




Performance of Decision Tree C4.5 Algorithm in Student Academic Evaluation

Edy Budiman¹(✉), Havaluddin¹(✉) , Nataniel Dengan¹,
Awang Harsa Kridalaksana¹, Masna Wati², and Purnawansyah²

¹ Faculty of Computer Science and Information Technology,
Mulawarman University, Samarinda, Indonesia
edy.budiman@fkti.unmul.ac.id, havaluddin@gmail.com,
ndengen@gmail.com, awangkid@gmail.com

² Faculty of Computer Science, Universitas Muslim, Makassar, Indonesia
masnawati.ssi@gmail.com, purnawansyah@gmail.com

Abstract. Student academic evaluation is part of academic information system (AIS) performance, in order to control student learning progress is necessary. Furthermore, the evaluation showing whether the student will pass or fail would benefit the student/instructor and act as a guide for future recommendations/evaluations on performance. An in depth study on the student academic evaluation technique by using Decision Tree C4.5 has been conducted. Specific parameters including age, place of birth, gender, high school status (public or private), department in high school, organization activeness, age at the start of high school level, and progress GPA (pGPA) and Total GPA (tGPA) from semester 1–4 with three times graduation criteria (i.e., *fast*, *on*, and *delay*) times have been defined and tested. The scope of the paper has been set for undergraduate programs. The experimental results show that accuracy algorithm (AC) of 78.57% with true positive rate (TP) of 76.72% by using quality training data of 90% have best performance accuracy value.

Keywords: Tree C4.5 · Confusion matrix · Student academic evaluation

1 Introduction

Learning process evaluation is a process to determine an academic performance level of students which comprehensive and continuous in accordance with educational regulations. Where, the student's achievement of subjects mastery are determined by quizzes, examinations, practicums, and other tasks that covering cognitive, affective, and psychometrics capacity [1–3]. Furthermore, in general, student academic assessment are based on progress report including progress GPA (pGPA) and Total GPA (tGPA). Where, pGPA and tGPA are calculated from course subjects values. Therefore, it is of great interest to identify the students to understand which factors have a larger influence on this. Hence, a data mining model is an appropriate tool for covering these tasks, i.e., classification [4], prediction [5, 6], cluster [7] etc. [8].

Furthermore, an application of data mining in the Educational context is referred as Educational Data Mining (EDM) that defined by the International Educational Data

Mining Society [9]. In other words, EDM is talked about fields of education and information or computer sciences [2, 10]. Numerous methods in data mining are widely applied in order to perform student academic evaluation tasks, including statistical and smart computing methods. [11] have implemented two classification techniques, namely Naïve Bayes and Decision Tree Classifier to model academic attrition (loss of academic status) at the Universidad Nacional de Colombia. This studied were used academic datasets 2007-II and 2012-II from two programs, Agricultural (AE) and Computer and Systems (CE) Engineering. The results showed that NBC and Decision Tree models can be used as models in the prediction of the loss of academic status. [12] have conducted research with C4.5 and ID3 algorithms of student dropout, predicting and characterizing students at the University Simón Bolívar. This experiment was used WEKA as a tools for data processing. The results of this study confirmed that these algorithms can be used as an alternatives model. [13] have conducted study Naive Bayes, the 1-NN and the WINNOWER algorithms in order to predict a student's performance. The results showed that this algorithm was the most appropriate to be used for the construction a software support tool.

The aim of this study is to investigate Tree C4.5 algorithm in order to student academic learning evaluate performance. Therefore, all students might improve and increase the learning process. It is expected that this model analysis can be used in order to support academic decisions. This paper is consists of four sections. Section 1 is the motivation to do the writing of the article. Next, the methodology and techniques is discussed in Sect. 2. Section 3 presents the experimental results and discussion, and finally Sect. 4 describes the research summaries and conclusion.

2 Methodology

2.1 Tree C4.5 Algorithm

Decision tree is a data structure consisting of nodes (i.e., root, branch, leaf) and edge. Tree C4.5 algorithm is a part of decision tree algorithm that supervised learning method [13, 14]. Tree C4.5 developed by Quinlan in the 1996s, which is derived from the algorithm *Iterative Dichotomiser* (ID3), efficient, powerful and popular [4]. In general, the C4.5 algorithm consists of two processes; preparation of decision tree and make the rules (structure and design). Then, calculate the entropy and information gain with the highest attribute is selected.

