ISSN 2723-6471

Data Mining Implementation with Algorithm C4.5 for Predicting Graduation Rate College student

Jeffri Prayitno Bangkit Saputra[1,*], Retno Waluyo[2]

Department Information System, Universitas Amikom Purwokerto, Indonesia
prayitnojeffry@amikompurwokerto.ac.id[1,*]; waluyo@amikompurwokerto.ac.id[2]
* corresponding author

(Received July 3, 2021 Revised August 11, 2021 Accepted August 28, 2021, Available online September 29, 2021)

Abstract

Academic evaluation and graduation of students are critical components of an academic information system's (AIS) effectiveness since they allow for the measurement of student learning progress. Additionally, the assessment stating whether the student passed or failed would benefit both the student and teacher by acting as a reference point for future performance suggestions and evaluations. Using Decision Tree C4.5, a comprehensive analysis of the student academic evaluation approach was conducted. Age, gender, public or private high school status, high school department, organization activity, age at high school admission, progress GPA (pGPA), and total GPA (tGPA) were all documented and evaluated from semester 1–4 utilizing three times the graduation criterion periods. The article's scope is confined to undergraduate programs. An accuracy algorithm (AC) with a performance accuracy of 79.60 percent, a true positive rate (TP) of 77.70 percent, and 91 percent quality training data achieved the highest performance accuracy value.

Keywords: Data Mining, C4.5, Education, Graduation Prediction

1. Introduction

A precise and continuing technique for determining a student's academic attainment level in accordance with educational laws is learning process evaluation. To measure a student's grasp of a course, quizzes, exams, practicums, and other activities addressing cognitive, emotional, and psychometrics capacity are used [1–3]. Additionally, in student academic evaluations, progress reports, which include both the progress GPA (pGPA) and the total GPA, are often used (tGPA). The grades from the course subjects are utilized to calculate the pGPA and tGPA. As a result, identifying the students is critical in determining which factors have the greatest influence. As a result, a data mining model can be utilized for classification [4, 6], prediction [5, 6], clustering [7], and other tasks [8].

According to the International Educational Data Mining Society, Educational Data Mining (EDM) is a data mining application that is employed in educational contexts [9]. To put it another way, education, information science, and computer science all fall under the umbrella of EDM [2, 10]. Numerous data mining technologies, including statistical and intelligent computer approaches, are commonly utilized to fulfill academic evaluation tasks for students. Academic attrition (loss of academic standing) was quantified at the Universidad Nacional de Colombia [11] using two classification approaches: Naive Bayes and Decision Tree Classifier. Between 2007 and 2012, this study investigated academic records from two programs, Agricultural (AE) and Computer and Systems (CE). The findings indicate that NBC and Decision Tree models can be used to forecast academic standing deterioration. At the University Simón Bolvar, [12] used the C4.5 and ID3 algorithms to predict and explain student dropout. WEKA was utilized in this experiment to process the data. According to the study's conclusions, these algorithms can be utilized in place of a model. [13] investigated Naive Bayes, 1-NN, and WINNOW algorithms for predicting student achievement. This strategy was shown to be the most successful for designing a software support tool. The purpose of this study is to explore the Tree C4.5 algorithm and determine how it might be utilized to assess academic learning performance of students. As a result, all pupils may be able to boost their learning efficiency and speed. The purpose of this example study is to assist students in making more informed academic choices.

2. Literature Review

A decision tree is a hierarchical data structure composed of nodes (root, branch, and leaf) and edges (connections between nodes). The decision tree algorithm [13, 14] includes the Tree C4.5 method. It is a method of learning that is supervised. In the 1990s, Quinlan developed Tree C4.5, which is based on the Iterative Dichotomiser (ID3) technique [4], which is efficient, powerful, and widely used. There are two aspects to the C4.5 approach: decision tree preparation and rule development (structure and design). The information gain with the highest attribute is picked following the entropy calculation.

