Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

C4.5 Forward Selection Based Algorithm for Class Level Classification of Nurul Jadid Islamic Boarding School Students

Muhammad Isomul Irfan
Universitas Dian Nuswntoro, Indonesia
Ishom.irfan99@gmail.com

Submitted :Feb 16, 2024 | **Accepted** : Mar 1, 2024 | **Published** : Apr 1, 2024

Abstract: Islamic Boarding School is an Islamic educational institution that plays a central role in the development of education in Indonesia. Although originally established for Islamic religious education, Islamic Boarding School has evolved into an educational institution that contributes to both scholarly and community service aspects. According to the regulations set by the Ministry of Religious Affairs of the Republic of Indonesia under Number 31 of 2020, Islamic Boarding School is a community-based institution that upholds the teachings of *Islam rahmatan lil'alamin* (Islam as a blessing for all) and the noble values of the Indonesian nation. Islamic Boarding School education is efficient because it is conducted in a boarding school setting, which shapes the character of its students or '*santri*.' However, the current method of determining the grade levels of *santri* is often inaccurate, relying solely on the average scores of entrance exams without considering essential aspects of subjects. This leads to a decrease in students' interest in learning and delays in achieving higher levels of education. By utilizing data mining techniques, such as the C4.5 algorithm based on Forward Selection, it is possible to address this issue and enhance the accuracy of placing *santri* into their appropriate grade levels at the Nurul Jadid Paiton Probolinggo Islamic Boarding School. This improvement can make the Islamic Boarding School education system more effective in managing student learning

Keywords: C4.5 Algorithm, Data Mining, Forward Selection, Classification, Islamic Boarding School.

INTRODUCTION

The educational process in Islamic boarding schools is carried out in the Islamic boarding school environment by developing a unique curriculum for each Islamic boarding school with an Islamic education pattern. *Mualimin* himself has *santri* as students, administrators as teachers, and *kiai* as leaders and professors at an Islamic boarding school. Education in Islamic boarding schools is very efficient considering that the daily lives of the students are far from the students' guardians so that not only science but character education is also developed in Islamic boarding schools, so that the students' affective and cognitive abilities are also well monitored (Sahlan, 2023).

The Islamic boarding school itself consists of several levels that its students will go through. So that students have a sustainable and systematic educational method in receiving the knowledge conveyed. There are many levels, but generally Islamic boarding school education has three levels of learning. The three levels are *Ula* as the initial level, *Wusto* as the middle level and for the high or final level, namely *Ulya* (Yuliani et al., 2024).

All students will be given a test before entering the Islamic boarding school to group students into levels. So that student learning can be right on target and without having to repeat what they have learned previously. However, these tests are sometimes not on target, because the grouping of scores is ineffective and only takes the average score of students without taking into account important aspects of the subjects to be taken in Islamic boarding school education. What happens is that there are students who have knowledge in the field of Islamic science but are still at an initial level (Anwar et al., 2023).

This actually makes students bored and fed up. So interest in learning decreases because what has been learned is taught again. Time efficiency and systematic learning become very small. Make students graduate late and understand higher learning. At the Nurul Jadid Paiton Probolinggo Islamic Boarding School, for example, the average score is the main reference in determining the class of students. Not a few students waste time and cannot graduate on time because of this. This system is very ineffective in determining class (Santos et al., 2023).

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

This value data does not seem very useful if the class classification of students is not on target. The dataprocessing process is actually very possible using data mining techniques. Data mining is a data analysis technique using databases and statistical calculations. Data processing of course goes through the stages of data cleaning, data transformation, data integrity and model evaluation required in data mining. Of course, the trial error system is often used in Islamic boarding schools. However, of course there are still many clear policies and weaknesses in determining student classes (Nurhalisha et al., 2023).

In 1993 J. Ross Uinlan started from ID3 and created the C4.5 Decision Tree. The C4.5 algorithm is one of the most influential decision trees available today. Compared with ID3, there is an improvement in the C4.5 algorithm. First, the C4.5 algorithm uses the information gain ratio as the attribute selection criterion, while the ID3 algorithm model takes from the subtree to use the information gain. Second, to avoid overfitting, this algorithm is able to create decision trees and pruning can be applied. Third, discrete data and complete data can be done with decision trees. The research we conducted regarding student class classification can be linked to the data mining classification method with the C4.5 algorithm (Saputra et al., 2023).

Referring to research conducted by Vista Anestiviya entitled Pattern Analysis Using the C4.5 Method Students' Main Interests Based on the Case Study Curriculum in State Senior High Schools 1. In this case study the researcher used the C4.5 method because he felt he was able to group interests prospective high school students. Because the C4.5 method has a structured and objective method in choosing data processing. Thus the chance of accuracy is higher, namely 85.40%. Previously, researchers compared it with the Navie Bayes method which had an accuracy of 85.09%. The dataset used in this research was 900 registrants. By carrying out data pre-processing stages and implementing them in Weka under the name j48. In this case, researchers also combined the C4.5 method which is felt to have higher accuracy with Particle Swarm Optimization. And the accuracy has a higher value, namely 87.61% and a precision value of 88.96% (Anestiviya et al., 2021).

One model for finding the best combination and correlation of a variable is Forward Selection. In the Forward Selection procedure, if a variable meets the criteria in an equation, then that variable cannot be removed. Apart from that, Forward Selection can also include independent variables that have the closest relationship to the independent variable. Then gradually enter the next potential independent variable and will stop until there are no more potential independent variables (Hamdani et al., 2024).

Based on the data above, the researcher wants to focus this research on the C4.5 algorithm in terms of classification. The C4.5 algorithm based on Advanced Selection for class level classification of Nurul Jadid Islamic Boarding School students is expected to be able to help the system in parsing errors. So it can support the classification of students at class level placement.

LITERATURE REVIEW

Research by X Wang, C. Zhou, X. Xu entitled Application of C4.5 Decision Trees for Scholarship Evaluation. Scholarships are often not well targeted for students because of the increasing number of students available. For this problem, researchers used the C4.5 method combined with Fuzzy mathematics and also ID3. From these results it turns out that the C4.5 algorithm has a high accuracy value even though it only has 428 student data for student scholarship applications. And also the C4.5 Method is very supportive for application in the education sector. Compared to other methods, the C4.5 algorithm is better than fuzzy math and ID3. ID3 has the lowest accuracy, namely 87.30%. Fuzzy mathematics has an accuracy of 90.45%. And the highest accuracy is owned by the C4.5 Algorithm with an accuracy of 91.59% (Yenila et al., 2022).