In principle, Tree C4.5 algorithm consists of four steps in order to generate decision tree. First, choose attribute as a root. Second, generate branch every value. Third, put dataset in branch, and. Four, repeat the second process until every class have the same value. Formula of Entropy is shown below where S is entropy, and p is class proportion in the output.

$$\text{Entropy}(S) = \sum_{i=1}^n -p_i * \text{Log}_2 p_i \quad (1)$$

Furthermore, the attribute with the highest gain value is used as the root attribute. Equation 2 shows the formula of the gain where, S is a set of case; A is an attribute of case; $|S_i|$ is a number of cases to i ; and $|S|$ is number of cases in the set.

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

The pseudocode of Tree C4.5 is shown as follow:

```

Input: an attribute-valued dataset  $D$ 
Tree = [14]
if  $D$  is "pure" or other stopping condition met then
    stop
end if
for all attribute  $a \in D$  do
    Compute information-theoretic condition if we split on  $a$ 
end for
 $a_{best}$  = Best attribute according to above computed condition
Tree = Create a decision node that tests  $a_{best}$  in the root
 $D_v$  = Induce sub-datasets from  $D$  based on  $a_{best}$ 
for all  $D_v$  do
    Tree $_v$  = C4.5 ( $D_v$ )
    Attach Tree $_v$  to the corresponding branch of Tree
end for
return Tree

```

2.2 Datasets

In this study, the student dataset includes biographical, academic portfolios, course duration, and student participation in the organization's activities has been used. The data were collected from academic information system (AIS) in 2014–2017 (279 samples data). Before training, all datasets will be normalization by using cleaning, integration and transformation, Fig. 1. First, cleaning process; total data collected of 459, then 180 data have been cleaned because some attribute value uncompleted. Second, integration and transformation process; total attribute value of 15, then 11 attribute have been applied in order to reduce and integrated unconditional attributes.

Furthermore, the performance of Tree C4.5 algorithm is measured by using the confusion matrix (CM) in which the true positive rate (TP) has been applied. Then, Rapid Miner Studio 7.3 software for the process of calculation and modeling has been used (Table 1).

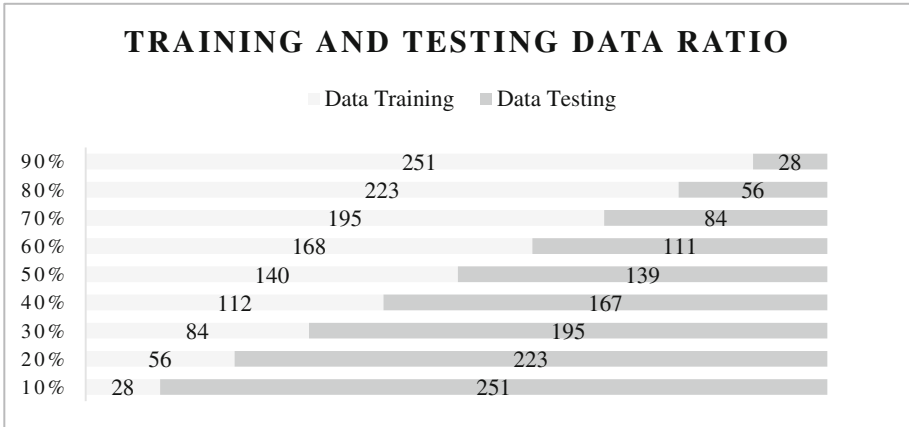


Fig. 1. Distribution of training data

Table 1. Data attribute after integration and transformation

No.	Attributes	Scale	Argument
1	Sex	Nominal	Male, Female or (M, F)
2	Age	Ordinal	Student age
3	Place of birth	Nominal	Town, Village
4	School status	Nominal	State, Private
5	School program	Nominal	Science, Non-science
6	GPA semester 1	Ordinal	1 ($GPA \leq 1.5$)
7	GPA semester 2	Ordinal	2 ($1.5 < GPA \leq 2.5$)
8	GPA semester 3	Ordinal	3 ($2.5 < GPA \leq 3.5$)
9	GPA semester 4	Ordinal	4 ($3.5 < GPA \leq 4.0$)
10	Organization	Nominal	Activist, Non-activist
11	Graduation time	Ordinal	Delay-time (>4,6 years) On-time (4–4,6 years) Fast-time (<4 years)