For generating a decision tree, the Tree C4.5 technique involves four parts. As a starting point, select attribute as the root attribute. In the second step, make a branch for each value. Make a branch for the dataset in the end. Repeat steps two and three until all of the classes have the same value. The following is the entropy formula, where S stands for entropy and p refers for the output's class proportion.

$$\text{Entropy(S)} = \Sigma_{i=1}^{n} - p_i * Log_2 Pi \quad (1)$$

Additionally, the root characteristic has the highest gain value. Equation 2 illustrates the gain formula,where S is a set of cases, A signifies a case attribute, |Si| denotes the number of cases associated with I, and |S| denotes the set's total number of cases.

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

3. Methodology

Personal information, academic portfolios, course duration, and student engagement in the organization's activities comprise the student dataset used in this study. Between 2014 and 2017, data from kaggle was collected using the academic information system (AIS) (279 samples data). As indicated in Fig. 1, cleaning, integration, and transformation will be utilized to normalize all datasets prior to training. The first step was to clean the data; a total of 459 data points were collected, and 180 of them were cleaned because certain attribute values were missing. Second, with a total attribute value of 15, the integration and transformation approach was used to reduce and integrate unconditional characteristics. Finally, 11 attributes were employed to decrease and integrate unconditional attributes. The confusion matrix (CM) with the true positive rate (TP) is also used to evaluate the Tree C4.5 method's performance. The Jupyter program was then used for the computing and modeling process (Table 1).

Table. 1. Following integration and transformation (data attribute)

Variables	Measure	Value
Gender	Number	M : Male F : Female
Age	Order	Student age
Birth place	Number	Town, Village
Education status	Number	State, Private
Education program	Number	Science, Non-science
GPA 1	Order	1 (GPA < 1.5)
GPA 2	Order	2 (1.5 < GPA 2.5)

GPA 3	Order	3 (2.5 < GPA 3.5)
GPA 4	Order	4 (3.5 < GPA 4.0)
Organization	Number	Activist, Non-activist
Graduation time	Order	Delay (>4,6 years) On-time (4–4,6 years) Fast (<4 years)

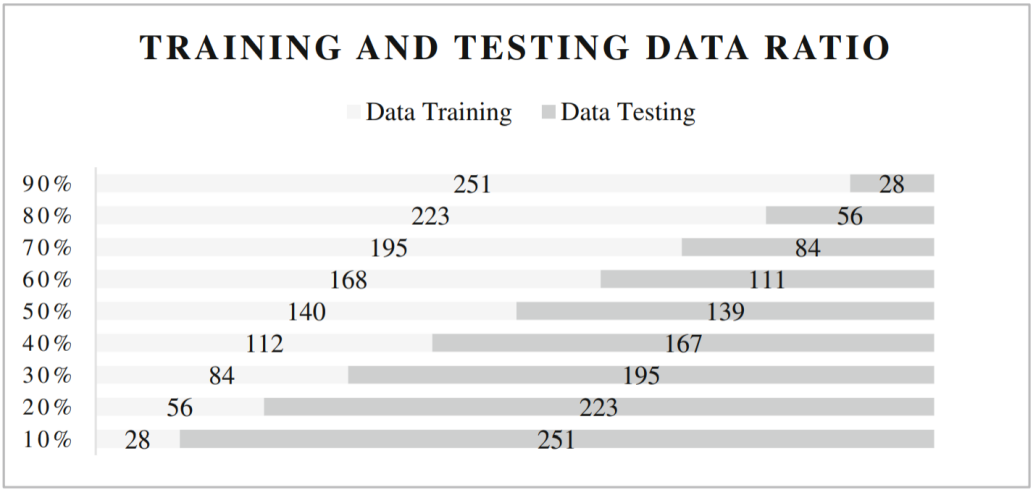

Figure. 1. Training data distribution

The confusion matrix (CM) and true positive rate were used to evaluate the Tree C4.5 model in this study (TP). As indicated in Tables 2 and 3, CM is a matrix of predictions that will be compared to the input's original class. In other words, the matrix incorporates both real-world data and categorization predictions [15]. The following equation is used to determine accuracy (AC): The total training data set size is N, and the precise number of forecasts for the "Fast-Time" graduation is An; the precise number of forecasts for the "On-Time" graduation is B; and the precise number of forecasts for the "Delay-Time" graduation is C.

$$AC = \frac{a+b+c}{N} \quad (3)$$

Where a stands for the correct number of negative forecasts, b for the erroneous number of negative forecasts, c for the inaccurate number of positive forecasts, and d for the proper number of negative forecasts.

