JPPIPA 9(7) (2023)

Jurnal Penelitian Pendidikan IPA

Journal of Research in Science Education

http://jppipa.unram.ac.id/index.php/jppipa/index

The Node Selection Method for Split Attribute in C4.5 Algorithm Using the Coefficient of Determination Values for Multivariate Data Set

Muhsi[1,2*], Suprapto[2], Rofiuddin[3]

[1]Department of Information System, Universitas Islam Madura, Pamekasan, Indonesia
[2]Department of Electronics and Informatics Education, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[3]Department of Informatics Engineering, Universitas Islam Madura, Pamekasan, Indonesia

Received: May 26, 2023
Revised: July 19, 2023
Accepted: July 25, 2023
Published: July 31, 2023

Corresponding Author:
Muhsi
muhsi@uim.ac.id

DOI: 10.29303/jppipa.v9i7.4031

Abstract: The split attribute in the decision tree algorithm, especially C4.5, has an important influence in producing a decision tree performance that has high predictive performance. This study aims to perform an attribute split in the C4.5 algorithm using the value of the termination coefficient (R^2/R Square) which is combined with the aim of increasing the performance of the model performance produced by the C4.5 algorithm itself. The data used in this research are public datasets and private datasets. This study combines the C4.5 algorithm developed by Quinlan. The results in this study indicate that the use of the R^2 value in the C4.5 algorithm has good performance in terms of accuracy and recall because three of the four datasets used have a higher value than the C4.5 algorithm without R^2. Whereas in the aspect of precision, it has quite good performance because only two datasets have a higher value than the performance results of the algorithm without R^2.

Keywords: Algorithm; Attribute; Multivariate

Introduction

The decision tree algorithm that is commonly used to predict has its own challenges related to its accuracy and scalability (Putra et al., 2023). Decision tree is a technique that assists in making decisions that resemble a tree or hierarchical shape (Ishak et al., 2019). The resulting model is later in the form of a recursive procedure, in which a set of a statistical unit is progressively divided into several groups based on division rules that aim to maximize the homogeneity or purity of the size of the response variable in each group obtained (Taylor et al., 2023). Several algorithms that can be used in decision trees are (Mienye et al., 2019) CHAID (Chi-squared Automatic Interaction Detection), CART (Classification and Regression Tree), ID3 (Iterative Dichotomiser 3), and C4.5 which is a development of ID3 (Idriss & Lawan, 2019). Then it was developed by several researchers into Credal DT and Credal C4.5 (Mantas et al., 2016). This algorithm can be implemented in data mining techniques in the form of classification of very large amounts of data with the aim of extracting knowledge through understanding the characteristics in the data (Muhsi, 2021).

The C4.5 algorithm is a decision tree algorithm that is widely used in data mining classification research because it is easy to interpret (Muttaqien et al., 2021). The weakness that is often found in the C4.5 algorithm is in the overfitting aspect so that it is good from a training point of view but is weak when implemented on unseen data. In addition, the classification performance in the C4.5 algorithm still experiences misclassification costs which usually occur due to poor attribute split factors (Wang & Gao, 2021). This study aims to perform an attribute split in the C4.5 algorithm using the value of the termination coefficient (R^2/R Square) which is combined with the aim of increasing the performance of

How to Cite:
Muhsi, M., Suprapto, S., & Rofiuddin, R. (2023). The Node Selection Method for Split Attribute in C4.5 Algorithm Using the Coefficient of Determination Values for Multivariate Data Set. *Jurnal Penelitian Pendidikan IPA*, *9*(7), 5574–5583. https://doi.org/10.29303/jppipa.v9i7.4031

the model performance produced by the C4.5 algorithm itself.

C4.5 algorithm is an algorithm that used to construct a decision tree Sulistiani & Aldino (2020), which is a method of classification and prediction very powerful and famous. Tree method decision turns a very big fact into decision trees that represent rules can be easily understood in language experience. Algorithm is a sequence of logical completion steps arranged systematically Consideration in choosing an algorithm is an algorithm has true value (Theofani & Sediyono, 2022), has efficiency which means the algorithm is used because can provide correct values (Loftus et al., 2022). There has been a lot of research on the development of decision tree algorithms related to split attributes.

Delgado-Bonal & Marshak (2019), uses the gain method which measures the probability with the value of the bits of the base 2 algorithm as a minus, then all existing classes are summed with their frequency. In addition, redeveloped the Info Gain Ratio method to overcome the weaknesses of the previous gain method (Albulayhi et al., 2022; Mao & Zhang, 2021). Furthermore, the development of the C4.5 algorithm has been carried out in terms of split attributes such as the imprecise info-gain ratio method Madadipouya (2017), (Credal-C4.5) using the Imprecise Probability Theory, Bosting gain ratio (C5.0), bagging techniques and average gain. There are also studies that combain the RelieF algorithm with the C4.5 algorithm. Abellan uses Imprecise Info Gain (IIG) in splitting attributes where the data set used is calculated as an imprecise probability and uncertainty measure. Threshold Pruning and Cost Complexity Pruning methods for splitting attributes in the C4.5 algorithm.