2.3 Performance of Evaluation

In this study, confusion matrix (CM) and with true positive rate (TP) for evaluation of Tree C4.5 model have been implemented. Where, CM is a matrix of prediction that will be compared with the original class of input, Tables 2 and 3. In other words, the matrix contains the actual value information and predictions on the classification [15]. Then, the equation of the accuracy (AC) measurement is shown as follows, where, AC is accuracy percentage proportion of predictions correct number; a is the exact number of predictions for the “Fast-Time” graduation; b is the exact number of predictions for the “On-Time” graduation; c is the exact number of predictions for the “Delay-Time” graduations; and N is total training data.

$$AC = \frac{a + b + c}{N} \tag{3}$$

Where, *a* is correct number of predictions, that negative instance; *b* is wrong number of predictions, that negative instance; *c* is wrong number of predictions, that positive instance; and *d* is correct number of predictions, that negative instance.

Table 2. Confusion matrix 2 class

		Predicted	
		Negative	Positive
Actual	Negative	<i>a</i>	<i>b</i>
	Positive	<i>c</i>	<i>d</i>

Source: [15]

Table 3. Confusion matrix of Tree C4.5 algorithm

Confusion matrix	Time		
	Fast	On	Delay
Fast-time	115	27	8
On-time	19	11	46
Delay-time	6	10	9

In this study, total course subject has been used as a student academic evaluation in Year 1, 2, and 3. In other words, student will be through the next level by this evaluation. The student evaluation term can be seen in Table 4.

Table 4. Student evaluation term

Evaluation	Degree	
Year I	Total course subject	24
	Total GPA	2,00
Year II	Total course subject	48
	Total GPA	2,00
Year III	Total course subject	72
	Total GPA	2,00

Meanwhile, true positive rate (TP) is also implemented for measured training data of Tree C4.5 model. The formula of TP as follows.

$$TP = \frac{\sum_{i=1}^3 \frac{a_i}{n_i}}{3} \tag{4}$$

Where, TP is a percentage of predictions correct number; a_i is the exact number of predictions for the “fast, on, delay” graduation time; n_i is total of training data for the “fast, on, delay” graduation time. In this study, Receiver Operating Characteristic (ROC) was not chosen for evaluation the model because ROC analysis is particularly useful for threshold selection of CM and TP. Furthermore, in this study, analysis stages using Tree C4.5 algorithm is shown in Fig. 1.

3 Experimental and Results

This section describes the test of student academic evaluation variables using Tree C4.5 models. Based on predetermined rules, nine training and testing classes’ dataset have been established. In this experiment, the dataset among others students’ academic

Table 5. Training and testing dataset

Confusion matrix	Training data		
	Fast-time	On-time	Delay-time
10%	101	21	4
	20	14	41
	19	13	18
20%	78	13	8
	32	18	24
	15	11	24
30%	72	10	3
	22	15	6
	15	12	40
40%	63	8	2
	26	17	12
	5	7	28
50%	47	5	3
	14	9	19
	17	12	13
60%	43	8	6
	13	9	2
	6	4	20
70%	35	8	5
	9	6	2
	3	2	14
80%	24	5	3
	3	5	0
	4	1	11
90%	14	3	1
	1	2	0
	1	0	6

performance evaluation variables including age, place of birth, gender, high school status (public or private), department in high school, organization activeness, age at the start of high school level, and *pGPA* and *iGPA* from semester 1–4. Furthermore, 10% to 90% of CM as a quality training data has been explored. Meanwhile, in order to get the best accuracy, CM as a performance of Tree C4.5 algorithm by using three times criteria (i.e., *fast*, *on*, and *delay* times) has been utilized, Table 5.