Table. 2. Matrix of Class 2 Confusion

		Predicted	
		-	+
Original	-	a	b
	+	c	d

Table. 3. The confusion matrix for the Tree C4.5 algorithm

Confusion matrix	Graduation		
	Fast	On	Delay
Fast-time	113	28	7
On-time	21	12	47
Delay-time	5	13	8

In this study, the total course subject was used to evaluate students' academic performance in Years 1, 2, and 3. In other words, by the end of this assessment, the student will have progressed to the next level. Table 4 shows the student evaluation period.

Table. 4. Term for student evaluation

Overview (Year)	Degree	
I	Subject of the entire course	24
	GPA	2,00
II	Subject of the entire course	48
	GPA	2,00
III	Subject of the entire course	72
	GPA	2,00

Meanwhile, the Tree C4.5 model's observed training data is subjected to a true positive rate (TP). The following is the TP formula.

$$TP = \frac{\Sigma_{i=1}^{3} \frac{a_i}{n_i}}{3} \quad (4)$$

Where TP stands for the percentage of correct predictions, ai stands for the precise number of predictions for the "fast, on, delay" graduation time, and in the total amount of training data stands for the total amount of training data for the "fast, on, delay" graduation period. The Receiver Operating Characteristic (ROC) analysis was omitted from this study's model evaluation because ROC analysis is critical for setting CM and TP thresholds. Additionally, Figure 1 illustrates the phases of analysis in this work using the Tree C4.5 approach.

4. Implementation and Result

This section describes how to examine student academic evaluation variables with the use of Tree C4.5 models. A dataset for nine training and testing classes was constructed using predefined procedures. From semester 1 to semester 4, this experiment assessed students' academic performance by age, birthplace, gender, high school status (public or private), high school department, organization participation, age at the start of high school, and pGPA and

tGPA. Furthermore, between 10% and 90% of CM has been looked into as a possible source of high-quality training data. Meanwhile, the CM as a Tree C4.5 technique was tested utilizing three time criteria (quick, on, and wait periods), as shown in Table 5, in order to achieve the best accuracy.

Table. 5. Training data

Confusion matrix (%)	Data on training		
	Fast-time	On-time	Delay-time
10	103	23	9
	24	16	44
	22	15	20
20	77	14	7
	31	17	25
	14	10	25
30	71	14	1
	24	16	5
	16	11	41
40	64	7	4
	27	18	8
	4	6	41
50	46	4	3
	15	7	20
	16	11	14
60	41	9	8
	12	8	1
	6	4	20
70	37	7	4
	8	5	3
	4	1	13
80	25	7	4
	2	4	1

ISSN 2723-6471

	5	2	13
90	16	4	3
	2	5	1
	3	2	7

According to the experiment's results, the Tree C4.5 algorithm's best accuracy value is 79.60 percent AC with 77.70 percent TP on 91 percent quality training data. As demonstrated in Table 6 and Fig. 2, the Tree C4.5 technique achieves the highest accuracy when 91 percent of the training data is used.