Likewise with research conducted by Gaurav L. Agrawal, Prof. Hitesh Gupta entitled Optimizing C4.5 for data mining, that C4.5 can be pre-processed data with other methods, which in this research is the L'Hospital Role method. This algorithm becomes better when given the L'Hospital Role method so it can be concluded that this algorithm is better and more accurate when given additional methods. In this research, C4.5 was given an additional method, namely Future Selection (Saputra et al., 2023).

Apart from that, Arief Saputro discusses how to determine student study completion by linking initial student entry data with the academic achievement of Community Academy students using the C4.5 Algorithm based on Advanced Selection and how to increase prediction accuracy. uses the C4.5 algorithm based on Forward Selection. The aim of this research is to predict the graduation of community college students using the Forward Selection-based C4.5 algorithm method approach by conducting experiments on training data, observing the ability to consume a model in the form of a decision tree. And this research is expected to improve the performance accuracy of the C4.5 algorithm using the Forward Selection method. The research results show that from 178 datasets, the C4.5 method has an accuracy of 95.84%. In the C4.5 algorithm which is based on forward selection, there is a slight increase in accuracy, namely 95.89% (Wibowo et al., 2023).

METHOD

This research uses experimental research. Experimental research using a dataset of new student exam scores at the Nurul Jadid Islamic Boarding School. This research aims to test the classification of Nurul Jadid's *santri* groups using the C4.5 algorithm based on forward selection which is expected to be able to improve the results of

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

classifying *santri* in finding group classes in teaching and learning activities. In the context of research, method refers to an organized approach to solving a problem, which includes: Data Collection, Formulation of Hypotheses or Propositions, Hypothesis Testing, Interpretation of Results, and Conclusions (Haryono, 2023).

Cross Industry Standard Process for Data Mining (CRISP-DM)

This method uses the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. This is a method for dealing with problem solving strategies using minig data (Omari Firas, 2023). Because research that is recognized/accepted must follow recognized rules, this research was carried out with the stages of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Application as in Figure 2, as follows:

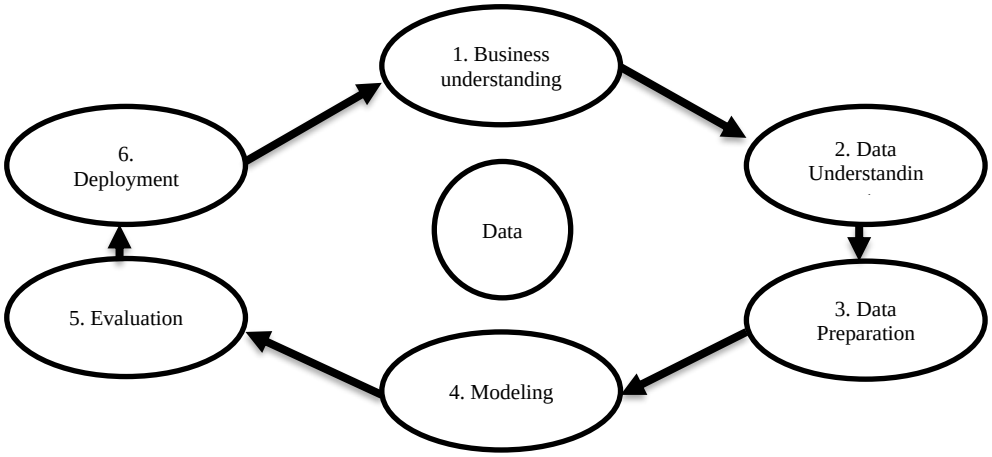


Figure 2 Researcher property

Validation with Cross Validation

The depiction of the model algorithm in Figure 3.2 will be a reference for comparison in this research (Yulhendri et al., 2023). The dataset is divided into X according to the validation value (X-Fold Cross Validation), each part of the model (1/X) will be used as test data, the rest will be used as training data. From each model the training data is processed according to the selected data pre-processing, as the training results of each selected learning algorithm. Next, the algorithm is tested and trained with testing to carry out validation. The validation results were used to measure the performance of the two research models, and a comparison was carried out to find the model that had the best performance using accuracy confusion matrix analysis.

Evaluation with Confusion Matrix

Evaluation using a confusion matrix is able to obtain accuracy, precision and recall values (Sheth et al., 2022). Accuracy in classification is the percentage of accuracy of data records that are classified correctly after testing the classification results.

Correct Classification	*Classfied as*	
	+	-
+	*True Positives*	*False Negatives*
-	*False Positives*	*True Negatives*

Table 1 Confusion Matrix

The following is the confusion matrix model equation:

$$akurasi = \frac{tp+tn}{tp+tn+fp+fn} \quad \text{“(1)”}$$

$$sensitivity = \frac{tp}{tp+fn} \quad \text{“(2)”}$$

$$specificity = \frac{tn}{tn+tp} \quad \text{“(3)”}$$

$$PPV = \frac{tp}{tp+fp} \quad \text{“(4)”}$$

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$NPV = \frac{tn}{tn + fn} \quad \text{“(5)”}$$

RESULT

PreProcessing Data

Before the data is processed, preprocessing is the initial step where the dataset is cleaned from values that are still inconsistent, such as noisy data, such as missing data or so on.

Name	Mark Fikh	Mark Moral	Mark *Aqidah*	Mark BTQ	Nila Mathematics	Mark IPA	Mark IPS	Mark ING	Graduation class
Muhammad Gharlan Al Jazair	-	-	-	-	3	3	1	1	*wustho*
Alya Azzahrah Althafunnis	-	4	2	3	3	2	4	1	*wustho*

Table 2 Table contains Noisy data such as Missing Values

After that, prepare the 2023/2024 Nurul Jadid Islamic Boarding School New Student Test dataset, then format the attributes and determine the data type according to the value, such as numeric, polynominal and label types. Here, we make the name attribute the ID and make the Class the label or target class.