Based on the experiment conducted, the CM of Tree C4.5 algorithm shows that 78.57% AC with 76.72% TP of 90% quality training data have best accuracy value. It means that the best accuracy of Tree C4.5 algorithm is obtained when using 90% of training data ratio as shown in Table 6 and Fig. 2.

Table 6. Confusion matrix and true positive rate of Tree C4.5 algorithm

Algorithm evaluation	Algorithm Accuracy (AC)	True Positive rate (TP)
Training data ratio	10%	52.99%
	20%	53.81%
	30%	65.13%
	40%	64.29%
	50%	49.64%
	60%	64.86%
	70%	65.48%
	80%	71.43%
	90%	78.58%

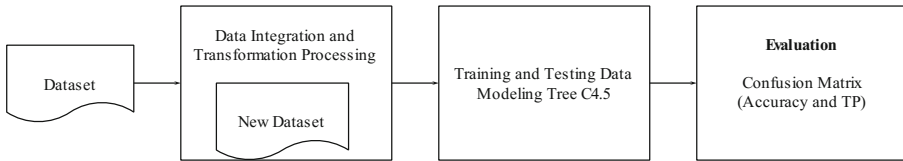


Fig. 2. Analysis stages of Tree C4.5 algorithm

The best performance of entropy and gain by using training data of 251 has been established. Where, GPA semester 4 as an initial node (root) has been settled. Detailed of entropy and gain values with 90% training data can be seen in Table 7. Based on Table 7, the highest gain value for the initial node of the manual calculation on the GPA semester 4 variables with 1.019 has been found. In other words, the initial node has corresponded with modeling (Fig. 3).

Table 7. Entropy and gain values with 90% training data

Root	Total graduation	Fast-time	On-time	Delay-time	Entropy	Gain
	251	140	48	63	1,43	
Sex						0,94
M	182	97	38	47	0,49	
F	69	43	10	16	0,47	
Age						0,96
16	1	0	1	0	0	
17	36	22	4	10	0,43	
18	156	83	35	38	0,49	
19	46	26	7	13	0,46	
20	5	3	1	1	0,49	
...	
23	2	2	0	0	0	
Place of birth						0,94
Town	107	59	22	26	0,49	
Village	144	81	26	37	0,48	
School status						0,95
State	193	101	43	49	0,49	
Private	58	39	5	14	0,41	
School program						0,94
Science	173	111	29	33	0,48	
Non-Science	78	29	19	30	0,48	
Organization						0,95
Activist	103	47	19	37		
Non-Activist	148	93	29	26		
GPA Sem. 1						0,96
1	0	0	0	0	0	
2	0	0	0	0	0	
3	170	73	42	55	0,50	
4	81	67	6	8	0,38	
GPA Sem. 2						1,01
1	0	0	0	0	0	
2	0	0	0	0	0	
3	156	56	39	61	0,49	
4	95	84	9	2	0,28	
GPA Sem. 3						0,99
1	0	0	0	0	0	
2	0	0	0	0	0	
3	173	74	38	61	0,49	
4	78	66	10	2	0,31	

(continued)

Table 7. (continued)

Root	Total graduation	Fast-time	On-time	Delay-time	Entropy	Gain
	251	140	48	63	1,43	
GPA Sem. 4						1,02
1	0	0	0	0	0	
2	0	0	0	0	0	
3	135	43	32	60	0,49	
4	116	97	16	3	0,31	