Meanwhile, Hart (2017), proposed a method for smoothing called m-estimation where the use of this method is intended to obtain the best estimate of probability. In this study, the value of the coefficient of determination or R-square is used in the form of the value of the measurement results of the regression formula. This value is used to see the ability of the regression model to explain how much influence the independent variables have on the dependent variable (Theofani & Sediyono, 2022). The coefficient of determination is also used to determine the extent to which the contribution value of the independent variables in the regression model has the ability to explain the variation of the dependent variable (Nawawi, 2020). In calculating regression statistics, a coefficient of determination value will be found by looking at the R-square value in the form of a range of 0 – 1. On this basis, the R^2 value for each data attribute will be processed using the C4.5 algorithm combined with the gain ratio value for selection split determination data to be used as a root or leaf node with the aim of improving the performance of the resulting model.

Method

The data used in this research are public datasets from UCI. The information and characteristics of the dataset used are as shown in Table 1.

Table 1. Dataset information

Information	Data sets			
	Breast Cancer Coimbra	Occupancy Detection	Heart Disease	Earlystage diabetes risk
Data Set Characteristics:	Multivariate	Multivariate, TimeSeries	Multivariate	Multivariate
Attribute Characteristics:	Integer	Real	Categorical, Integer, Real	Integer
Associated Tasks:	Classification	Classification	Classification	Classi fication
Number of Instances:	116	11852	303	520
Number of Attributes:	10	7	14	17
Missing Values?	N/A	N/A	N/A	Yes
Area:	Life	Computer	Life	Computer
Date	06/03/	29/02/	01/07/	07/12/
Donated	2018	2016	1988	2020
Number of Web Hits:	131792	177228	2201311	114057

This study combines the C4.5 algorithm developed by Quinlan with the coefficient of determination (R^2) of each attribute that is correlated with the class label. The use of the R^2 value is intended for the selection of split attributes to be used as a root or leaf node so that the decision tree obtained has a better performance value after being applied to the data set. Figure 1 below illustrates the process of using the R^2 value in the c4.5 algorithm. In Figure 1 it can be shown that the development of the proposed C4.5 algorithm is as follows, 1) Preprocessing data by separating training and testing data. Determine the x and y variables. 2) Calculate the entropy of the dataset with formula 1.

$$entropy(S) = \sum_{i=1}^{n} -pi \times \log_2 pi \quad (1)$$

Where S is dataset, n is count of S partitions and pi is the proportion of S_i to S.

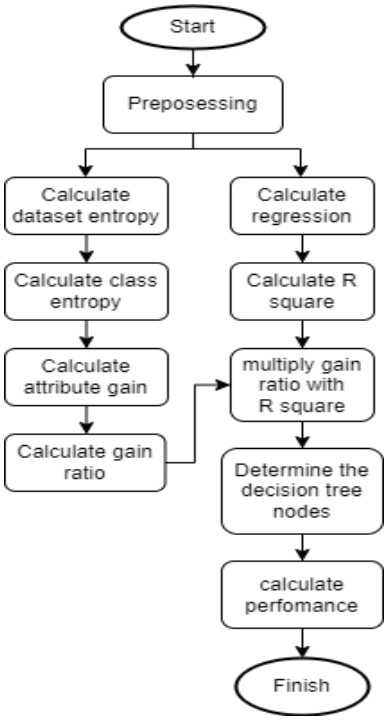


Figure 1. development stages

1. Calculate the entropy of the class with formula:
$$entropy(S_i) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times entropy(S) \quad (2)$$
Where S is dataset, n is count of S partitions, $|S_i|$ is the count of cases on i partition and $|S|$ is the count of S cases.
2. Calculate gain with formula:
$$Gain(S, A) = entropy(S) - entropy(S_i) \quad (3)$$
where S is dataset, A attribute, S_i count of sample for i attribute.
3. Calculate split info with formula:
$$splitinfo(S, A) = \sum_{i=1}^{n} \frac{S_i}{S} log_2 \frac{S_i}{S} \quad (4)$$
where where S is dataset, A attribute, S_i count of sample for i attribute.
4. Calculate gain rasio with formula:
$$gainratio(A) = \frac{gain(S,A)}{splitinfo(S,A)}, \quad (5)$$
where where S is dataset, and A attribute.
5. Calculate R^2 with formula:
$$R^2 = \frac{((n)(\sum xy) - (\sum x)(\sum y))^2}{(n(\sum x^2) - (\sum x)^2)\,(n(\sum y^2) - (\sum y)^2)} \quad (6)$$
6. Multiply gain ratio (6) with R^2.
7. Determine the decision tree node by selecting the highest value.
8. Calculate performance.

Thus, for each stage of determining the next decision leaf, it is carried out using processes 1 to 9. The process stops until the existing attributes can no longer be derived to produce more detailed decision leaves.