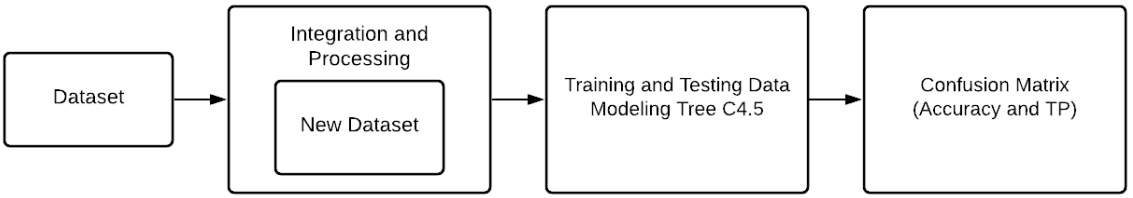

Fig. 2. Tree C4.5 algorithm stages of analysis

Table. 6. Tree C4.5 algorithm's confusion matrix and true positive rate

		Algorithm Accuracy (AC)	True Positive rate (TP)
Ratio of training data	10%	52.89%	45.94%
	20%	53.83%	51.37%
	30%	65.23%	59.77%
	40%	65.29%	62.50%
	50%	49.74%	46.95%
	60%	63.86%	59.97%
	70%	65.49%	61.63%
	80%	71.33%	78.75%
	90%	78.68%	76.73%

Using 251 training data, the best entropy and gain performance was determined. GPA semester 4 has been designated as the beginning point (root). The results for entropy and gain derived from 90% training data are detailed in Table 7. According to Table 7, the maximum gain value obtained by doing a manual calculation on the GPA semester 4 variables is 1.019 at the starting node. In other words, the starting node is associated with modeling (Fig. 3).

ISSN 2723-6471

Table. 7. Result after 90% training data, entropy and gain values are calculated

Root	Total graduation	Fast-time	On-time	Delay-time	Entropy	Gain
	251	140	48	63	1,43	
Sex						0,74
M	192	95	39	37	0,59	
F	67	53	20	17	0,37	
Age						0,86
16	0	2	3	1	1	
17	37	23	6	12	0,53	
18	136	73	36	48	0,47	
19	45	16	6	33	0,66	
20	6	2	0	2	0,47	
...	...	...	...	...	...	
23	3	1	1	2	2	
Place of birth						0,95
Town	117	49	23	36	0,59	
Village	154	80	46	27	0,51	
School status						0,96
State	183	121	53	47	0,46	
Private	57	38	7	13	0,51	
School program						0.96
Science	174	131	19	35	0.38	
Non-Science	68	27	39	32	0.68	
Organization						0,94
Activist	104	57	29	36		
Non-Activist	149	73	49	27		
GPA Semester 1						0,97

1	2	3	1	1	2	
2	1	2	3	3	1	
3	180	75	62	45	0,40	
4	91	47	7	9	1,38	
GPA Semester 2						1,11
1	1	3	2	2	1	
2	3	2	2	1	1	
3	256	55	49	71	0,39	
4	97	94	7	3	0,38	
GPA Semester 3						0,89
1	1	3	2	1	3	
2	3	2	2	1	2	
3	183	64	37	51	0,39	
4	88	67	20	4	0,51	
GPA Semester 4						1,12
1	1	02	3	3	1	
2	1	10	1	2	1	
3	145	53	33	50	0,47	
4	126	98	26	15	0,51	

