Vol 5-Sep2021
ISSN: 2541-3023

Prediction of Student Graduation Using Decision Tree Method

Harun Mukhtar, Januar Al Amien, Fitra Dewi

Faculty of Computer Science, Universitas Muhammadiyah Riau, Indonesia

E-mail : harunmukhtar@umri.ac.id, januaralamien@umri.ac.id, 160401133@student.umri.ac.id

Abstract. Students are one of the important parameters in the evaluation of study programs. Student attendance, student achievements, and graduation profiles must receive serious attention. More and more students are accepted every year, but not a few are able to complete their studies on time. Predictions for the current graduation rate are needed, considering the number of students experiencing delays is assessed to have reached 92%. Based on the data studied, therewere 25 students in one class, only 2 students who graduated on time were confirmed to be late. To make this prediction, the writer uses Decision Tree. Testing student graduation using validation data as many as 110 lines of data, consisting ofregular students A and regular B. There are 50 students graduating on time and 60 students graduating late. Then the technique of balancing the number of classes on the data is carried out, namely the Synthetic Minority Oversampling Technique (SMOTE), so that the number of validation data increases to 120 rows of data. This test resulted in a score of 96.67%, 96.67% precision, 96.67% recall, and 96.67% f1-score. Based on the ROC curve, the evaluation results that have been achieved in predicting future student graduation using the Decision Tree method, including a very good classification.

Keywords: *Prediction of Student Graduation, Decision Tree, Synthetic Minority Oversampling Technique (SMOTE), Confusion Matrix*

INTRODUCTION

More and more students are accepted every year, but not a few are able to complete their studies on time. The university database stores academic, administrative and student biodata. If the data is dug properly, patterns or knowledge can be known to make decisions [1]. Data on graduating students can provide useful information for universities if they are used to the fullest. One way to utilize data about graduating students is to process it using data mining. With this data mining process, patterns or rules can be found that can be used to produce information such as predictions of student graduation [2].

Predictions for the current graduation rate are needed, considering the percentage of students who experience delays is assessed to reach 92%. Class of 2016 students, there are 25 students in one class, of which only 2 students graduated on time, and 23 students were confirmed to have graduated late. This data is obtained in our class A2 regular class. To make this prediction, the writer uses Decision Tree. Decision Tree uses the principle of rules that are formed based on the decision tree model. The existence of alumni student graduation data, data processing canbe carried out, finding branching flow information patterns that can classify students with timely and late graduation. Decision Tree is a data mining technique that can be used to predict student graduation [3].

Problems in predicting student graduation, because the information has not been found from a number of student data who have graduated. In addition, it is difficult to know which final semester students can be predicted to graduate on time and late in the future. It is very important for the university if it can minimize student graduation failure [4]. The use of decision trees in this study is useful for utilizing student data, and finding certain patterns between several input attributes and a target attribute. Student data used as input attributes include NIM, type of school of origin, area of origin, parental occupation, class, and graduation.

The purpose of this study is to predict student graduation using the Decision Tree algorithm, with the acquisition of the maximum level of prediction accuracy, find a decision tree pattern that matches the data phenomenon, the accuracy and accuracy of grouping graduation on time and late. This study took a sample of 213 alumni students for the 2012-2017 graduation period. Sourced from references, some of the data attributes used include gender, GPA, last education, concentration [5], regional origin [6] [7], study period [8], and regular.

LITERATURE REVIEW

Several previous research studies observing student graduation predictions include those quoted from Amirah Mohamed Shahiria, Wahidah Husaina, Nur'aini Abdul Rashida, 2015, in a study entitled "A Review on Predicting Student's Performance Using Data Mining Techniques", that student performance is an important part of the institution. higher education. This is one of the criteria for evaluating quality universities based on their excellent academic record. Another assessment can also be seen from how much achievement the number of student graduations can be achieved at the right time [9]. Kolo David Koloa, Solomon A. Adepojub, and John Kolo Alhassan, 2015, in the research "A Decision Tree Approach for Predicting Students Academic Performance", used the Decision Tree algorithm approach in predicting student academic performance, with attributes such as grades, graduation status, type of gender, finance, and learning motivation [3].

