Application of the C5.0 Algorithm for Student Graduation Integrated with Academic Information Systems

Abdi Dharma, Sonia N. Lase, Yenny, Yonata Laia, Mardi Turnip

***Abstract*— A series of processes that must be passed by each student, including completing a predetermined number of courses, carrying out practical fieldwork, proposal seminars, and final project seminars, all of which are carried out towards graduation. However, not all students can graduate on time according to the study period taken, so sometimes it causes a pile of students who do not graduate according to their graduation period. These students' graduation time cannot always be detected early, leading to delays in graduates. To overcome this, a technique of the C5.0 algorithm forms a classification tree is proposed to predict student graduation. The C5.0 algorithm has a pruning technique that can trim the decision tree to be simpler, shorter, and boosting is used to measure the level of percentage accuracy. From the test results by applying the C5.0 algorithm using 300 students data from the class of 2019 (with 210 for training and 90 for data test), the percentage of recall, precision, accuracy, and error values were about 94.28%, 80.48%, 88 .89%, and 11.11%, respectively.**

***Index Terms*— C5.0 Algorithm, Pruning, Boosting, Graduation.**

I. INTRODUCTION

IN education, the Bachelor Program (SI) has a load of approximately 144 - 160 cumulative credits, which must be completed within eight semesters [1]. Each student must pass the series of stages of the academic process until graduation by completing education within a predetermined period, otherwise, the student is declared Drop-Out [3]. Every year the quota of students accepted is increasing, but not all of them can graduate on time according to the study period taken so sometimes it results in a buildup of some students who do not graduate according to their graduation period [23].

Late graduation is likely to cause an extra workload for faculty members because they must supervise more students at a time. Therefore, the university usually has a strategy to improve and maintain the on-time graduation rate [5][6]. The time of graduation of these students cannot always be detected early so that it can lead to delays in graduates. To handle this, a technique to predict student graduation is needed [7].

The technique used in data mining and classification method will be used to predict student graduation [22]. Data mining can extract educational data to improve the quality of the education process [8] and identify strategies for improving students' performance [9]. There are two aspects of students' performance: academic achievement and learning progressions, and this can be used to predict their success in finishing the study on time [10][11] or to design interventions to prevent failure [12]. Data mining has three main functions, which are clustering data [13][14], classifying data [19][20], and identifying association rules patterns [15]. The current student performance prediction study shows that student performance prediction is challenging due to educational data variants [16] [17]. One of the data mining algorithms that can be used is the C5.0 algorithm [19], the algorithm is the latest version of the ID3 algorithm development and the C4.5 algorithm was discovered by John Ross Quinlan [18]. Making a decision tree in C5.0 is similar to C4.5 [20], the difference is that in C5.0 there is a boost. Boosting is a learning algorithm to reduce bias [21] and variance. Boosting is a collection of algorithms that can turn weak students into strong students [24] [25].

Several studies on the C5.0 algorithm have been carried out, where the results obtained are that the C5.0 algorithm is better at classifying, such as research on the comparison of the performance of the Decision Tree C5.0, CART, and CHAID algorithms, where it was found that the C5.0 algorithm is better than other algorithms and accurate in doing credit scoring [18]. Research on the comparison of the C5.0 Algorithm and CART Classification. The results obtained that the accuracy of the C5.0 algorithm is better than the CART algorithm [19]. This study tries to apply the C5.0 algorithm, which has been done before but with different variables and is applied to student assessment data. The C5.0 algorithm has several techniques that are simpler than other algorithms because the C5.0 algorithm has a decision tree pruning and can increase the accuracy of a prediction [20]. The results of this algorithm will be implemented into the student academic information system. If this student's graduation can be classified early, the study program can identify the characteristics of students who have the potential to graduate on time or not and take persuasive

Manuscript received October 9, 2020.