Row No.	*Name (polynominal) Id*	*Regular Fiqh (Integer) value*	*Moral Value (Integer) regular*	*Regular Aqidah (Integer) value*	*Regular btq (Integer) value*	*Regular Social Sciences (Integer).*	*Sciences (Integer) regular*	*Regular (Integer) Mathematics*	*ENGLISH (Integer) regular*	*Class (polynominal) label*
1	Ahmad Danial	1	1	1	1	4	1	4	1	*Ula*
2	Ahmad Ardhan S.	1	2	2	2	3	3	4	2	*Ula*
3	Aisyah Nur F.	3	2	2	2	4	4	4	3	*Ula*
4	Amanda Dwi A.	1	2	2	2	3	4	4	4	*Ula*
5	Angga Efendi P	2	2	1	1	4	3	4	3	*Ula*
6	Bella Nawangsari	1	2	2	2	3	4	4	2	*Ula*
7	Bilhakqi Maha	1	1	1	1	4	4	4	4	*Ula*
8	Bilhakqi Maha	3	2	2	2	4	4	4	4	*Ula*
9	Cesar Putra P.	2	2	1	1	3	2	4	2	*Ula*

Table 3 Process of changing attributes and data types in RapidMiner

Missing Value "Null" data cleaning aims to clean inappropriate data in the dataset, there is some data that contains missing values. RapidMiner detected 5 Missing Value data, namely two fiqh value data, one moral value data, faith and BTQ

NAME	TYPE	MISSING
Mark Fikh	Integer	2
Mark Moral	Integer	1
Mark *Aqidah*	Integer	1
Mark Btq	Integer	1
Class	Polynominal	0
Name	Polynominal	0
IPS	Integer	0

Table 4 RapidMiner detects Missing Value data

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Then each data is cleaned using the replace missing value operator in RapidMiner. With the parameter Attribute = All and the default value Averange, the goal is to eliminate all inappropriate values in all Missing Table attributes, as shown in the following image:

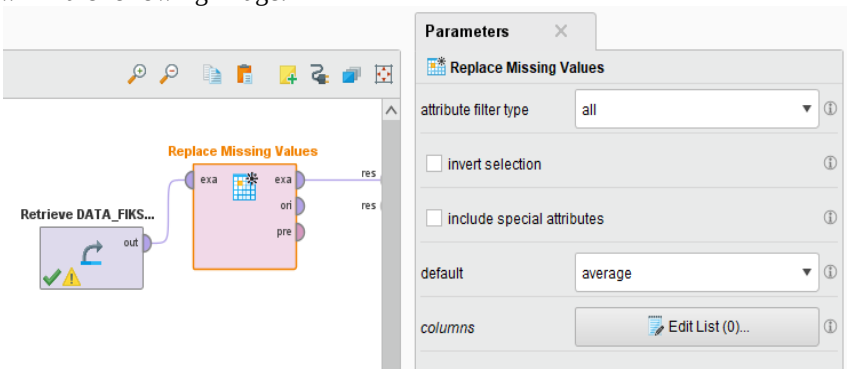


Figure 6 Rapid Miner operator deletes Missing Value data

So that the data changes and has clean data from noisy data where in the previous table there were missing values

Name	Mark Fikh	Mark Moral	Mark *Aqidah*	Mark BTQ	Nila Mathematics	Mark IPA	Mark IPS	Mark ING	Graduation class
Muhammad Gharlan Al Jazair	3	3	3	3	3	3	1	1	*wustho*
Alya Azzahrah Althafunnis	3	4	2	3	3	2	4	1	*wustho*

Table 3 Example of a results table that has been cleaned from Noisy data

Then validation will be carried out using k-Fold Cross Validation (Bates et al., 2023). Before the data is predicted, it is first validated using k-fold cross validation, meaning the data is divided into 10 parts randomly. The test data and training data part can also be called 10-fold cross validation. In this research, testing was carried out using 10-fold cross validation, namely by separating the data into two parts, using 394 data sets with test data and training data. This operation was carried out 3, 5, 10 times, and the final result was obtained, namely the accuracy value from 3 trials so that an average was formed. The resulting accuracy is accuracy: 91.86% +/- 6.13% (micro average: 91.88%). Quite significantly different from the validation results before data preprocessing, namely accuracy: 86.09% +/- 14.84% (micro average: 86.04%).

Modeling and Classification Method Using the C4.5 Algorithm

Modeling is carried out by entering the results of data preprocessing into algorithm calculations. First carry out the process using the C4.5 algorithm. The C4.5 algorithm is an algorithm used to carry out predictive classification or segmentation or grouping. The C4.5 algorithm model was generated from a dataset consisting of 394 data and 8 attributes which are presented in Table 4.1 for the attribute level classification of Nurul Jadid Islamic Boarding School students. In general, the C4.5 algorithm takes one attribute as the root and creates branches for each value. The selection of an attribute as the root is based on the highest validation score based on the existing attributes. The first is to calculate the entropy value using equation (1):

$$Entropi\ (total) = \left(-\frac{37}{394} * log_3\left(\frac{37}{394}\right)\right) + \left(-\frac{169}{394} * log_3\left(\frac{169}{394}\right)\right) + \left(-\frac{188}{394} * log_3\left(\frac{188}{394}\right)\right)$$
$$= 0.85404$$

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Then calculate the Entropy of all cases which are divided based on the attributes "fiqh value", "faith value", "moral value", "BTQ value", "MTK value", "social studies value", "knowledge value", and "science value". BING Value" in the same way, namely with equation (1). After that, testing was carried out using RapidMiner with a cross-validated Decision tree operator to validate the model simultaneously using parameters k = 10 and using the following operators:

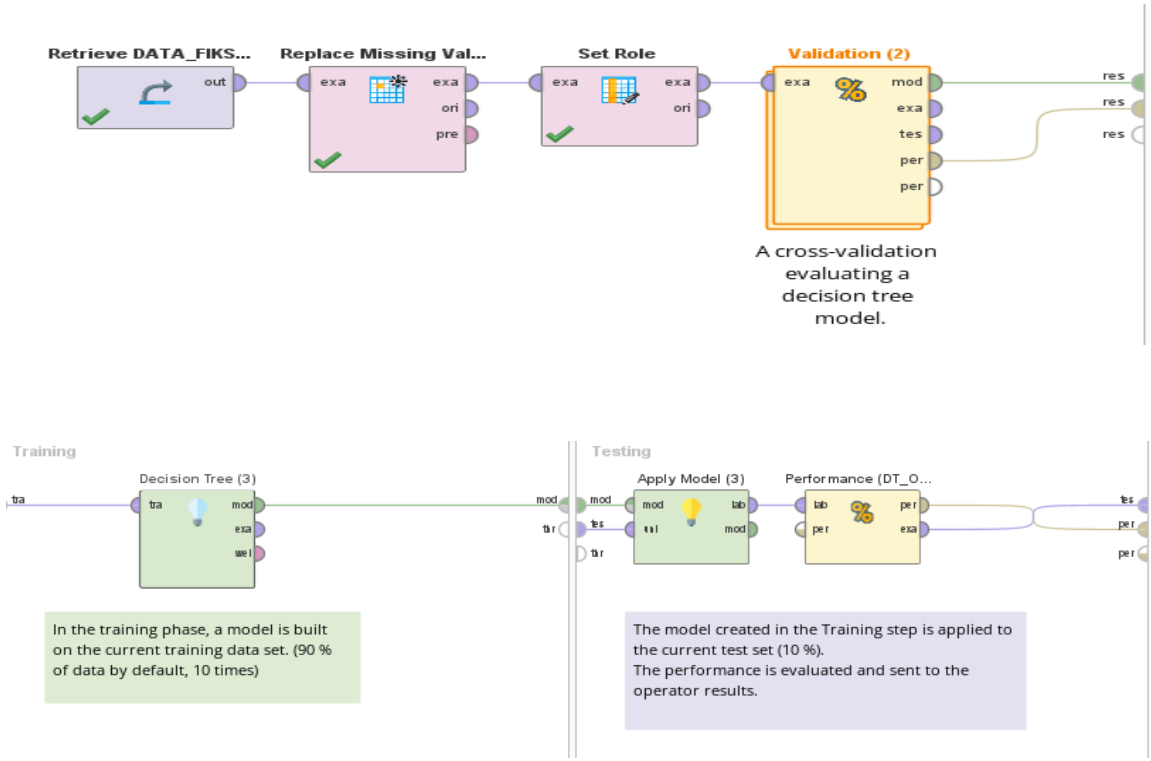


Figure 7 Operators used for modeling Algorithm C4.5

Using the RapidMiner tool, the C4.5 algorithm was found with the following overall root results:

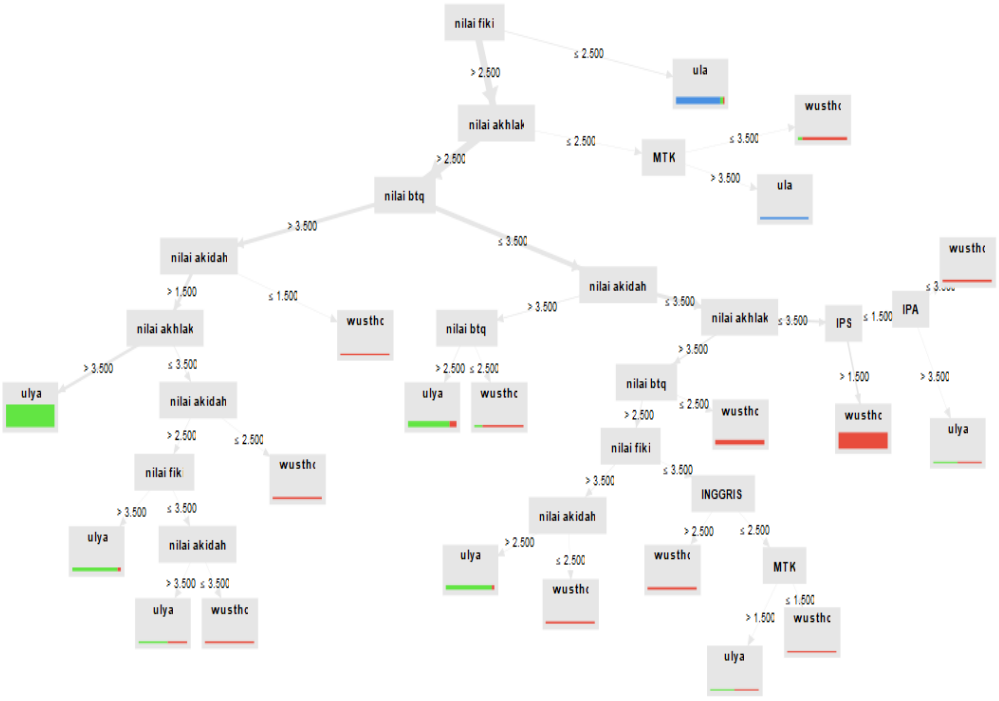


Figure 8 C4.5 Algorithm Decision Tree with RapidMiner

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In the picture above, it can be seen that fiqh is the highest root because it has the highest Gain Ratio value with a Gain Ratio value (fiqh value) = 0.7097929761116226.

After that the data is processed using the Rapid Miner tool. To test the dataset in RapidMinner using the crossvalidation operator by applying the parameter k = 10, and using the C4.5 algorithm produces an accuracy value: 90.85% +/- 6.66% (micro average: 90.86%).

accuracy: 90.85% +/- 6.66% (micro average: 90.86%)

	true ula	true ulya	true wustho	class precision
pred. ula	36	2	2	90.00%
pred. ulya	0	173	18	90.58%
pred. wustho	1	13	149	91.41%
class recall	97.30%	92.02%	88.17%	

Figure 9 Accuracy Value of Algorithm C4.5

After implementing the C4.5 Algorithm model, an evaluation will be carried out. The confusion matrix model will produce a matrix consisting of true positive or positive tuples and true negative or negative tuples. Then enter the test data that has been provided into the confusion matrix so that results such as Mean Precision and Recall are obtained as in the following image:

weighted_mean_precision: 90.96% +/- 7.48% (micro average: 90.66%), weights: 1, 1, 1

	true ula	true ulya	true wustho	class precision
pred. ula	36	2	2	90.00%
pred. ulya	0	173	18	90.58%
pred. wustho	1	13	149	91.41%
class recall	97.30%	92.02%	88.17%	