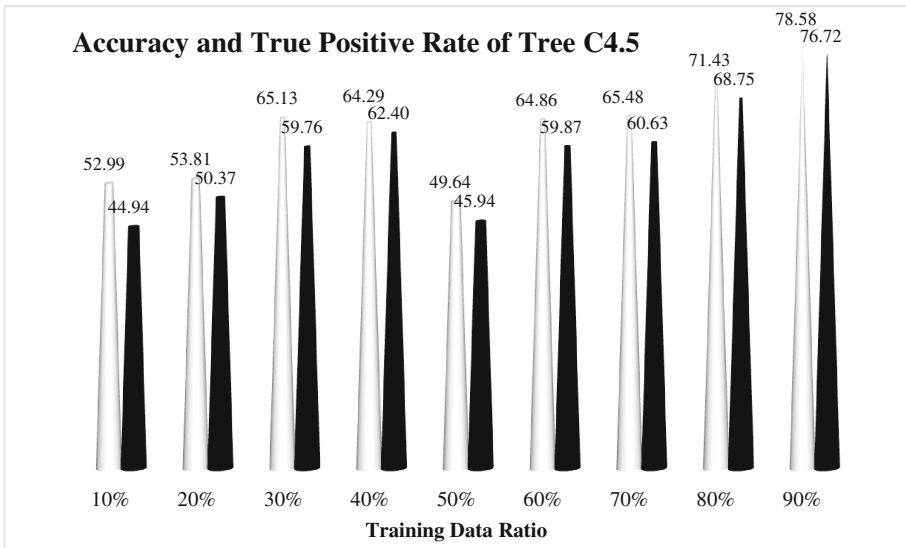


Fig. 3. Graphic of confusion matrix and true positive rate of Tree C4.5 algorithm

4 Conclusion

This paper has presented the Tree C4.5 algorithm in order to evaluate students' academic performance. Based on experiment, particular variables includes student activeness in the organization (activist and non-activist), place of birth, and age have been influence in student academic performance. This study indicated that Tree C4.5 algorithm have an accuracy better on evaluate students' academic performance. In other words, Tree C4.5 algorithm could be applied as an alternative model in student academic evaluation. Therefore, one of the planned future works is to implement Naïve Bayes Classifier (NBC), K-Means Cluster and Support Vector Machine (SVM) algorithms in order to get the better accuracy performance.

References

1. Ktona, A., Xhaja, D., Ninka, I.: Extracting relationships between students' academic performance and their area of interest using data mining techniques. In: 2014 Sixth International Conference on Computational Intelligence, Communication Systems and Networks. IEEE (2014)
2. Xu, B., et al.: Clustering educational digital library usage data: a comparison of latent class analysis and K-means algorithms. *J. Educ. Data Min.* **5**(2), 38–68 (2013)
3. Mahboob, T., Irfan, S., Karamat, A.: A machine learning approach for student assessment in e-learning using Quinlan's C4.5, Naïve Bayes and random forest algorithms. IEEE (2016)
4. Lakshmi, B.N., Indumathi, T.S., Ravi, N.: A study on C.5 decision tree classification algorithm for risk predictions during pregnancy. In: International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST-2015), Procedia Technology (2016)
5. Haviluddin, et al.: Modelling of network traffic usage using self-organizing maps techniques. In: 2016 2nd International Conference on Science in Information Technology (ICSITech). IEEE (2016)
6. Haviluddin, et al.: A performance comparison of statistical and machine learning techniques in learning time series data. *Adv. Sci. Lett.* **21**(10), 3037–3041 (2015)
7. Purnawansyah, Haviluddin: K-means clustering implementation in network traffic activities. In: 2016 International Conference on Computational Intelligence and Cybernetics, Makassar, Indonesia. IEEE (2016)
8. Pandey, M., Taruna, S.: Towards the integration of multiple classifier pertaining to the student's performance prediction. *Perspect. Sci.* **2016**(8), 364–366 (2016)
9. Yunianta, A., et al.: Data mapping process to handle semantic data problem on student grading system. *Int. J. Adv. Intell. Inform. (IJAIN)* **2**(3), 157–166 (2017)
10. Dangi, A., Srivastava, S.: Educational data classification using selective Naïve Bayes for quota categorization. In: 2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE). IEEE (2014)
11. Guarín, C.E.L., Guzmán, E.L., González, F.A.: A model to predict low academic performance at a specific enrollment using data mining. *IEEE Rev. Iberoam. Tecnol. Aprendiz.* **10**(3), 119–125 (2015)
12. Amaya, Y., Barrientos, E., Heredia, D.: Student dropout predictive model using data mining techniques. *IEEE Lat. Am. Trans.* **13**(9), 3127–3134 (2015)
13. Kotsiantis, S., Patriarchas, K., Xenos, M.: A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowl. Based Syst.* **23**, 529–535 (2010)
14. Jiawei, H., Kamber, M.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann Ltd., San Francisco (2001)
15. Gorunescu, F.: *Data Mining*. Intelligent Systems Reference Library, vol. 12. Springer, Craiova (2011)