Abdi Dharma, Faculty of Technology and Computer Science, Universitas Prima Indonesia Medan, Indonesia (e-mail: abdidharma@unprimdn.ac.id).

Sonia Novel Lase, Faculty of Technology and Computer Science, Universitas Prima Indonesia Medan, Indonesia (e-mail: sonianovel140620@gmail.com).

Yenny, Faculty of Technology and Computer Science, Universitas Prima Indonesia Medan, Indonesia (e-mail: tanyenny012@gmail.com).

Yonata Laia, Faculty of Technology and Computer Science, Universitas Prima Indonesia Medan, Indonesia (e-mail: yonata@gmail.com).

Mardi Turnip, Faculty of Technology and Computer Science, Universitas Prima Indonesia Medan, Indonesia (e-mail: marditurnip@unprimdn.ac.id).

steps.

II. Procedure for Paper Submission

The research method can be seen in Fig. 1, divided into six stages: data collection, data preprocessing, data sharing, classification with the C5.0 algorithm, and calculation accuracy testing. The data preprocessing stage is an important step used to convert raw data into a format that allows for applying data mining techniques, using the C5.0 algorithm, and improving data quality. This study uses student graduation data from the faculty of technology and computer science of a private university with the name Universitas Prima Indonesia with Grade B in 2019.

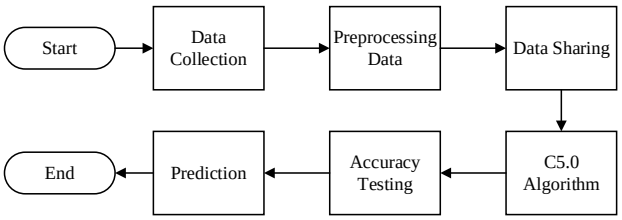

Fig. 1. Stages of the Research Process

A. Preprocessing Data

Preprocessing is an important process used to convert raw data into a format that is used to apply data mining techniques, use the C5.0 algorithm and improve data quality. [18]. The result of this stage is to get data that is ready to use. In the classification process, there are several stages in preprocessing the data, to retrieve the relevant data used as research and convert the numerical data into categorical. The dataset contains six attributes, namely Student Status, GPA, Cumulative Credits, and Attendance, which will be converted into categorical data. In table 1, the dataset attribute explains the attributes that will be used to simplify the calculation of the C5.0 algorithm.

TABLE I
Dataset Attribute

ATTRIBUTE	CATEGORY
Student Status	- Active - Not Active
Cumulative Credit	- Less (0-42) - Medium (43-87) - Good(88-134) - Very Good(135-160)
GPA	- Low (0-2,4) - Medium(2,5-3) - High (3,1-3,5) - Very High (3,6-4)
Presence	- Less (>=4) - Medium(2-3) - Good(1) - Very Good(0)
Output	- On Time - Not On Time

All attributes are divided into several categories, one of which is the Cumulative Credit attribute taken from each universe to get low, medium, high, and very high categories. Categorizing at the initial stage of preprocessing will make it easier to process data to produce a decision tree that can predict the graduation of 2019 students.

B. Data Sharing

In classification, the data is divided into two parts, namely training data and test data, where training data is used to build a decision tree. The training data that already has the results are then used to calculate the error rate at the pruning stage and the weight value at the boosting stage. The test data is used to calculate the accuracy of the decision tree. The test data also has a result value that is used to test the last classification tree formed and selected during the C50 algorithm process from a total of 300 data. About 70% for training data and 30% for test data so that it can produce a low error ratio.