The resulting decision tree model is then calculated for its performance using the confusion matrix to see the values for accuracy, precision, and recall because it only has two classes for each data. In the use of the confusion matrix there are terms that will be used to calculate the level of performance of the model that has been produced. True positive (TP) is used for data that is predicted to be the same as reality in a certain class while True negative (TN) is the amount of data that is predicted to be the same as reality but for certain other classes. False positive (FP) is the amount of data that is predicted not to be the same as reality for one class while false negative (FN) is the amount of data that is predicted to be different from reality for another class. As shown table 2.

Table 2. Confusion Matrix

	Actually Positive	Actually Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

From the values in the Confusion Matrix, performance can be calculated in the form of accuracy, precision, and recall. The formula for each performance is as follows:

$$akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$presisi = \frac{TP}{TP+FP} \quad (8)$$

$$recall = \frac{TP}{TP+FN} \quad (9)$$

Result and Discussion

Preparation Datasets

Each dataset downloaded from the ICS UCI page is prepared by selecting the attributes to be used and fixing the missing data values. After the data is cleaned, it is then divided into two in the form of training data and testing data with a proportion of 80% and 20%.

Proposed Method

Based on the previously prepared data, the calculation of the C4.5 algorithm and the termination coefficient is carried out to get the nodes with the following results:

Calculate the entropy value of each dataset

This process is carried out to see the diversity of the dataset used. The results are as shown in table 3.

Table 3. Entropy values for each dataset

Datasets	Qty		Class	Entropy
Occupancy Detection	6073	0 5167	1 906	0.608
Breast Cancer Coimbra	93	1 40	2 53	0.986
Heart Disease	242	0 110	1 132	0.994
Early-stage diabetes risk	416	1 160	2 256	0.961

Calculate the entropy value of each class

At this stage it is intended to see the diversity of each data class in the dataset. However, before doing a split mapping for each attribute then calculating the entropy of each class. The results are as shown in table 4.

Table 4. The highest gain ratio value for each dataset

Dataset	Attribute	Gain Ratio
Occupancy Detection	Light	0.39
Breast Cancer Coimbra	Age	0.15
heart disease	Slope	0.50
Early-stage diabetes Risk	Polyuria	0.35

Calculate the gain ratio

Gain ratio is a comparison of the gain value with the splitinfo value of each attribute. The results of calculating the highest gain ratio for each attribute in the dataset are shown in table 5. Based on the results of calculating the attribute gain ratio with the C4.5 algorithm, the root node for each dataset is obtained, namely for the Occupancy Detection dataset is the light attribute, the Breast Cancer Coimbra dataset is the age attribute, the heart disease dataset is the slope attribute, and the Early-stage diabetes risk dataset is the polyuria attribute.

Table 5. Entropy values for each dataset attribute

Datasets	Attribute	Entropy
Occupancy Detection	Temp	0.74
	Humidity	0.59
	Light	0.99
	CO2	0.60
	Humidity	0.60
Breast Cancer Coimbra	Age	0.98
	BMI	0.98
	Glucose	0.98
	Insulin	0.98
	Homa	0.98
	Leptin	0.98
	Adipo Nectin	0.98
	Resistin	0.98
	MCP.1	0.98
Heart Disease	Age	0.99
	Gender	0.99
	Chest pain	0.99
	Resting blood	0.99
	Cholestoral	0.99
	Fasting blood Sugar	0.99
	Resting Electrocar	0.99
	Maximum Heart Rate	0.99
	Exercise Induced Angina	0.88
	Depression Induced	0.99
	The Slope	0.99
	Number of Major	0.99
	Thall	0.99
Early-stage diabetes risk	Age	0.96
	Gender	0.98
	Polyuria	0.87
	Polydipsia	0.92
	Sudden	0.98
	Weakness	0.99
	Polyphagia	0.99
	Genital thrush	0.97

Datasets	Attribute	Entropy
	visual blurring	0.99
	Itching	0.96
	Irritability	0.99
	delayed healing	0.97
	partial paresis	0.98
	muscle stiness	0.98
	Alopecia	0.98
	Obesity	0.97

4) Calculates the value of R^2

Testing the coefficient of determination was carried out with the intention of measuring the ability of the model to explain how the effect of the independent variables jointly (simultaneously) affects the dependent variable which can be indicated by the value of adjusted R – Squared (Perwitasari, 2022). The coefficient of determination shows the extent to which the contribution of the independent variables in the regression model can explain the variation of the dependent variable.

The coefficient of determination can be seen through the value of R-square (R^2) in the Model Summary table. According to Jenkins & Quintana-Ascencio (2020), a small coefficient of determination means that the ability of the independent variables to explain the dependent variable is very limited. Conversely, if the value is close to one and away from 0 (zero), it means that the independent variables have the ability to give all information needed to predict the dependent variable (Andrade, 2021).