Figure 10 Mean Precision Value

weighted_mean_recall: 92.58% +/- 5.37% (micro average: 92.49%), weights: 1, 1, 1

	true ula	true ulya	true wustho	class precision
pred. ula	36	2	2	90.00%
pred. ulya	0	173	18	90.58%
pred. wustho	1	13	149	91.41%
class recall	97.30%	92.02%	88.17%	

Figure 11 Mean Recall Value

Of the 394 data, there are data on the number of true positive (TP) *ula* 36, true positive (TP) *ulya* 173, true positive (TP) *wustho* 149. Based on these data, evaluation using a confusion matrix can obtain accuracy, precision and recall values. The processed data is listed in the following table:

	Value
Accuracy	0.908
Mean_Precision	0.909
Mean_Recall	0.925

Tabel 4 Nilai *accuracy, Precision,* dan *recall dari* Agoritma *C4.5*

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Testing Method Using Forward Selection with the C4.5 Algorithm

To improve the calculation results of the C4.5 algorithm model, the Forward Selection method will be added so that it is hoped to get better accuracy results (Suhendar et al., 2024). The modeling stage aims to compare between C4.5 and C4.5 based on forward selection in determining student class classification using RapidMiner software. For Forward selection modeling using Rapidminer as follows:

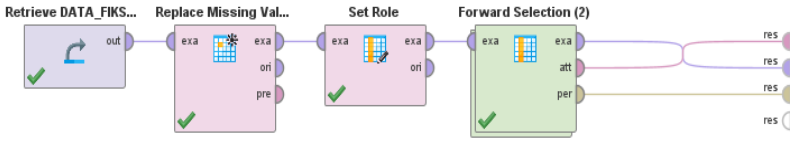


Figure 12 Modeling of the C4.5 Algorithm and Forward Selection using RapidMiner tools

Obtained modeling results with RapidMiner using five influential attributes. The following are the attributes of the results of applying Feature Forward Selection which are presented in the image:

Attribbute	Fikih	Akhlak	Akidah	BTQ	MTK	Nama	IPS	IPA	Inggris
Weight	1	1	1	1	1	0	0	0	0

Figure 13 Most influential attributes

Using the Rapidminer tool helps in the calculation process to find the value of each attribute. The results obtained from using the Forward Selection operator in Rapidminer produce values for each attribute which are expressed in binary form 0 and 1. After knowing which attributes have an influence, the next step is to model the C4.5 algorithm using the Forward Selection method. By using Rapidminer, with previously processed data, a Forward Selection operator will be added. After adding the Forward Selection operator, then adding cross validation using the parameter k = 10, then adding the C4.5 algorithm to process the dataset.

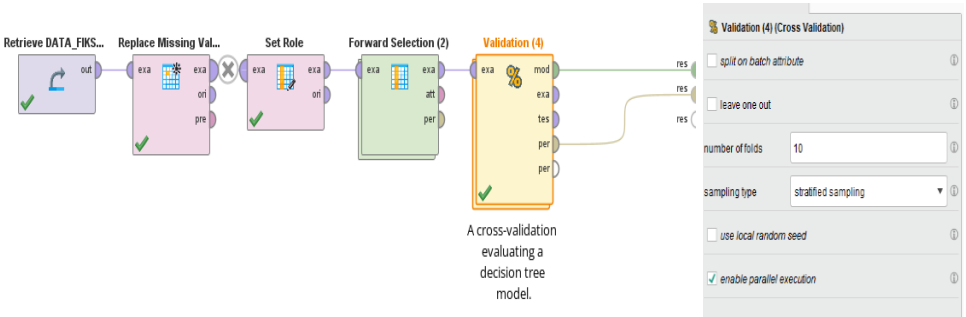


Figure 14 C4.5 modeling operators and Forward Selection in RapidMiner

The result of combining Forward Selection with the C4.5 Algorithm is accuracy: 94.68% +/- 2.49% (micro average: 94.67%) as seen in Figure 15:

accuracy: 94.68% +/- 2.49% (micro average: 94.67%)

	true ula	true ulya	true wustho	class precision
pred. ula	37	2	1	92.50%
pred. ulya	0	180	12	93.75%
pred. wustho	0	6	156	96.30%
class recall	100.00%	95.74%	92.31%	

Figure 15 Results of the C4.5 Algorithm and Forward Selection

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

For an overview of the decision tree and roles in the C4.5 Algorithm using Forward Selection, see Figure 4.7 as follows:

Gambar 16 Pohon dari Algoritma C4.5 dan *Forward Selection*


```
nilai fiki > 2.500
|   nilai akhlak > 2.500
|   |   nilai btq > 3.500
|   |   |   nilai akidah > 1.500
|   |   |   |   nilai akhlak > 3.500: ulya {ula=0, ulya=122, wustho=0}
|   |   |   |   nilai akhlak ≤ 3.500
|   |   |   |   |   nilai akidah > 2.500
|   |   |   |   |   |   nilai fiki > 3.500: ulya {ula=0, ulya=14, wustho=1}
|   |   |   |   |   |   nilai fiki ≤ 3.500
|   |   |   |   |   |   |   nilai akidah > 3.500: ulya {ula=0, ulya=3, wustho=2}
|   |   |   |   |   |   |   nilai akidah ≤ 3.500: wustho {ula=0, ulya=0, wustho=5}
|   |   |   |   |   nilai akidah ≤ 2.500: wustho {ula=0, ulya=0, wustho=6}
|   |   |   nilai akidah ≤ 1.500: wustho {ula=0, ulya=0, wustho=2}
|   |   nilai btq ≤ 3.500
|   |   |   nilai akidah > 3.500
|   |   |   |   nilai btq > 2.500: ulya {ula=0, ulya=24, wustho=4}
|   |   |   |   nilai btq ≤ 2.500: wustho {ula=0, ulya=1, wustho=5}
|   |   |   nilai akidah ≤ 3.500
|   |   |   |   nilai akhlak > 3.500
|   |   |   |   |   nilai btq > 2.500
|   |   |   |   |   |   nilai akidah > 2.500
|   |   |   |   |   |   |   nilai fiki > 3.500: ulya {ula=0, ulya=19, wustho=1}
|   |   |   |   |   |   |   nilai fiki ≤ 3.500: wustho {ula=0, ulya=0, wustho=6}
|   |   |   |   |   |   nilai akidah ≤ 2.500: wustho {ula=0, ulya=1, wustho=11}
|   |   |   |   |   nilai btq ≤ 2.500: wustho {ula=0, ulya=0, wustho=21}
|   |   |   |   nilai akhlak ≤ 3.500: wustho {ula=0, ulya=1, wustho=94}
|   nilai akhlak ≤ 2.500
|   |   MTK > 3.500: ula {ula=6, ulya=0, wustho=0}
|   |   MTK ≤ 3.500: wustho {ula=0, ulya=1, wustho=10}
nilai fiki ≤ 2.500: ula {ula=31, ulya=2, wustho=1}
```