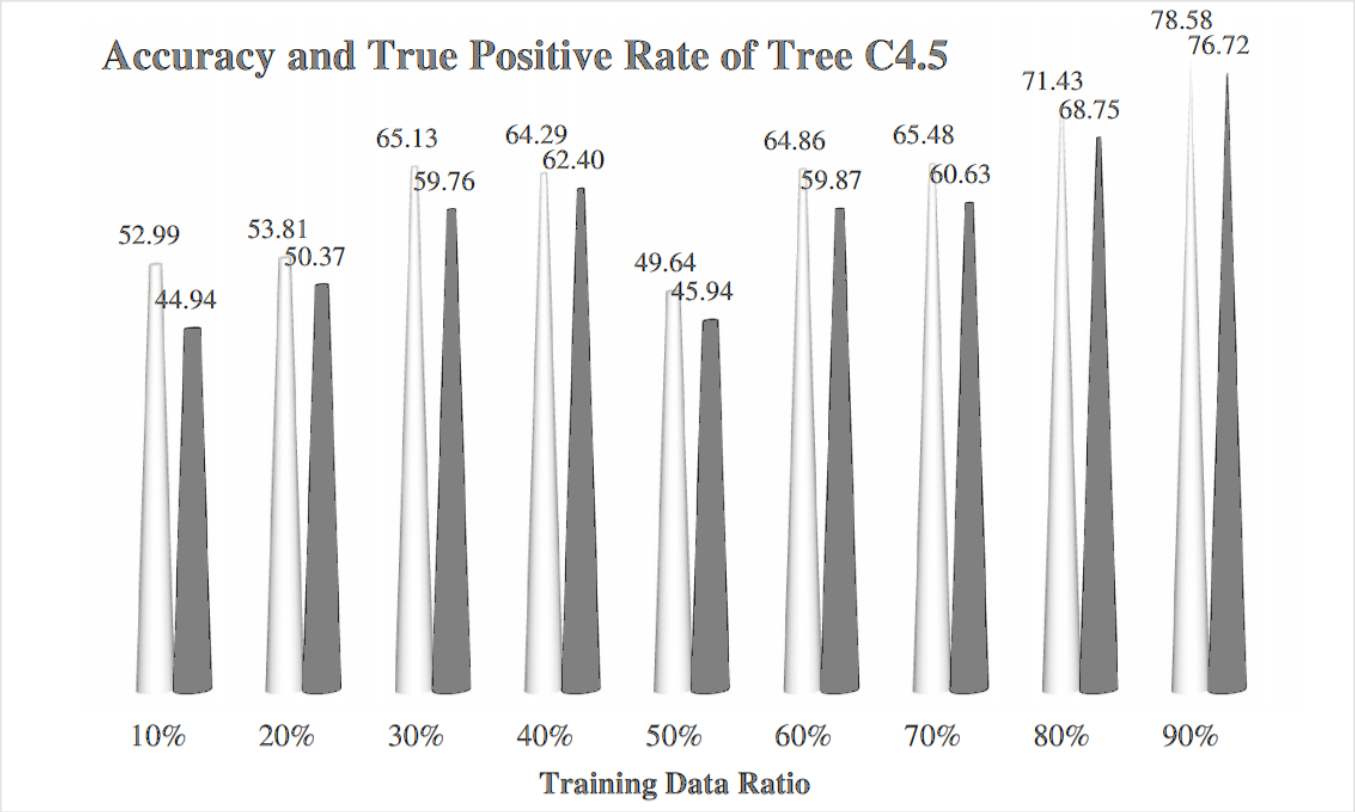

Figure. 3. The confusion matrix for the Tree C4.5 algorithm and the true positive rate graph

5. Conclusion

The intellectual achievement of the children in this study was determined using the Tree C4.5 approach. According to the experiment's findings, various variables such as student organization participation (activist and non-activist), birthplace, and age all have an effect on student academic success. In this study, it was discovered that the Tree C4.5 algorithm was more accurate at evaluating students' academic performance. In other words, the Tree C4.5 method might be used in place of the traditional paradigm for assessing student academic development. To improve accuracy, it is proposed that future initiatives incorporate the Naive Bayes Classifier (NBC), K-Means Clustering, and Support Vector Machine (SVM) algorithms.

References

[1] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," Heliyon, vol. 5, no. 2, p. e01250, 2019, doi: 10.1016/j.heliyon.2019.e01250.

[2] C. K. Lo, H. C. Chen, P. Y. Lee, M. C. Ku, L. Ogiela, and C. H. Chuang, "Smart dynamic resource allocation model for patient-driven mobile medical information system using C4.5 algorithm," J. Electron. Sci. Technol., vol. 17, no. 3, pp. 231–241, 2019, doi: 10.11989/JEST.1674-862X.71018117.

[3] H. Bin Wang and Y. J. Gao, "Research on C4.5 algorithm improvement strategy based on MapReduce," Procedia Comput. Sci., vol. 183, pp. 160–165, 2021, doi: 10.1016/j.procs.2021.02.045.

[4] L. G. Plata, C. G. Ramos, M. L. Silva Oliveira, and L. F. Silva Oliveira, "Release kinetics of multi-nutrients from volcanic rock mining by-products: Evidences for their use as a soil remineralizer," J. Clean. Prod., vol. 279, p. 123668, 2021, doi: 10.1016/j.jclepro.2020.123668.