C. C5.0 Algorithm

The initial stage of this algorithm is to calculate the information obtained from SKS, GPA, and Status attributes. Equations 1 and 2 are the process flow in the C5.0 algorithm. Training data is the input in the calculation of the C5.0 algorithm where there are 300 data, 70% of the data is taken for training data, and 30% for test data. This algorithm starts with all the data used as the root of the decision tree while the selected attribute becomes the divisor for the sample [18]. Equation (1) is used to calculate the entropy.

$$Entropi\ (S1, S2, \dots, Sn) = \Sigma\ Pi * log2(Pi) ni=1 \qquad (1)$$

with description S is Case Set; n is Number of samples; Pi is class proportion. The gain value is obtained using equation 2.

$$\text{Gain(A)} = \text{Entropi(S)} - \sum_{i-1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \qquad (2)$$

with S is Case Set; A is Attribute; n is Number of Samples; |Si| is Number of Cases on Partition I; |S| is Number of Cases in S.

1) Pruning

In this pruning process, it will receive input in the form of attributes that become the parent node and each will become a branch node. Error rates for parent nodes and branch nodes are calculated based on these values. If the branch node error value is greater than the parent node, pruning is performed and subtree formation stops.

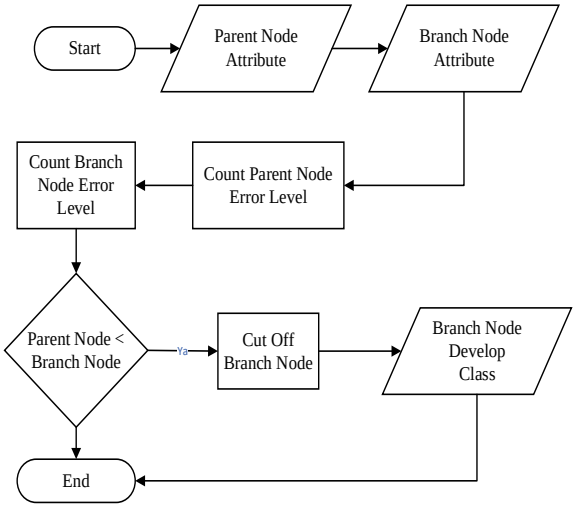

Fig. 2. Pruning Process Flowchart

In Fig. 2, the output in the form of a branch node has an output class if there is pruning. If this process has not produced an output class, it will continue calculating the result information for each remaining attribute. Pruning is a process done to cut or remove some branches that are not needed. Pruning methods are divided into pre-pruning and post-trimming. The pruning technique generally uses a statistical approach. After pruning, the decision tree will have fewer results, making it easier to understand. Below is the formula for calculating the estimated error:

$$e = \frac{f+\frac{z^2}{2N}+z\sqrt{\frac{f}{N}+\frac{f^2}{N}+\frac{z^2}{2N^2}}}{1+\frac{z^2}{N}} \quad (3)$$

with f is The quotient of the number of incorrectly classified data by the number of sample data; N is Number of sample data with z is a constant value about 0.69.

2) Boosting

The process of boosting receives the form of a decision tree classification, at this stage, resulting in new training data being re-calculated to form a new decision tree again. Before performing the first iteration boosting for each data sample will be initialized with the same weight, which is 1 / N for any weight of samples to be updated according to whether or not the data sample classified. The weight that continues to be updated in each iteration will continue to increase in value. The process of increasing the weight value is one of the goals at the boosting stage.

The first calculation by giving the same weight to each data, with a calculation like this equation 4.

$$w = \frac{1}{N} \quad (4)$$

In equation 4 w is the weight, N is the amount of data used to form the classification model. Then determine the midpoint between the total weight for the misclassified sample and half of the total weight, using the following formula:

$$Midpoint = \frac{1}{2}\left[\frac{1}{2}(s_- + s_+) - s_+\right] = \frac{1}{4}(s_- + s_+) \quad (5)$$

Then, the data that can be classified correctly will experience a change in the weight value through the following equation.

$$W_k = W_{k-1} \times \frac{s_+ - midpoint}{s_+} \quad (6)$$

While the value of the weight will change the data that is classified incorrectly through equation 7

$$W_k = W_{k-1} \times \frac{midpoint}{N_-} \quad (7)$$

With, midpoint = mean value of wk = data weight on the kth boosting iteration; wk-1 = data weight on the k-1 boosting iteration; N- = number of misclassified data; S+ = the number of data weights that are correctly classified; S- = number of data weights that are misclassified.