In the picture above, it can be seen that jurisprudence is still the highest root because apart from having the highest Gain Ratio value, the jurisprudence value in the Forward selection method has a weight value that influences the student dataset. And it turns out that the Advanced selection method states that the values of fiqh, akhlak, btq, aqidah, and MTK have relevant weights for classifying students.

Figure 17 description of the C4.5 Algorithm Role and Forward Selection

Based on the resulting rules, 16 rules are formed from the C4.5 Algorithm decision tree and Advanced Selection has a total of 2 rules for the ula class, 9 rules for the wustho class and 5 rules for the ulya class in predicting the class level of Nurul Jadid Islamic Boarding School students. After implementing the C4.5 model and the Forward Selection Algorithm, an evaluation will be carried out on the Confusion Matrix model which will produce a matrix consisting of true positive or positive tuples and true negative or negative tuples. Then enter the test data provided in the confusion matrix so that results are obtained as in the following image:

accuracy: 93.91% +/- 3.21% (micro average: 93.91%)

	true ula	true ulya	true wustho	class precision
pred. ula	37	2	1	92.50%
pred. ulya	0	179	14	92.75%
pred. wustho	0	7	154	95.65%
class recall	100.00%	95.21%	91.12%	

Figure 18 Accuracy value of the C4.5 Algorithm and Forward Selection

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

weighted_mean_recall: 95.48% +/- 2.49% (micro average: 95.45%), weights: 1, 1, 1

	true ula	true ulya	true wustho	class precision
pred. ula	37	2	1	92.50%
pred. ulya	0	179	14	92.75%
pred. wustho	0	7	154	95.65%
class recall	100.00%	95.21%	91.12%	

Figure 19 Mean Precision Value of the C4.5 Algorithm and Forward Selection

weighted_mean_precision: 94.24% +/- 3.57% (micro average: 93.63%), weights: 1, 1, 1

	true ula	true ulya	true wustho	class precision
pred. ula	37	2	1	92.50%
pred. ulya	0	179	14	92.75%
pred. wustho	0	7	154	95.65%
class recall	100.00%	95.21%	91.12%	

Figure 20 Mean Recall Value from the C4.5 and Forward Selection Algorithms

In accordance with the image above which was produced by processing the confusion matrix evaluating the C4.5 Algorithm and Forward selection using the RapidMiner tool, the accuracy value obtained is accuracy: 93.91% +/- 3.21% (micro average: 93.91%), mean Recall is Weighted_mean_recall: 95.48 % +/- 2.49% (micro average: 95.45%), weight: 1, 1, 1 and average Precision is Weighted_mean_precision: 94.24% +/- 3.57% (micro average: 93.63%) , weight: 1, 1, 1. Of the 394 data there is data on the number of true positive (TP) *ula* 37, true positive (TP) *ulya* 179, true positive (TP) *wustho* 154. Based on this data an evaluation was carried out using the Confusion Matrix which was able to obtain accuracy, precision and recall values. The processed data is listed in the following table:

	Nilai
Accuracy	0.939
Mean_Precision	0.942
Mean_Recall	0.954

Table 5 accuracy, precision, and recall of C4.5 and Forward Selection

DISCUSSIONS

From the test results, it can be seen that there are 5 attributes that influence the test, namely moral values, aqidah, fiqh, BTQ and MTK. Comparison of the test results of the C4.5 algorithm model without attribute selection with the C4.5 model with attribute selection using forward selection is shown in Table 6 below:

	Algorima C4.5	Algorima C4.5 dan Forward Selection
Accuracy	**0.908**	**0.939**
Mean_Precision	**0.909**	**0.942**
Mean_Recall	**0.925**	**0.954**

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 6 Testing C4.c5 and C4.5 using Forward Selection

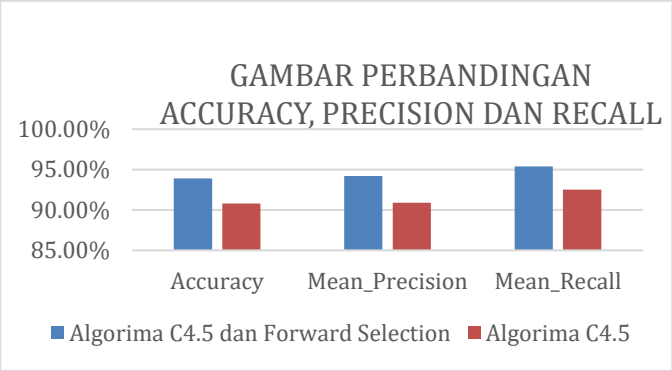
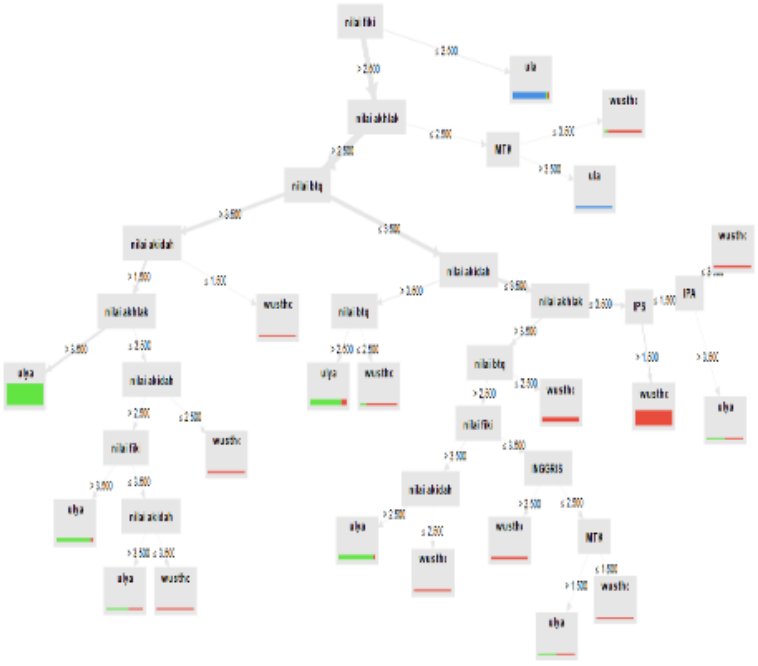