Regular attributes in each dataset are used as independent variables (x) and class attributes as dependent variables (y) (Demisse et al., 2017). The results of calculations with the R^2 formula obtained the value of R^2 for each dataset attribute as shown in table 6.

Table 6. R^2 values for each dataset attribute

Datasets	Attribute	R^2
Occupancy Detection	Light	0.81
	Temperature	0.25
	HumidityRatio	0.06
	Humidity	0.01
	CO2	0.00
Breast Cancer Coimbra	Glucose	0.14
	Homa	0.08
	Insuline	0.07
	Resistin	0.04
	MCP	0.01
	BMI	0.01
	Adiponectin	0.00
	Age	0.00
	Leptin	0.00
Heart Disease	Age	0.06
	Gender	0.07
	Chest pain	0.18
	Resting Blood	0.03
	Cholestoral	0.00
	Fasting Blood Sugar	0.00
	Resting Electrocar	0.00
	Maximum Heart Rate	0.21
	Exercise Induced Angina	0.18
	Depression Induced	0.15
	The Slope	0.11
	Number Of Major	0.19
	Thall	0.09
Early-stage diabetes risk	Age	0,00
	Gender	0.25
	Polyuria	0.43
	Polydipsia	0.40
	Sudden	0.18
	weakness	0.05

Polyphagia	0.08
Genital thrush	0.01
Visual Blurring	0.06
Itching	0.00
Irritability	0.09
Delayed Healing	0.00
Partial Paresis	017
Muscle Stiness	0.00
Alopecia	0.07
Obesity	0,0053

Then the gain ratio value for each attribute is multiplied by the R^2 value to determine the split for each attribute to be used as a root node or leaf node. The highest multiplication result value is selected as a split in the attributes and nodes as shown in table 7.

Table 7. Split selected based on the value of the gain ratio and R^2

Dataset	Attribute	Split	Gain Ratio & R^2
Breast Cancer Coimbra	Glucose	118.50	0.03
Occupancy Detection	Light	364.50	0.71
heart disease	Maximum heart rate achieved	110	0.04
Early-stage diabetes risk	Polyuria	0.50	0.15

Based on the multiplication of the gain ratio with R^2, the split attribute is obtained as the root node dataset, namely for the Occupancy Detection dataset is the light attribute, the Breast Cancer Coimbra dataset is the glucose attribute, the heart disease dataset is the maximum heart rate achieved attribute and the Early-stage diabetes risk dataset is the attribute polyuria. Processes (a) through (d) are repeated until a gain value = 0 is obtained from calculating all the remaining attributes.

Decision Tree Results

The decision tree is one of the most popular classification methods because it can be easily interpreted by humans (Lamrini, 2021). A decision tree is a structure that can be used to divide a large data set into smaller record sets apply a set of decision rules (Lee et al., 2022) .

The results of calculations with the C4.5 algorithm which uses the value of the termination coefficient will get a decision tree. The decision tree can be used as a rule of knowledge model to be used as a prediction. The decision tree of each dataset (Riansyah et al., 2023) in chart form is shown in Figure 2.

Performance Results

The performance of the rule model resulting from combining the termination coefficient values in the C4.5 algorithm compared to the C4.5 algorithm without using the termination coefficient value for the Occupancy Detection dataset is as shown in table 8 and Figure 3.

Table 8. Comparison of method performance for the Occupancy Detection dataset

		Performance (%)	
Method	Accuracy	Precision	Recall
Algo. C4.5 & Koef.	98.18	97.63	88.10
Algo. C4.5	94.70	77.22	83.82

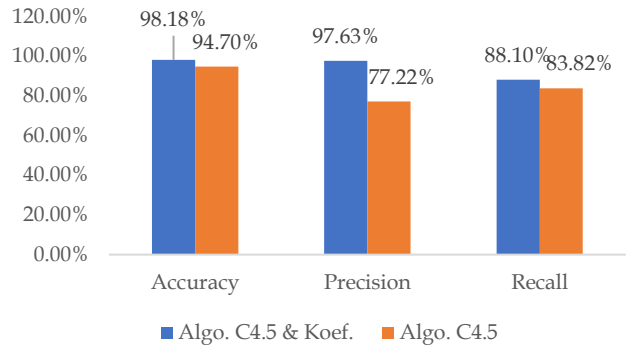


Figure 3. Graph of Method Performance Comparison for Occupancy Detection Dataset

The performance of implementing the two attribute split methods in the decision tree algorithm for Occupancy Detection data in Figure 2 shows that the C4.5 algorithm using the termination coefficient value (R^2) has higher performance compared to the C4.5 algorithm without R^2 both from the aspect of accuracy, precision, and recall. Comparison of the performance percentage values of the two methods, namely accuracy of 97.63% and 94.70% with a positive difference of 2.93%, precision of 97.63% and 77.22% with a positive difference of 20.41% and recall of 88.10% and 83.82% with a positive difference of 4.28%. The performance for the Breast Cancer Coimbra dataset is shown in Table 9 and Figure 4.