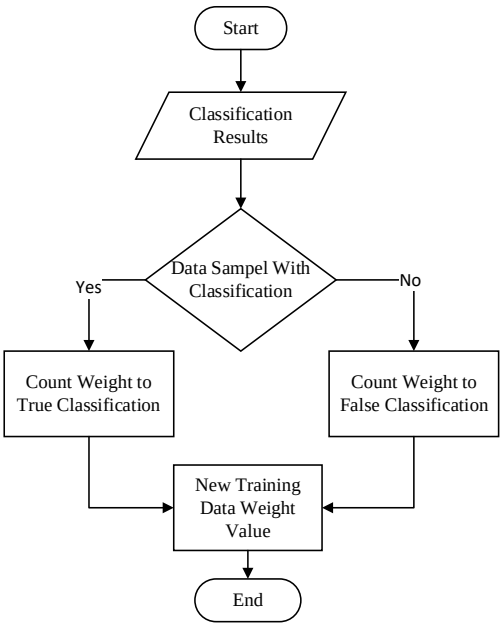

Fig. 2. Boosting Process Flowchart

III. RESULTS AND DISCUSSION

In this study, the analysis uses the C5.0 Algorithm system. In the calculation process, student data is needed, the data contained in Table 2 is the 2019 student dataset table that will be used. In the table, there are the results of the calculation of entropy and information gain which are calculated using the equations in equation (1) and equation (2).

TABLE II
Entropy and Gain Calculation

	Amount	**On-Time**	**Not On Time**	**Entropy**	**Gain**
Total	210	169	41	0.71230	
Status					**0.618511**
A	165	133	32	0.04638	
NA	45	36	9	0.26761	
Credit					**0.62158**
VG	69	54	15	0.10929	
G	103	86	17	0.07886	

M	26	0	26	0.23517	
L	5	0	5	0,72192	
Presence					**0.60433**
VH	148	122	26	0.05843	
H	52	37	15	0.13709	
M	7	7	0	0.59168	
L	3	3	0	0.91833	
GPA					**0.59720**
VH	17	15	3	0.32275	
H	110	101	23	0.07470	
M	57	50	15	0.12742	
L	26	3	0	0.62346	

After getting the gain value in Table 2, the next is to determine the attribute's parent node with the highest gain value. As shown in Table 3 the highest gain value is the Cumulative Credit attribute with a value of 0.62158. For attribute values with sub-attributes M and L, it has been said that the result is incorrect, and the example of the parent node is shown in Fig. 3.

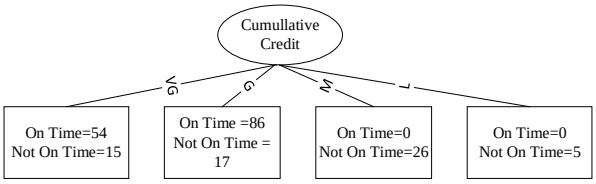

Fig. 3. Parent Node Example

A. *Pruning*

The way pruning works is by first calculating the information gain value to find out the value of the parent node and branch node, after the parent node and branch node are known then, the error value is calculated, if the branch node error value is smaller than the parent node, the parent node forms a subtree again. If the branch node error value exceeds the parent node, pruning is performed, and subtree formation stops. When the error value at the branch node is greater than the error value at the parent node, After knowing the gain value, then calculating the pruning process where this pruning process calculates the level of error in the VG and G sub-attributes only, because the VG and G sub-attributes have produced incorrect output. From the results of pruning calculations using equation 3, the pruning results obtained from the overall data are 0.545 with a total data of 220 students where 41 students are not on time. From the VG sub-attribute with a total of 69 students, 15 students were not on time, so the pruning result was 0.061 with information that the data was not trimmed. As for the G attribute with a total data of 103 students, 17 students enter the output, not on time, the pruning result is 0.894 with the data description being trimmed. For the next stage, namely the formation of a decision tree and the final classification rule, where this stage has gone through the pruning process, and the tree becomes smaller than it was in the beginning.