Figure 21 Comparison of C4.5 and C4.5 with Forward Selection

Based on the results of testing with the Evaluation Confusion Matrix, it is proven that testing carried out by optimizing the C4.5 algorithm with Forward Selection has a higher accuracy score than using only the C4.5 algorithm (Rahmayani, 2023). The accuracy score produced by the C4.5 algorithm model is accuracy: 90.85% with a mean recall of 92.58% and a mean precision of 90.96%. And the accuracy score of the C4.5 model using the Forward Selection method has an accuracy value of 93.91% with a mean recall of 95.48% and a mean precision of 94.24%. From this score, it can be seen that the difference in accuracy is 03.06%. Meanwhile the difference from recall is 02.90% and the difference from precision is 03.55%.

Analysis of the C4.5 Algorithm and C4.5 Algorithm with Forward Selection

From the results obtained, it can be seen that the C4.5 algorithm itself has a lower level of accuracy than the C4.5 algorithm combined with the Forward Selection method. So these results prove that the Forward Selection method can add a better level of accuracy so that Forward Selection can be compared with the C4.5 Algorithm. The advantage of the C4.5 algorithm is that decision making is quite simple and specific compared to previous decision making which was complex and global (Sunarto et al., 2023). And the C4.5 algorithm fixes unnecessary calculations, because when creating a decision tree the data samples are only tested based on certain class criteria.



Apart from that, the advantage of the C4.5 algorithm is that it is flexible in selecting features from different internal nodes, the selected features will differentiate a criterion compared to other criteria in the same node. The flexibility of the decision tree method improves the quality of the resulting decisions when compared to more conventional one-stage calculation methods. [Hamsir Saleh (2014]

* Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13514
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Figure *22 Algorima C4.5*

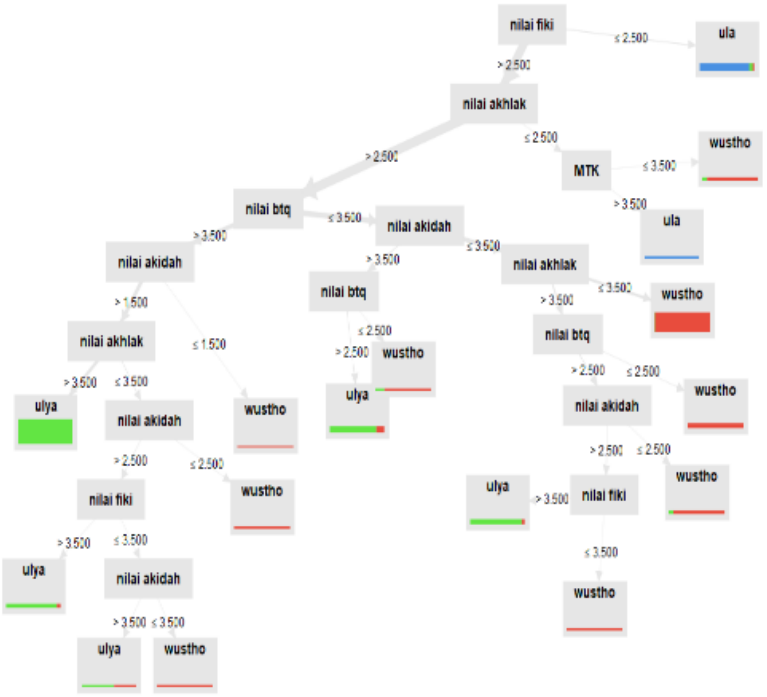


Figure *23 C4.5 and Forward Selection Algorithm*

The C4.5 algorithm it self cannot be better than when combined with forward selection because according to research by Moh Efendi Lasulika and Andi Bode (2021), the C4.5 algorithm is unstable. Because this makes even small changes to the data can change the results of the decision tree of the C4.5 algorithm to a large extent. that a normal event user obtains, may be different from what has been processed with the C4.5 algorithm. Apart from that, another thing that causes low accuracy values is the accumulated number of errors from each level in a large decision tree. This causes when the C4.5 algorithm stands alone without other methods, making a decision tree produces excessive error results, causing the accuracy value to decrease quite significantly.

So, because of the shortcomings above, it is deemed necessary to add a Forward Selection method to increase the accuracy of the calculation results of the C4.5 algorithm. Feature Selection which aims to get value from better accuracy, as well as getting an attribute model by applying Feature Selection. Forward Selection or forward selection is used to analyze attributes or forward selection features that can be initiated without predictors in the model to help improve results rather than accuracy and determine the most influential attributes. From here the Forward selection method helps reduce the load on the C4 algorithm. 5 is analyzed because the Forward selection method has been selected first. Forward selection helps reduce the accumulated number of errors from each level in the decision tree, and helps stabilize the algorithm due to the impact of selection.

CONCLUSION

From the test results it can be concluded that there are five attributes that have a significant influence on the test, namely moral values, aqidah, fiqh, BTQ, and MTK. The research results show that the C4.5 algorithm has a lower level of accuracy compared to the Forward Selection method. Therefore, these findings support the view that Forward Selection can improve the level of accuracy better, so that the Forward Selection method can be considered equivalent to the C4.5 Algorithm. In this context, the Forward Selection method contributes to reducing the burden of analysis with the C4.5 algorithm, because this method has carried out an initial selection. In addition, Forward Selection plays a role in reducing the accumulated number of errors from each level of the decision tree, as well as helping to maintain algorithm stability by reducing the impact of the selection process.