[5] G. Lesinski and S. Corns, "Multi-objective evolutionary neural network to predict graduation success at the United States Military Academy," Procedia Comput. Sci., vol. 140, pp. 196–205, 2018, doi: 10.1016/j.procs.2018.10.329.
[6] C. Grac, X. Dolques, A. Braud, M. Trémolières, J. N. Beisel, and F. Le Ber, "Mining the sequential patterns of water quality preceding the biological status of waterbodies," Ecol. Indic., vol. 130, 2021, doi: 10.1016/j.ecolind.2021.108070.
[7] J. de J. Costa, F. Bernardini, D. Artigas, and J. Viterbo, "Mining direct acyclic graphs to find frequent substructures — An experimental analysis on educational data," Inf. Sci. (Ny)., vol. 482, pp. 266–278, 2019, doi: 10.1016/j.ins.2019.01.032.
[8] L. Bonilla Mejía, "Mining and human capital accumulation: Evidence from the Colombian gold rush," J. Dev. Econ., vol. 145, no. July 2019, 2020, doi: 10.1016/j.jdeveco.2020.102471.
[9] K. R. Pradeep and N. C. Naveen, "Lung Cancer Survivability Prediction based on Performance Using Classification Techniques of Support Vector Machines, C4.5 and Naive Bayes Algorithms for Healthcare Analytics," Procedia Comput. Sci., vol. 132, pp. 412–420, 2018, doi: 10.1016/j.procs.2018.05.162.
[10] M. Sharma, S. Joshi, and K. Govindan, "Issues and solutions of electronic waste urban mining for circular economy transition: An Indian context," J. Environ. Manage., vol. 290, no. October 2020, p. 112373, 2021, doi: 10.1016/j.jenvman.2021.112373.
[11] V. Fernandez, "Innovation in the global mining sector and the case of Chile," Resour. Policy, vol. 68, no. January, p. 101690, 2020, doi: 10.1016/j.resourpol.2020.101690.
[12] A. Hira and J. Busumtwi-Sam, "Improving mining community benefits through better monitoring and evaluation," Resour. Policy, vol. 73, no. December 2020, p. 102138, 2021, doi: 10.1016/j.resourpol.2021.102138.
[13] R. Benkercha and S. Moulahoum, "Fault detection and diagnosis based on C4.5 decision tree algorithm for grid connected PV system," Sol. Energy, vol. 173, no. July, pp. 610–634, 2018, doi: 10.1016/j.solener.2018.07.089.
[14] B. Sen and E. Ucar, "Evaluating the achievements of computer engineering department of distance education students with data mining methods," Procedia Technol., vol. 1, pp. 262–267, 2012, doi: 10.1016/j.protcy.2012.02.053.
[15] P. Dabhade, R. Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar, "Educational data mining for predicting students' academic performance using machine learning algorithms," Mater. Today Proc., no. xxxx, 2021, doi: 10.1016/j.matpr.2021.05.646.
[16] R. Campagni, D. Merlini, R. Sprugnoli, and M. C. Verri, "Data mining models for student careers," Expert Syst. Appl., vol. 42, no. 13, pp. 5508–5521, 2015, doi: 10.1016/j.eswa.2015.02.052.
[17] X. Meng, P. Zhang, Y. Xu, and H. Xie, "Construction of decision tree based on C4.5 algorithm for online voltage stability assessment," Int. J. Electr. Power Energy Syst., vol. 118, no. December 2019, p. 105793, 2020, doi: 10.1016/j.ijepes.2019.105793.
[18] G. Lesinski, S. Corns, and C. Dagli, "Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy," Procedia Comput. Sci., vol. 95, pp. 375–382, 2016, doi: 10.1016/j.procs.2016.09.348.
[19] S. J. Lee, Z. Xu, T. Li, and Y. Yang, "A novel bagging C4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making," J. Biomed. Inform., vol. 78, pp. 144–155, 2018, doi: 10.1016/j.jbi.2017.11.005.