To realize a data classification and prediction system, it is first necessary to analyze and understand the existing data and the problems that must be solved. solvable, because data may be incomplete, inconsistent or noisy, i.e. zero values, outliers, therefore pre-processing of data and problems is very important for accurate classification and prediction system [10]. To understand the existing dataset, it must be explored statistically, as well as visualize it using plots and graphic diagrams. This step in data mining is very important as it allows the researchers as well as the readers to understand the data before diving into implementing more complex data mining tasks and algorithms [11].

Developed a decision support system with a decision tree algorithm, with an average accuracy of 86.77% in predicting student graduation. The system received an assessment from respondents of 4.69 and was very acceptable for implementation [12]. Ace C. Lagman, Lourwel P. Alfonso, Marie Luvett I. Goh, Jay-ar P. Lalata, Juan Paulo H. Magcuyao, and Heintjie N. Vicente, 2020, on the research “Classification Algorithm Accuracy Improvement for Student Graduation Prediction Using Ensemble Model”, usingthe ensemble technique approach in predicting student graduation analysis, and getting the best accuracy with a value of 88.3% on the test data [6]

This artificial neural network model is built based on students' cognitive and non-cognitive measures, along with background information [13]. Adewale OpeoluwaOgunde, 2014 conducted a research entitled “A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm”. In this study, a data mining system was built to predict college studentgraduation, using the Iterative Dichotomiser 3 (ID3) decision tree algorithm with desktop-based Java programming [2]. Fujie Zhou, Lingxiao Xue, Zhiqiang Yan, and Yongxian Wen, 2019 with the research "Research on college graduates employment prediction model based on C4.5 algorithm", predicting student graduation job decisions using the C4.5 decision tree algorithm, and using Cross Validation evaluation to verify the accuracy of the model, and gradually achieve high accuracy and excellent predictive models [14]

Compared the Decision Tree, KNN, and SVM algorithms on predicting student graduation performance, and obtained SVM has an accuracy value with the highest score of 95%, but the Decision Tree algorithm, can also predict very well if it uses the complexity parameter (cp) = 0.6689113 model [15]. Receiver Operating Characteristics (ROC) curves can also be used to evaluate classification algorithms. The ROC curve measures the performance of the model by plotting the true positive rate against thefalse positive rate. Perfectly discriminated tests have an ROC plot that passes the upper test angle (100% sensitivity, 100% specificity) [16]

Predicts the graduation of students from a university using Logistic Regression and Random Forest algorithms, and the number balancing algorithm. classes in the data, including SMOTE, Adasyn, and Modified Adasyn [17]. In other research studies using different algorithms, a modelto predict student academic achievement using decision trees and the K-Means algorithm has better accuracy and is easily implemented in higher institutions to predict student performance [18]. Farid Jauhari, Ahmad Afif Supianto, 2018, to increase the best performance score on the decision tree, in predicting student performance, three boosting algorithms are proposed (C5.0, adaBoost.M1, and adaBoost.SAMME) [19].

RESEARCH METHODOLOGY

In this study, the raw data were obtained from the Head of the Informatics Engineering Study Program, Muhammadiyah University of Riau. This dataset includes private data stored in Microsoft Excel, in .xls format and has been modified with the permission of the Head of Study Program, as data processing material for research. The dataset contains alumni students of the Informatics Engineering study program, namely student data from the class of 2008 to 2013 totaling 213 rows of data, which are stored in Microsoft Excel and named the file dataset_mahasiswa_full.csv. This data is used to predict student graduation in the future, with the stages of data mining described in the chart below.