ACKNOWLEDGMENT

The author would like to thank for providing data from the Nurul Jadid Islamic boarding school and have given us permission to carry out this research until completion, as well as support from facilities from Dian Nuswntoro University, without which this research would not have been carried out impossible to happen.

* Corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

REFERENCES

Anestiviya, V., Ferico, A., & Pasaribu, O. (2021). Analisis Pola Menggunakan Metode C4.5 Untuk Peminatan Jurusan Siswa Berdasarkan Kurikulum (Studi Kasus : Sman 1 Natar). *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(1), 80–85.

Anwar, H. S., Denata, R., & Firdaus, A. I. I. (2023). Digitalisasi Pendidikan Pesantren melalui Sistem Pembayaran Cashless Menggunakan Ngabar Smart Payment di Pondok Pesantren Wali Songo Ngabar. *MA'ALIM: Jurnal Pendidikan Islam*, *4*(1), 43–53. https://doi.org/10.21154/maalim.v4i1.6678

Bates, S., Hastie, T., & Tibshirani, R. (2023). Cross-Validation: What Does It Estimate and How Well Does It Do It? *Journal of the American Statistical Association*, 1–43. https://doi.org/10.1080/01621459.2023.2197686

Hamdani, R., Sriani, & Darta, A. (2024). ANALISIS TINGKAT KEPUASAN PASIEN DI KLINIK PRATAMA. *Journal of Science and Social Research*, *4307*(1), 273–280.

Haryono, E. (2023). METODOLOGI PENELITIAN KUALITATIF DI PERGURUAN TINGGI KEAGAMAAN ISLAM. *An Nur*, *13*(2), 23–30.

Lasulika, M. E., & Bode, A. (2021). Komparasi Algoritma Data Mining Menggunakan Forward selection Pada Prediksi Harga Jagung. *Jurnal Tecnoscienza*, *5*(2), 157. https://doi.org/10.51158/tecnoscienza.v5i2.392

Nurhalisha, T., Pranoto, W. J., Saputra, H., & Latipah, A. J. (2023). The application of particle swarm optimization (PSO) to improve the accuracy of the naive bayes algorithm in predicting floods in the city of Samarinda. *Journal of Intelligent Decision Support System*, *6*(3), 138–146.

Omari Firas. (2023). A combination of SEMMA & CRISP-DM models for effectively handling big data using formal concept analysis based knowledge discovery: A data mining approach. *World Journal of Advanced Engineering Technology and Sciences*, *8*(1), 009–014. https://doi.org/10.30574/wjaets.2023.8.1.0147

Rahmayani, F. (2023). Penerapan Algoritma C4.5 Dengan Feature Forward Selection Untuk Analisis Capaian Indikator Kinerja Utama Berdasarkan Tracer Study (Studi Kasus: Fasilkom Unsika). *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(4), 2732–2738.

Sahlan, M. (2023). *Konsep Dakwah Pluralisme Dalam Trilogi Santri Maqolahhusnu Al-Adab Ma'a Allah Wa Ma'a Al-Kholqi Di Pondok Pesantren Nurul Jadid*. UNIVERSITAS NURUL JADID.

Santos, M. S., Abreu, P. H., Japkowicz, N., Fernández, A., & Santos, J. (2023). A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research. *Information Fusion*, *89*(June 2022), 228–253. https://doi.org/10.1016/j.inffus.2022.08.017

Saputra, M., Sinaga, I. P. L., & Sinaga, W. (2023). Analisis penerapan algoritma c4.5 untuk mengidentifikasi kelancaran kredit nasabah pada cu. dosnitahi pinang sori. *Jurnal TEKINKOM*, *6*(2), 737–743. https://doi.org/10.37600/tekinkom.v6i2.943

Sheth, V., Tripathi, U., & Sharma, A. (2022). A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. *Procedia Computer Science*, *215*, 422–431. https://doi.org/10.1016/j.procs.2022.12.044

Suhendar, A. H., Rohmawati, A. A., & Prasetyowati, S. S. (2024). Performance of CART Time-Based Feature Expansion in Dengue Classification Index Rate. *Sinkron: Jurnal Dan Penelitian Teknik Informatika*, *9*(1), 1–9.

Sunarto, A., Kencana, P. N., Munadjat, B., Dewi, I. K., Abidin, A. Z., & Rahim, R. (2023). Application of Boosting Technique with C4.5 Algorithm to Reduce the Classification Error Rate in Online Shoppers Purchasing Intention. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *14*(2), 01–11. https://doi.org/10.58346/JOWUA.2023.I2.001

Wibowo, A., Wardani, S., Dewantoro, R. W., Wesly, W., & Leonardo. (2023). Komparasi Tingkat Akurasi Random Forestdan Decision TreeC4.5 Pada Klasifikasi Data Penyakit Infertilitas. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *4*(1), 218–224. https://doi.org/10.30865/klik.v4i1.1115

Yenila, F., Wahyuni, S., Rianti, E., Marfalino, H., & Gusmita, D. (2022). Sistem Pakar Deteksi Hemangioma pada Batita menggunakan Metode Hybrid. *Jurnal Informasi Dan Teknologi*, *4*(4), 265–270. https://doi.org/10.37034/jidt.v4i4.250

* Corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Yulhendri, Y., Malabay, M., & Kartini, K. (2023). Correlated Naïve Bayes Algorithm to Determine Healing Rate of Hepatitis Patients. *International Journal of Science, Technology & Management*, *4*(2), 401–410. https://doi.org/10.46729/ijstm.v4i2.776

Yuliani, W., Studi, P., Agama, P., Bukittinggi, D., Wati, S., Studi, P., Agama, P., Bukittinggi, D., Pesantren, P., & Thawalib, S. (2024). Sistem Pendidikan Pesantren Modern Studi Kasus Pendidikan Pesantren Di Pondok Pesantren Sumatera Thawalib Parabek. *Jurnal Pendidikan Dan Keguruan*, *2*(1), 54–63.

* Corresponding author