Table 9. Comparison of method performance for the Breast Cancer Coimbra dataset

Method	Performance (%)		
	Accuracy	Precision	Recall
Algo. C4.5 & Koef.	78.26	75.00	81.82
Algo. C4.5	69.57	75.00	54.55

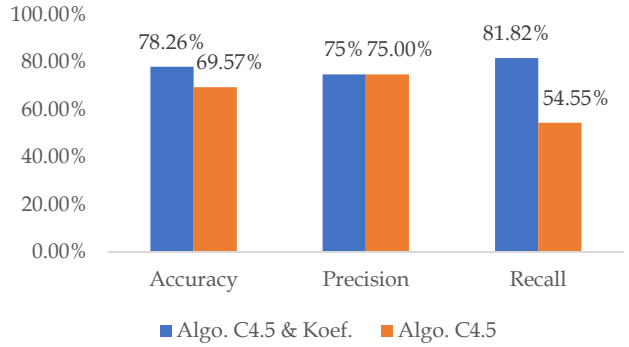


Figure 4. Graph of Method Performance Comparison for Breast Cancer Coimbra Dataset

In Figure 3, the performance of applying the two attribute split methods in the decision tree algorithm for Breast Cancer Coimbra data in algorithm C4.5 using the termination coefficient value (R^2) has higher performance compared to algorithm C4.5 without R^2 both in terms of accuracy and recall. Meanwhile, accuracy performance has the same value, namely 75%. Comparison of the percentage of precision performance values is 78.26% and 69.57% with a positive difference of 8.69% and recall performance is 81.82% and 54.55% with a positive difference of 27.27%. The performance for the heart disease dataset is shown in Table 10 and Figure 5.

Table 10. Comparison of method performance for the heart disease dataset

Method	Performance (%)		
	Accuracy	Precision	Recall
Algo. C4.5 & Koef.	77.05	78.79	78.79
Algo. C4.5	75.41	78.26	64.29

The application of the two attribute split methods in the decision tree algorithm for heart disease data in the C4.5 algorithm using the termination coefficient value (R^2) has a higher performance Compared to the C4.5 algorithm without R^2 both in terms of accuracy, precision and recall as shown in the figure 5. Comparison of the percentage value of each performance is for accuracy of 77.05% and 75.41% with a positive difference of 1.64%, precision of 78.79% and 78.26% with a positive difference of 0.53% and recall of 78.79% and 64.29% with a positive difference of 14.5%. The performance for the Early-stage diabetes risk dataset as shown in Table 11 and Figure 6.

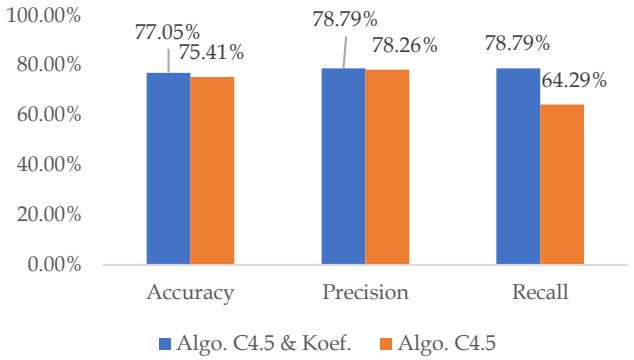


Figure 5. Graph of Method Performance Comparison for Heart Disease Dataset

Table 11. Comparison of method performance for the Early-stage diabetes risk dataset

Method	Performance (%)		
	Accuracy	Precision	Recall
Algo. C4.5 & Koef.	93.27	98.31	90.63
Algo. C4.5	95.19	98.36	93.75

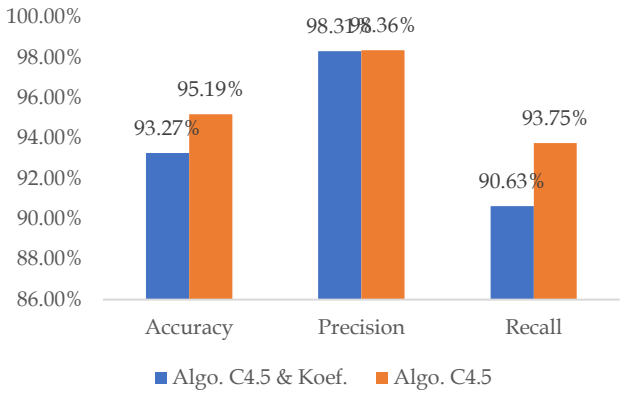


Figure 6. Graph of Method Performance Comparison for Early-stage Diabetes Risk Dataset