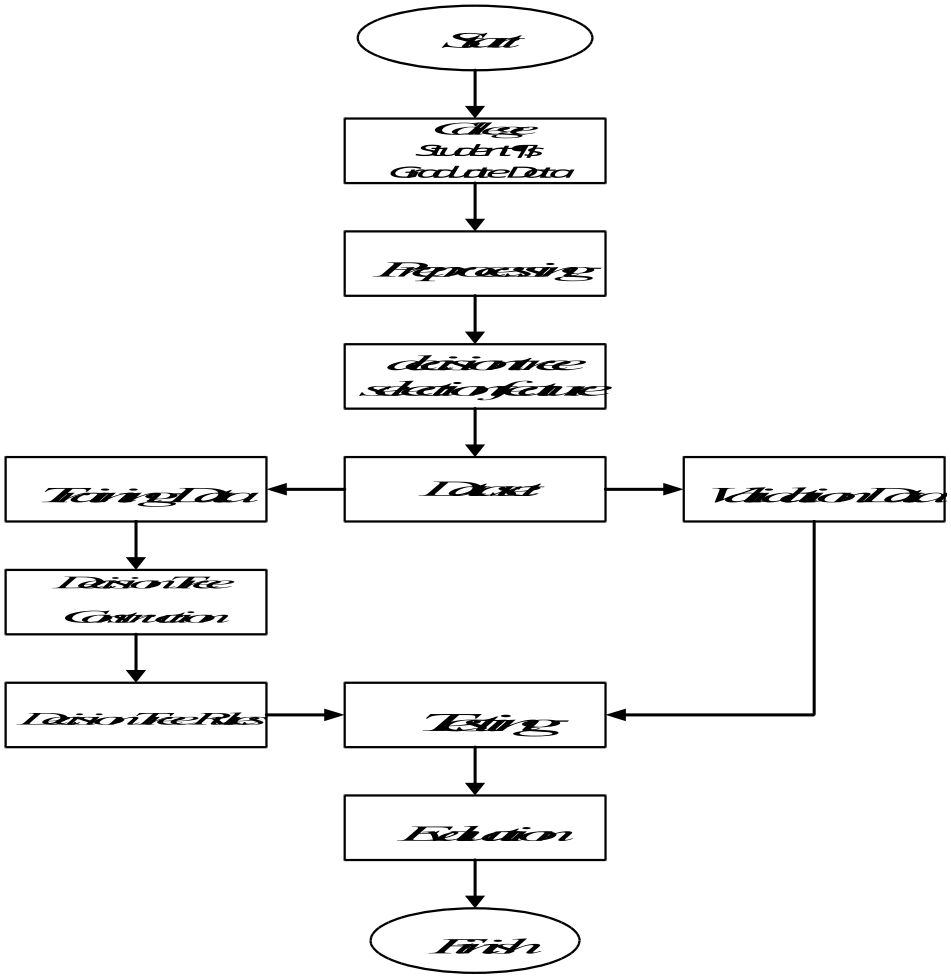

***FIGURE 1**. Research Flow*

Raw Data

The data consists of 213 rows of alumni student data, and 10 attribute columns, namely serial number, Student Identification Number (NIM), gender, GPA value, specialization or concentration, regular class category, regional origin, last education, year of graduation, and period of time. studies. Table 1. below is an example of 20 lines of raw data for alumni students, which must go through the preprocessing stage first, in order to get quality data that can be used in the data mining process flow according to Figure 1. above.

***TABLE 1.** Sample Raw Dataset*

No	NIM	Sex	Grade-Point Average	Concentration	Regular	Origin	Last Education	Graduation Year	Graduation

1	010	Male	3.56	Software Engineering	B	Outside Riau Province	Vocational School	2012	4 Years
2	013	Female	2.83	Software Engineering	A	Riau Province	Senior High School	2017	> 4 Years
3	015	Male	3.64	Computer Networking	A	Outside Riau Province	Islamic Senior High School	2013	4 Years
4	012	Male	3.74	Computer Networking	A	Riau Province	Senior High School	2017	> 4 Years
5	013	Female	3.28	Software Engineering	A	Riau Province	Vocational School	2014	> 4 Years
6	014	Male	2.95	Computer Networking	A	Outside Riau Province	Senior High School	2012	4 Years
7	013	Male	3.71	Computer Networking	A	Outside Riau Province	Senior High School	2014	4 Years
8	016	Female	2.82	Software Engineering	B	Riau Province	Vocational School	2015	4 Years
9	021	Female	2.78	Software Engineering	B	Outside Riau Province	Vocational School	2014	4 Years
10	025	Male	2.91	Computer Networking	A	Riau Province	Senior High School	2014	> 4 Years
11	030	Male	3.69	Software Engineering	B	Outside Riau Province	Vocational School	2014	4 Years
12	028	Female	2.77	Software Engineering	A	Riau Province	Senior High School	2012	> 4 Years
13	035	Male	3.28	Computer Networking	A	Outside Riau Province	Islamic Senior High School	2015	4 Years
14	036	Female	2.85	Software Engineering	B	Outside Riau Province	Vocational School	2012	4 Years
15	005	Male	2.99	Computer Networking	A	Riau Province	Senior High School	2016	> 4 Years
16	006	Male	2.82	Computer Networking	A	Outside Riau Province	Senior High School	2017	4 Years
17	024	Male	2.82	Computer Networking	A	Outside Riau Province	Senior High School	2014	4 Years
18	030	Male	3.42	Computer Networking	A	Riau Province	Senior High School	2012	> 4 Years