B. *Boosting*

The boosting stage displays the correct data classification and misclassified data, where the correct data classification is the correct or correct data in the previous stage, namely the pruning stage, while the misclassified data is true or false data at the pruning stage. This time the weighting of each data is given the same weight value in every existing data. After assigning a weight to each data, the true data weight is 0.91152 and the false data weight is 0.053816. Next is to calculate the correct data classification and incorrect data classification. It is known that the data whose attribute results are the same as the pruning results, which are declared correct, has a more recent weight of 0.007898, while data whose attribute results are different from the pruning results, which are declared misclassified, has the latest weight of 0.05643. We also get the midpoint results is 0,31234.

C. *Testing*

This research was conducted to test the system using test data about 30% of the dataset, namely a total of 60 test data that will be used in this study. The 60 data already have class or attribute values that will be compared with the results of the output program or decision tree. The test in this study uses the confusion matrix method, which will later get the values of precision, recall, accuracy, and error. In the results of the confusion matrix calculation, were found that 33 students were True Positive (TP), eight students were False Positive (FP), two students were False Negative (FN), and 47 students were True Negative (TN). From the calculation of the confusion matrix we got the recall value is 94.28%, the precision value is 80.48%, the accuracy value is 88.89%, and the error value is 11.11%.

IV. CONCLUSION

A. *Conclusion*

Based on the results of analysis and testing of student graduation predictions using the C5.0 algorithm with five attributes (not all attributes are used as decision tree nodes), at this stage, the decision tree pruning process becomes simpler than before the pruning process. The results of testing the C5.0 algorithm found that the recall value is 94.28%, the precision value is 80.48%, the accuracy value is 88.89% and the error value is 11.11%.

B. *Suggestion*

From the results of the research on student graduation predictions using the C5.0 algorithm, it is suggested that the dataset used can be further developed, from the amount of data or attributes used and the web-based student graduation prediction system using the C5.0 algorithm can be developed more dynamically and in desktop form and mobile.

ACKNOWLEDGMENT

With all humility, we say Thank You to the Faculty of Technology and Computer Science, Universitas Prima Indonesia as a case study material in this research. Thanks also to the editors and reviewers for their useful comments and suggestions.

REFERENCES

[1] H. Yuliansyah, Hafsah, I. Arfiani, and R. Umar, "Discovering Meaningful Pattern of Undergraduate Students Data using Association Rules Mining," in 2019 Ahmad Dahlan International Conference Series on Engineering and Science (ADICS-ES 2019), 2019, pp. 13-17.

[2] Mardi Turnip. "An application of modified filter algorithm fetal electrocardiogram signals with various subjects," *International Journal of Artificial Intelligence*, vol.18, Mar 2020.

[3] A. Alhassan, B. Zaffar, and A. Mueen, "Predict students"academic performance based on their assessment grades and online activity data," *International Journal of Advanced and Computer Science and Applications*, vol. 11, no. 4, 2020.

[4] C. Aina and G. Casalone, "Early labor market outcomes of university graduates: Does time to degree matter?," *Socioecon. Plann. Sci.*, p. 100822, Mar 2020.

[5] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," *Comput. Human Behav.*, vol. 98, pp. 166–173, Sep 2019.

[6] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Comput. Educ.*, vol. 113, pp. 177–194, 2017.

[7] R. Campagni, D. Merlini, R. Sprugnoli, and M. C. Verri, "Data mining models for student careers," *Expert Syst. Appl.*, 2015.

[8] 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, Feb 2019.