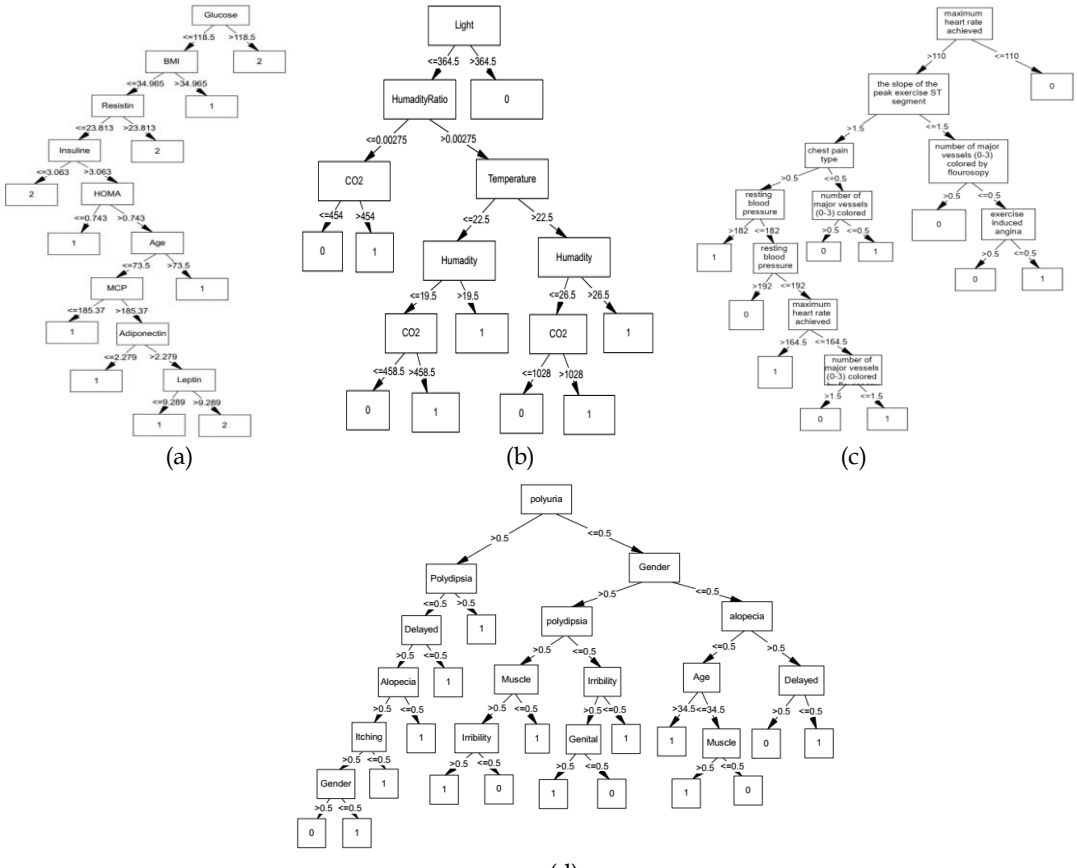
Whereas the application of the two attribute split methods in the decision tree algorithm to the Early-stage Diabetes Risk Dataset shows that the C4.5 algorithm using the termination coefficient value (R^2) has lower performance compared to the C4.5 algorithm without R^2 both in terms of accuracy, precision and recall as shown in Figure 6. Comparison of the percentage value of each performance is for accuracy of 93.27% and 95.19% with a negative difference of -1.92%, precision of 98.31% and 98.36% with a negative difference of -0.05% and recall of 90.63% and 93.75% with a negative difference of -3.12%.

Recapitulation of the performance comparison of the implementation of the C4.5 algorithm that uses the termination coefficient (R^2) with the C4.5 algorithm without R^2 on public datasets is as shown in table 12.

Table 12. Recapitulation of performance comparison

Datasets	Method	Accuracy (%)	Precision Performance (%)	Recall (%)
Breast Cancer Coimbra	C4.5	69.57	75.00	54.55
	C4.5 & R^2	78.26	75.00	81.82
Occupancy Detection	C4.5	94.70	77.22	83.82
	C4.5 & R^2	98.18	97.63	88.10
Heart Disease	C4.5	75.41	78.26	64.29
	C4.5 & R^2	77.05	78.79	78.79
Early-stage Diabetes Risk	C4.5	95.19	98.36	93.75
	C4.5 & R^2	93.27	98.31	90.63

In Table 12 the accuracy performance of the C4.5 algorithm which uses a termination coefficient value (R^2) is higher than the C4.5 algorithm without an R^2 value in trials on the three public datasets used. This shows that the knowledge model rule obtained from the C4.5 algorithm which uses the R^2 value can predict well for each class of the entire data. While the precision performance is higher in the C4.5 algorithm which uses the R^2 value compared to the C4.5 algorithm without the R^2 value in the two public datasets used and one dataset has the same precision value. This shows that the knowledge model rule obtained from the C4.5 algorithm that uses the R^2 value can predict quite well the class that occurs compared to the predicted results. For recall performance, it shows that the C4.5 algorithm that uses a higher R^2 value compared to the C4.5 algorithm without an R^2 value in trials on the three public datasets used. This shows that the knowledge model rule obtained from the C4.5 algorithm that uses the R^2 value can predict a certain class well compared to the actual reality (Kerckhoffs et al., 2019).