19	032	Female	3.34	Software Engineering	A	Riau Province	Vocational School	2017	> 4 Years
20	036	Female	3.72	Software Engineering	A	Riau Province	Senior High School	2015	> 4 Years

Preprocessing

Data Selection

It is an attribute selection process when building a student graduation prediction decision tree, not all attributesin the raw data are used, in this case discarding the attribute along with the data row No. and NIM. The predictor attributes selected were gender, GPA, concentration, regularity, regional origin, and last education, then the class attribute was the study period [20].

Data Transformation

The first stage of data transformation is the data row to be changed is the numeric value data on the GPA attribute. GPA >=3.51 is categorized as “Very Satisfactory”, GPA >=3.00 && <3.51 is categorized as “Satisfactory”, and GPA < 3.00 “Unsatisfactory”. Then the name of the attribute of the study period was changed to graduation which was divided into two categories, namely = 4 Years "On Time" and > 4 Years "Late".

The second stage of data transformation is carried out again with data on predictor attributes with character data types. Implementation of data mining calculations using Decision Tree in python can only process data of integer or numeric type, therefore, data transformation needs to be done by converting categorical variable values into integer data types such as 0, 1 and so on [21]. Dataset

Data Training

Training data or training data is used to build a machine learning model in data mining, in this case the Decision Tree method, making a decision tree. The data used is the entire dataset totaling 213 rows of data, stored in Microsoft Excel and named the *file college_student_dataset.csv.*

Data Validasi

Validation data is testing data or test data outside of the dataset, with the aim of testing and knowing the results of predictions by the decision tree model. The validation data contains data from the 2016 batch of students totaling 110 people, consisting of several regular students A and regular B, stored in Microsoft Excel and named the *file college_student_data_validation.csv*

Student Graduation Decision Tree

The decision tree is built using training data from a total of 213 datasets, with attributes of gender, GPA, concentration, regularity, and regional origin. At a node or decision tree symbol there is an attribute name, the result of calculating the entropy value, the total number of samples, and the number of samples for each class.

