} \times 100\% \quad (13)$$

3. RESULTS AND DISCUSSION

3.1. Descriptive Statistical Analysis

A descriptive statistical analysis technique is a statistical analysis that provides an overview of the data being analyzed. Information regarding descriptive statistical analysis can be found in Table 3. Based on Table 3, it is known that head injury cases mainly occur in male patients; the cause of head injury generally occurs in cases of traffic accidents. This is caused by a lack of vigilance in driving, which can cause accidents. The results of descriptive statistics state that head injury cases occur mostly in patients in their late teens aged 17 years to 25 years, with the highest length of hospitalization being ≤ 9 days. Of the 70 patients with head injuries, there were 32 patients with a mild level of Glasscow Coma Scale. This shows that head injury patients are declared capable of providing eye, motor, and verbal responses with normal SSC conditions in as many as 52 patients, which shows that the prediction of death exit status still tends to be minimal from the overall sample results used.

Table 3. Descriptive Statistical Analysis

No.	Variable	Category	Amount	Percentage
1	Exit Status	1. Live	54	77.14%
		2. Deceased	16	22.86%
2	Age	1. Toddlers	43	61.43%
3	Gender	e	43	61.43%
4	Length of Hospitalization	1. ≤ 9 Day’s	55	78.57%
5	Etiology of Head Injury	1. Traffic Accident	63	90.00%
6	Suprasellar Cistern (SSC)	1. 0	52	74.29%
7	Glasscow Coma Scale (GCS)	1. Lightweight	32	45.71%

3.2. Chi-Square Feature Selection Method

Chi Square feature selection is a filter feature selection method used to measure the distribution value of variables that can affect or not at all affect the value of the dependent variable used. The Chi Square feature selection calculation is performed using Equation (1) presented in Table 4. After calculating the Chi Square value, the next step is to determine the decision making that will be used in the Chi Square test. This decision is determined by comparing χ^2_{table} and χ^2_{count} which has been obtained. The determination of this decision is adjusted to the value of the previously determined significance level. In addition, determining of the degree of freedom value is also one of the references in calculating the results of the decision to be taken at the final stage in determining the table value that will be used later. In the comparison of values χ^2_{table} and χ^2_{count}, if the result $\chi^2_{count} > \chi^2_{table}$ then the hypothesis H_0 is rejected. The hypotheses used are:

H_0 : There is no significant relationship between the independent variable and the dependent variable.

H_1: There is a significant relationship between the independent variable and the dependent variable.

The results of calculating the Chi-Square table value and Chi-Square count for all independent variables are listed in Table 4.

Table 4. Comparison Table $\chi^2_{count} > \chi^2_{table}$

Variable	χ^2_{count}	Degree of Freedom	χ^2_{table}	Decision
Variable X_1 to Y	13.24	$df : (9-1)(2-1) = 8$	15.50	Fail to Reject H_0
Variable X_2 to Y	0.01	$df : (2-1)(2-1) = 1$	3.84	Fail to Reject H_0
Variable X_3 to Y	0.98	$df : (2-1)(2-1) = 1$	3.84	Fail to Reject H_0
Variable X_4 to Y	2.30	$df : (3-1)(2-1) = 2$	5.99	Fail to Reject H_0
Variable X_5 to Y	27.82	$df : (3-1)(2-1) = 2$	5.99	Reject H_0
Variable X_6 to Y	22.62	$df : (3-1)(2-1) = 2$	5.99	Reject H_0

Based on the data in Table 4, four variables have the decision to fail to reject H_0, so that the four variables are declared not to have a significant effect on the dependent variable. In comparison, the two rejected independent variables are declared to significantly impact the dependent variable H_0, which means that both variables significantly influence the dependent variable, namely the Suprasellar Cistern and Glasscow Coma Scale variables.

3.3. Mutual Information Feature Selection Method

Mutual Information feature selection is a computationally efficient feature selection stage using a simple formula, which aims to show the relationship or attachment of each independent variable to the dependent variable. Mutual Information value calculation can be done using Equation (2). The results of the Mutual Information value calculation for each independent variable used are presented below:

Table 5. Result Table Mutual Information

Variable	$\boldsymbol{MI}\ (\boldsymbol{X}, \boldsymbol{Y})$
Suprasselar Cistern	0.28
Glassgow Coma Scale	0.18
Age	0.13
Etiology of Head Injury	0.04
Length of Hospitalization	0.01
Gender	0

Based on the Mutual Information feature selection rule, if the feature selection value obtained is greater, the relationship between the independent variable and the dependent variable is stronger. Based on Table 5, it can be stated that five independent variables have a significant influence, namely the variables of age, etiology of head injury, length of hospitalization, Suprasellar Cistern and Glasscow Coma Scale, while the gender variable is stated to have no significant relationship with the dependent variable because it has a Mutual Information feature selection value of 0.