(a) (b) (c) (d)

Figure 2. Decision tree for each dataset (a) Breast Cancer Coimbra, (b) Occupancy Detection, (c) heart disease, (d) Early-stage diabetes risk

Conclusion

The technique of using the termination coefficient value (R^2) as a method in determining the split attribute in algorithm C4.5 is done by multiplying it with the gain ratio value that has been obtained previously. Furthermore, the highest value of the multiplication result is used as the basis for determining a root or leaf node of a decision tree. The implementation of the split attribute determination method in the C4.5 algorithm using the R^2 value for 4 public datasets shows that the performance is good in three datasets, namely Breast Cancer Coimbra, Occupancy Detection, and heart

disease both in accuracy, precision, and recall. Comparison of the performance percentage values of the two methods for the Occupancy Detection Dataset is an accuracy of 97.63% and 94.70% with a positive difference of 2.93%, precision of 97.63% and 77.22% with a positive difference of 20.41% and a recall of 88.10% and 83.82% with a positive difference of 4.28%. While the performance of the Coimbra Breast Cancer dataset has the same accuracy value, namely 75%, precision is 78.26% and 69.57% with a positive difference of 8.69% and recall performance is 81.82% and 54.55% with a positive difference of 27.27%. Comparison of performance on the heart disease dataset for accuracy is 77.05% and 75.41% with a positive difference of 1.64%, precision is 78.79% and 78.26% with a positive difference of 0.53%, and recall is 78.79% and 64.29% with a positive difference of 14.5%. Furthermore, the performance comparison for the Early-stage diabetes risk dataset for accuracy is 93.27% and 95.19% with a negative difference of -1.92%, 98.31% and 98.36% precision with a negative difference of -0.05%, and 90.63% and 93.75% recall with a negative difference of -3.12 %.

Acknowledgments
Thanks to all parties who have supported the implementation of this research especially Universitas Negeri Yogyakarta. I hope this research can be useful.

Author Contributions
Conceptualization: Muhsi, data curation: Suprapto, funding acquisition: Rofiuddin, methodology: Suprapto, visualization: Muhsi, writing-original draft: Muhsi, writing–review & editing: Rofiuddin.

Funding
This research was independently funded by researchers.

Conflicts of Interest
No Conflicts of interest.

References

Albulayhi, K., Abu Al-Haija, Q., Alsuhibany, S. A., Jillepalli, A. A., Ashrafuzzaman, M., & Sheldon, F. T. (2022). IoT Intrusion Detection Using Machine Learning with a Novel High Performing Feature Selection Method. *Applied Sciences, 12*(10), 5015. https://doi.org/10.3390/app12105015

Andrade, C. (2021). A Student's Guide to the Classification and Operationalization of Variables in the Conceptualization and Design of a Clinical Study: Part 1. *Indian Journal of Psychological Medicine, 43*(2), 177–179. https://doi.org/10.1177/0253717621994334

Demisse, G. B., Tadesse, T., & Bayissa, Y. (2017). Data Mining Attribute Selection Approach for Drought Modelling: A Case Study for Greater Horn of Africa. *International Journal of Data Mining & Knowledge Management Process, 7*(4), 1–16. https://doi.org/10.5121/ijdkp.2017.7401

Delgado-Bonal, A., & Marshak, A. (2019). Approximate Entropy and Sample Entropy: A Comprehensive Tutorial. *Entropy, 21*(6), 541. https://doi.org/10.3390/e21060541

Hart, J. D. (2017). Use of BayesSim and Smoothing to Enhance Simulation Studies. *Open Journal of Statistics, 7*(1), 153–172. https://doi.org/10.4236/ojs.2017.71012

Idriss, S., & Lawan, A. (2019). An Improved C4.5 Model Classification Algorithm Based on Taylor's Series. *Jordanian Journal of Computers and Information Technology, 5*(1). https://doi.org/10.5455/jjcit.71-1546551963

Ishak, A., Asfriyati, & Akmaliah, V. (2019). Analytical Hierarchy Process and PROMETHEE as Decision Making Tool: A Review. *IOP Conference Series: Materials Science and Engineering, 505*(1), 012085. https://doi.org/10.1088/1757-899X/505/1/012085

Jenkins, D. G., & Quintana-Ascencio, P. F. (2020). A solution to minimum sample size for regressions. *PLOS ONE, 15*(2), e0229345. https://doi.org/10.1371/journal.pone.0229345

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., & Vermeulen, R. C. H. (2019). Performance of Prediction Algorithms for Modeling Outdoor Air Pollution Spatial Surfaces. *Environmental Science & Technology, 53*(3), 1413–1421. https://doi.org/10.1021/acs.est.8b06038

Lamrini, B. (2021). *Contribution to Decision Tree Induction with Python: A Review*. Data Mining—Methods, Applications and Systems. IntechOpen. https://doi.org/10.5772/intechopen.92438