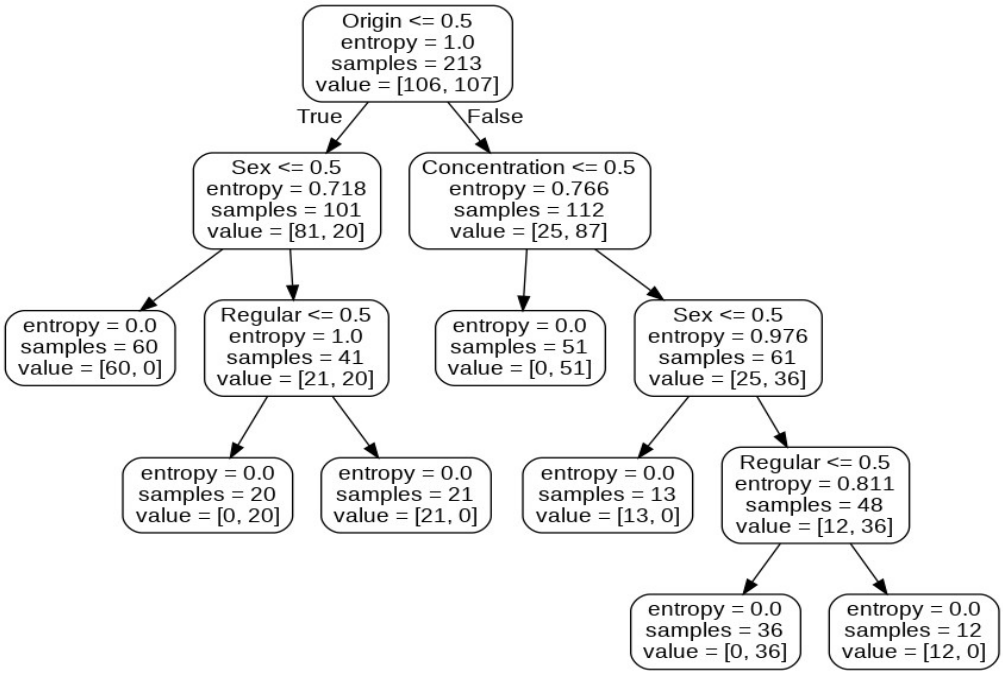

FIGURE 2. Student Graduation Decision Tree

How to get the decision tree structure in Figure 2. above, according to the steps for calculating the Decision Tree ID3 method. Starting from the node at the very top, the top node of the regional origin attribute, descends to the left branch until a decision node is obtained and is complete, then continues and is also completed on the right branch. The arrow lines indicate the condition that is notated by the content of the attribute data, for example, from the area outside Riau with the down arrow pointing to the left, and Riau with the down arrow pointing to the right. Branching nodes are indicated by the attribute names that are above in each of these nodes. The rules formed from the student graduation decision tree above can be shown in table 1. Below

TABEL 2. *Decision Tree Rules*

Left Side Decision Tree			**Right Side Decision Tree**			
Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Rule 7
True	True	True	False	False	False	False
True	False	False	True	False	False	False
On Time	True	False	Delayed	True	False	False
	Delayed	On Time		On Time	True	False
					Delayed	On Time

Decision Tree Performance Evaluation

Validation data totaling 110 rows of data is used to predict student graduation in the future. There are two labels on the graduation attribute, namely, 50 data on time and 60 data late. The imbalance in the comparison of the amount of data in the two tables is calculated as a difference of 10 rows of data. Then balancing the number of minority classes, namely on the label on time, using the Synthetic Minority Oversampling Technique (SMOTE) technique. After the SMOTE process, the number of timely and late label data becomes balanced, and the number ofvalidation data increases to 120 data. Table 3. below represents the evaluation of the confusion matrix of the prediction results of student graduation on the validation data

TABLE 3. Prediction Results of Student Graduation based on Confusion Matrix

Graduation Comparison	**Student Graduation Prediction Results**

Vol 5-Sep2021
ISSN: 2541-3023

	On time	Late

Vol 5-Sep2021
ISSN: 2541-3023

Actual Student Graduation

On time	58	2
Late	2	58

From table 4 confusion matrix above, there are 4 terms from the results of the prediction of student graduation predictions, namely:

1. True Positive, namely the number of student graduation data that were predicted to be correct on time, totaling 58data with a percentage value of 48.333%.
2. False Positive, namely the number of student graduation data that is predicted to be late, but the student validation data is on time, totaling 2 data with a percentage value of 1.667%.
3. False Negative, namely the number of student graduation data that is predicted to be on time, but in the validation data the student is late, totaling 2 data with a percentage value of 1.667%.
4. True Negative, namely the number of student graduation data that were predicted to be late, totaling 58 data witha percentage value of 48.333%.

The number of correct graduations on time is the same as true late graduation as many as 58 data. Errors in the prediction results of student graduation, both in false positives and false negatives, are both worth 2. Then the decision tree performance is calculated on the test results of predictions for student graduation, obtained 96.67% accuracy, 96.67% precision, 96.67% recall, and F1 -Score 96.67%. The assessment score is shown in table 3. below

TABLE 4. Evaluation of Decision Tree Performance

Rating Score

Vol 5-Sep2021
ISSN: 2541-3023

$$\text{Accuracy } \% = \frac{TP+TN}{TP+FP+FN+TN} = \frac{58+58}{58+2+2+58} = \frac{116}{120} = 96.67$$

$$\text{Precision } \% = \frac{TP}{TP+FP} = \frac{58}{58+2} = 58 = 96.67\%$$

$$\text{Recall } \% = \frac{TP}{TP+FN} = \frac{58}{58+2} = 58 = 96.67\%$$

$$F\ 1-\text{Score } \% = 2 * \text{Presisi} * \text{Recall} = 2 * 0.9667 * 0.9667 = 2 * 0.9345 = 96.6$$

$$\text{Presisi} + \quad 0.9667+0.966 \quad 1.933$$

In this test an error occurred in predicting student graduation. There are 2 student graduation data that are actually on time, but the prediction results are late, and there are 2 student graduation data that are actually late, but the prediction results are on time.