[9] S. Winiarti, H. Yuliansyah, and A. A. Purnama, "Identification of Toddlers" Nutritional Status using Data Mining Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 1, pp. 164–169, 2018.

[10] D. Chi, "Research on the Application of K-Means Clustering Algorithm in Student Achievement," in 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021, pp. 435–438.

[11] D. Kumalasari, A. B. W. Putra, and A. F. O. Gaffar, "Speech classification using combination virtual center of gravity and k-means clustering based on audio feature extraction," *J. Inform.*, vol. 14, no. 2, p. 85, May 2020.

[12] A. Namoun and A. Alshanqiti, "Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review," *Appl. Sci.*, vol. 11, no. 1, p. 237, Dec 2020.

[13] H. Yuliansyah and L. Zahrotun, "Designing web-based data mining applications to analyze the association rules tracer study at university using a FOLD-growth method," *Int. J. Adv. Comput. Res.*, vol. 6, no. 27, pp. 215–221, Oct 2016.

[14] A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Educ. Inf. Technol.*, vol. 26, no. 1, pp. 205–240, Jan 2021.

[15] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," in Procedia Computer Science, 2015.

[16] Daniyah Alkhadi et all. "Developing and Implementing Web-based Online University Facilities Reservation System," *International Journal of Applied Engineering Research*, ISSN 0973-4562, Vol.13, 2018, pp 2-4.

[17] Razan Aldaej et all. "Analyzing, Designing, and Implementing a Web-Based Auction online System," *International Journal of Applied Engineering Research*, ISSN 0973-4562, Volume 13, Number 10, 2018.

[18] W. Yusuf Y, "Perbandingan Performansi Algoritma Decision Tree C5.0, CART, Dan CHAID: Kasus Prediksi Status Resiko Kredit Di Bank X", Seminar, vol. 2007, no. Snati, pp. 0–3, 2007.

[19] P. N. Patil, R. Lathi, and V. Chitre, "Comparison of C5 . 0 & CART Classification algorithms using pruning technique," *Int. J. Eng. Res.Technol.*, vol. 1, no. 4, 2012, pp. 1–5.

[20] H. Munowaroh, B. Khusnul dan Y. Kustiyahningsih, "Perbandingan Algoritma ID3 Dan C5.0 Dalam Identifikasi Penjurusan Siswa SMA," *Jurnal Sarjana Teknik Informatika*, Vol. 1 No. 1, Juni 2013, pp. 1 – 13.

[21] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, July 2019, pp. 216–231.

[22] B. Bertaccini, S. Bacci, and A. Petrucci, "A graduates satisfaction index for the evaluation of the university overall quality," *Socioecon. Plann. Sci.*, May 2020, p. 100875.

[23] S. Helal et all., "Predicting academic performance by considering student heterogeneity," *Knowledge-Based Syst.*, vol. 161, Dec 2018, pp. 134146.

[24] K. P. Shaleena and S. Paul, "Data Mining Techniques For Predicting Student Performance," in ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology, 2015.

[25] M. Wibowo, F. Noviyanto, S. Sulaiman, and S. M. Shamsuddin, "Machine Learning Technique For Enhancing Classification Performance In Data Summarization Using Rough Set And Genetic Algorithm," *Int. J. Sci. Technol. Res.*, vol. 8, no. 10, 2019, pp. 1108–1119.

[26] F. Yang and F. W. B. Li, "Study on student performance estimation, student progress analysis, and student potential prediction based on data mining," *Comput. Educ.*, vol. 123, Aug 2018, pp. 97–108.

Mardi Turnip received S.Kom. degree in Informatics Engineering from the National Institute of Technology (ITENAS) Bandung, Indonesia, in 2010 and the M.Kom degree. in the Information System from STMIK LIKMI Bandung, Indonesia, in 2013. He worked at Universitas Prima Indonesia, Indonesia as a lecturer. The research area is information systems and artificial intelligence.