3.4. C4.5 Algorithm with Feature Selection

The formation of the decision tree in the C4.5 feature selection algorithm shows some differences at the variable selection stage. The C4.5 algorithm with feature selection is analyzed using several variables that are stated to have a relationship with the dependent variable. The following are the stages of forming the C4.5 Algorithm Decision tree with feature selection, namely as follows:

a. Prepare the data to be used, namely head injury patient data with a total sample size of 70 samples.

b. Calculating the total entropy value using Equation (6).

$$Entropy\ (total) = 0.77$$

c. Calculate the gain value using Equation (7). The following is an example of calculating the gain value for the gender variable.

$$Gain\ (Total, JK) = 0.00$$

d. After calculating the gain value, the next step is to calculate the splitinfo value which can be calculated using Equation (8). An example of calculating the splitinfo value of the gender variable is presented as follows:

$$SplitInfo(Total, JK)\ = 0.96$$

e. The next step is to calculate the Gain ratio value. The Gain ratio value can be calculated using Equation (9). An example of calculating the Gain ratio value for the gender variable is given:

$$Gain\ Ratio(Total, JK) = 0.00$$

The calculation of entropy, Gain, splitinfo, and Gain ratio values is also carried out on all independent variables used in the study, namely variables The results of the formation of a decision tree with Mutual Information feature selection are shown in Figure 1.

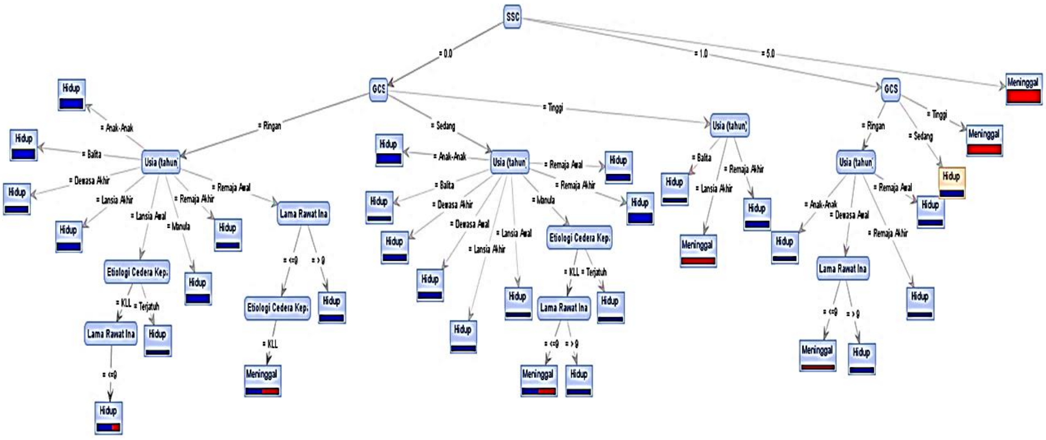

Figure 1. C4.5 Algorithm Decision Tree with Mutual Information feature selection

Based on the decision tree presented in Figure 1, the results of the formation of decisions in the form of rules (rules) are 29 rules consisting of 4 rules for the exit status of death and 25 rules for the exit status of life.

3.5. Model Evaluation with Confusion Matrix

A confusion matrix is an evaluation or testing method used to evaluate the results of a model that has been formed. Confusion matrix as a test will produce accuracy, sensitivity, and specificity values [14]. he results of creating a confusion matrix on the C4.5 Algorithm decision tree model without feature selection are presented in Table 6. The results of forming a confusion matrix on the C4.5 Algorithm decision tree model using the Chi-Square feature selection are presented in Table 7, and the results of forming a confusion matrix on the C4.5 Algorithm decision tree model using the Mutual Information feature selection are presented in Table 8.

Table 6. Confusion Matrix Table Model Without Feature Selection

Classification	Actual Life	Actual Death
Life Prediction	51	5
Prediction of Death	3	11

Table 7. Chi Square Feature Selection Model Confusion Matrix Table

Classification	Actual Life	Actual Death
Life Prediction	52	6
Prediction of Death	2	10

Table 8. Confusion Matrix Table Mutual Information Feature Selection Model

Classification	Actual Life	Actual Death
Life Prediction	52	4
Prediction of Death	2	12

The results of the formation of the confusion matrix in Table 6, Table 7, and Table 8 can be calculated accuracy value using Equation (11), sensitivity value using Equation (12), and specificity value using Equation (13). The results of model evaluation using the confusion matrix on the formation of the three resulting models are presented in Table 9.

Table 9. Result Table *Confusion Matrix*

Model	Accuracy	Sensitivity	Specificity
C4.5 Algorithm Without Feature Selection	88.57%	94.44%	68.75%
C4.5 Algorithm with Chi Square Feature Selection	88.57%	96.29%	62.50%
C4.5 Algorithm with Mutual Information Feature Selection	**91.42%**	96.29%	75.00%

Based on Table 9, it can be stated that of the three models formed, the highest accuracy value is obtained for the C4.5 Algorithm model with Mutual Information feature selection obtained an accuracy of 91.42%, a sensitivity value of 96.29%, and a specificity value of 75%.