Lee, S., Lee, C., Mun, K. G., & Kim, D. (2022). Decision Tree Algorithm Considering Distances Between Classes. *IEEE Access, 10*, 69750–69756. https://doi.org/10.1109/ACCESS.2022.3187172

Loftus, T. J., Tighe, P. J., Ozrazgat-Baslanti, T., Davis, J. P., Ruppert, M. M., Ren, Y., Shickel, B., Kamaleswaran, R., Hogan, W. R., Moorman, J. R., Upchurch, G. R., Rashidi, P., & Bihorac, A. (2022). Ideal algorithms in healthcare: Explainable, dynamic, precise, autonomous, fair, and reproducible. *PLOS Digital Health, 1*(1), e0000006. https://doi.org/10.1371/journal.pdig.0000006

Madadipouya, K. (2017). A Survey on Data Mining Algorithms and Techniques in Medicine. *JOIV: International Journal on Informatics Visualization, 1*(3), 61. https://doi.org/10.30630/joiv.1.3.25

Mantas, C. J., Abellán, J., & Castellano, J. G. (2016). Analysis of Credal-C4.5 for classification in noisy

domains. *Expert Systems with Applications, 61*, 314–326. https://doi.org/10.1016/j.eswa.2016.05.035

Mao, L., & Zhang, W. (2021). Analysis of entrepreneurship education in colleges and based on improved decision tree algorithm and fuzzy mathematics. *Journal of Intelligent & Fuzzy Systems, 40*(2), 2095–2107. https://doi.org/10.3233/JIFS-189210

Mienye, I. D., Sun, Y., & Wang, Z. (2019). Prediction performance of improved decision tree-based algorithms: A review. *Procedia Manufacturing, 35*, 698–703. https://doi.org/10.1016/j.promfg.2019.06.011

Muhsi, (2021). Model dan Analisa Faktor Eksternal Aktifitas Siswa Kelas X TKJ SMKN 1 Pakong Pamekasan Menggunakan Algoritma Decision Tree. *Jurnal Aplikasi Teknologi Informasi Dan Manajemen (JATIM), 2*(3), 94–106. https://doi.org/10.31102/jatim.v2i2.1239

Muttaqien, R., Pradana, M. G., & Pramuntadi, A. (2021). Implementation of Data Mining Using C4.5 Algorithm for Predicting Customer Loyalty of PT. Pegadaian (Persero) Pati Area Office. *International Journal of Computer and Information System (IJCIS), 2*(3), 64–68. https://doi.org/10.29040/ijcis.v2i3.36

Nawawi, M. (2020). Influence On Service Quality, Product Quality, Product Design, Price and Trust To Xl Axiata Customer Loyalty On Students Of Pgri Karang Sari Belitang Iii Oku Timur Vocational High School. *International Journal of Economics, Business and Accounting Research (IJEBAR), 4*(3). https://doi.org/10.29040/ijebar.v4i03.1251

Perwitasari, A. W. (2022). The Effect of Perceived Usefulness and Perceived Easiness towards Behavioral Intention to Use Fintech by Indonesian MSMEs. *The Winners, 23*(1), 1–9. https://doi.org/10.21512/tw.v23i1.7078

Putra, P. H., Azanuddin, A., Purba, B., & Dalimunthe, Y. A. (2023). Random forest and decision tree algorithms for car price prediction. *Jurnal Matematika Dan Ilmu Pengetahuan Alam LLDikti Wilayah 1 (JUMPA), 3*(2), 81–89. https://doi.org/10.54076/jumpa.v3i2.305

Riansyah, M., Suwilo, S., & Zarlis, M. (2023). Improved Accuracy in Data Mining Decision Tree Classification Using Adaptive Boosting (Adaboost). *SinkrOn, 8*(2), 617–622. https://doi.org/10.33395/sinkron.v8i2.12055

Sulistiani, H., & Aldino, A. A. (2020). Decision Tree C4.5 Algorithm for Tuition Aid Grant Program Classification (Case Study: Department of Information System, Universitas Teknokrat Indonesia). *Jurnal Ilmiah Edutic: Pendidikan dan Informatika, 7*(1), 40-50. https://doi.org/10.21107/edutic.v7i1.8849

Taylor, C. J., Pomberger, A., Felton, K. C., Grainger, R., Barecka, M., Chamberlain, T. W., Bourne, R. A., Johnson, C. N., & Lapkin, A. A. (2023). A Brief Introduction to Chemical Reaction Optimization. *Chemical Reviews, 123*(6), 3089–3126. https://doi.org/10.1021/acs.chemrev.2c00798

Theofani, G., & Sediyono, E. (2022). Multiple Linear Regression Analysis on Factors that Influence Employees Work Motivation. *SinkrOn, 7*(3), 791–798. https://doi.org/10.33395/sinkron.v7i3.11453

Wang, H.-B., & Gao, Y.-J. (2021). Research on C4.5 algorithm improvement strategy based on MapReduce. *Procedia Computer Science, 183*, 160–165. https://doi.org/10.1016/j.procs.2021.02.045