4. CONCLUSION

Based on the results and discussions that have been carried out in this study, it can be concluded that the Chi-Square feature selection obtained two independent variables that have a significant effect on the discharge status of head injury patients, namely the Suprasellar Cistern and Glasscow Coma Scale variables while for the Mutual Information feature selection results obtained five independent variables that have a significant effect, namely the age variable, head injury etiology, length of hospitalization, Suprasellar Cistern, and Glasscow Coma Scale. The model evaluation results using the Confusion Matrix show that the highest accuracy value of 91.42% is obtained among the three models formed, resulting from creating the C4.5 Algorithm classification model with Mutual Information feature selection.

ACKNOWLEDGMENTS

The researcher would like to thank all those who have contributed to this research, especially Bilal Wana Satria, S.Ked, who has provided essential data and information to complete the research.

REFERENCES

[1] M. F. B. Gunawan, S. Maliawan, T. G. B. Mahadewa, and I. W. Niryana, “Karakteristik Klinis Cedera Kepala Pada Pediatri Di Rsup Sanglah Denpasar Tahun 2020,” *Jurnal Medika Udayana*, vol. 11, no. 4, pp. 95–100, 2022, doi: 10.24843.MU.2022.V11.i5.P16.

[2] Badan Penelitian dan Pengembangan Kesehatan, “Laporan Provinsi Kalimantan Barat RISKESDAS 2018,” 2019. Accessed: Nov. 07, 2023. [Online]. Available: www.litbang.depkes.go.id

[3] A. Capizzi, J. Woo, and M. Verduzco-Gutierrez, “Traumatic Brain Injury: An Overview of Epidemiology, Pathophysiology, and Medical Management,” *Medical Clinics of North America*, vol. 104, no. 2. W.B. Saunders, pp. 213–238, Mar. 01, 2020. doi: 10.1016/j.mcna.2019.11.001.

[4] S. Tandean, J. Japardi, F. Kollins, and M. L. Loe, “Epidemiology of Traumatic Brain Injury in Neurosurgery Department of Tertiary Referral Hospital at North Sumatera,” 2019. doi: http://dx.doi.org/10.19166/med.v7i5.2471.

[5] R. Raj, J. Siironen, M. B. Skrifvars, J. Hernesniemi, and R. Kivisaari, “Predicting outcome in traumatic brain injury: Development of a novel computerized tomography classification system (Helsinki Computerized Tomography Score),” *Neurosurgery*, vol. 75, no. 6, pp. 632–646, 2014, doi: 10.1227/NEU.0000000000000533.

[6] C. Zai, “Implementasi Data Mining Sebagai Pengolahan Data,” *Portaldata.org*, vol. 2, no. 3, pp. 1–12, 2022.

[7] M. Adriansa, L. Yulianti, L. Elfianty, U. Dehasen Bengkulu, and J. Meranti Raya, “Analisis Kepuasan Pelanggan Menggunakan Algoritma C4.5,” *Jurnal Teknik Informatika Unika St. Thomas (JTIUST)*, vol. 07, no. 01, pp. 115–121, 2022.

[8] I. M. B. Adnyana, “Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa,” *Jurnal Sistem dan Informatika*, vol. 13, no. 2, pp. 72–76, 2019.

[9] L. A. C. Ahakonye, C. I. Nwakanma, J. M. Lee, and D. S. Kim, “SCADA intrusion detection scheme exploiting the fusion of modified decision tree and Chi-square feature selection,” *Internet of Things (Netherlands)*, vol. 21, pp. 1–17, Apr. 2023, doi: 10.1016/j.iot.2022.100676.

[10] F. Souza, C. Premebida, and R. Araújo, “High-order conditional mutual information maximization for dealing with high-order dependencies in feature selection,” *Pattern Recognit*, vol. 131, Nov. 2022, doi: 10.1016/j.patcog.2022.108895.

[11] K. Wang, W. Mao, W. Feng, and H. Wang, “Research on spam filtering technology based on new mutual information feature selection algorithm,” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1742-6596/1673/1/012028.

[12] S. Febriani and H. Sulistiani, “Analisis Data Hasil Diagnosa Untuk Klasifikasi Gangguan Kepribadian Menggunakan Algoritma C4.5,” *Jurnal Teknologi dan Sistem Informasi (JTSI)*, vol. 2, no. 4, pp. 89–95, 2021, doi: http://jim.teknokrat.ac.id/index.php/JTSI.

[13] P. B. N. Setio, D. R. S. Saputro, and B. Winarno, “Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C4.5,” *PRISMA, Prosding Seminar Nasional Matematika*, vol. 3, pp. 64–71, 2020, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[14] Hozairi, Anwari, and S. Alim, “Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes,” *Jurnal Ilmiah NERO*, vol. 6, no. 2, pp. 133–144, 2021.

[15] Y. Crismayella, N. Satyahadewi, and H. Perdana, “Algoritma Adaboost pada Metode Decision Tree untuk Klasifikasi Kelulusan Mahasiswa,” *Jambura Journal of Mathematics*, vol. 5, no. 2, pp. 278–288, Aug. 2023, doi: 10.34312/jjom.v5i2.18790.