Measuring Student Graduation Decision Tree Performance

Create a ROC (Receiver Operating Characteristic) curve in the test result, to visualize the accuracy graph with a value of 96.67%, between false positives as the x-axis and true positives as the y-axis. To be able to see the accuracymanually, a classification comparison is carried out using the ROC curve as a result of the expression from the confusion matrix [6]. The level of accuracy is measured as follows : *Accuracy 0.90 – 1.00 = Excellent classification, Accuracy 0.80 – 0.90 = Good classification, Accuracy 0.70 – 0.80 = Fair classification, Accuracy 0.60 – 0.70 = Poor classificatio, Accuracy 0.50 – 0.60 = Failure*

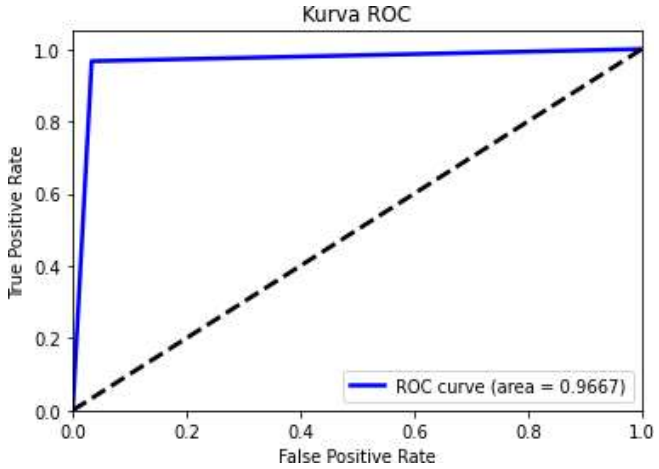

FIGURE 3. ROC curve

The ROC curve graph in Figure 3. shows that an accuracy rate of 0.9667 or 96.67%, which is produced in predicting student graduation on validation data, which is 120 lines of data using the Decision Tree method or decision tree, is classified as a very good type of classification.

IV. CONCLUSION

The research predicts future student graduation using the Decision Tree method, it can be concluded as follows: (1) student graduation prediction testing is carried out on a dataset of 213 data, where all of these datasets become training data for making decision trees. There are 110 validation data which are testing data outside of the dataset, totest predictions of student graduation with the technique of balancing the number of classes, namely the Synthetic Minority Oversampling Technique (SMOTE). This test resulted in an accuracy rate of 96.67%, precision 96.67%, recall 96.67%, and f1-score 96.67%; (2) evaluation of the performance of the decision tree in predicting student graduation, showing the occurrence of misclassification or prediction errors in 2 student graduation data which were actually on time, but the prediction results were late or termed as false positives, and there were also 2 student graduation data which were actually late , but the prediction result is on time or termed as false negative; (3) research results with the level of accuracy that has been achieved in the test in predicting the graduation of regular Aand regular B students in the future, using the Decision Tree method with ROC curve (Receiver Operating Characteristic) visualization, including the type of classification that is very good.

REFERENCEE

[1] M. A. Al-barrak and M. Al-razgan, “Predicting Students Final GPA Using Decision Trees : A Case Study,” *Int. J. Inf. Educ. Technol.*, vol. 6, no. 7, pp. 528–533, 2016.

[2] A. O. Ogunde, “A Data Mining System for Predicting University Students ’ Graduation Grades Using ID3Decision Tree Algorithm,” *Comput. Sci. Inf. Technol.*, vol. 2, no. 1, pp. 1–26, 2014.

[3] K. David, S. A. Adepoju, and J. Kolo, "A Decision Tree Approach for Predicting Students Academic Performance," *I.J. Educ. Manag. Eng.*, vol. 5, no. 5, pp. 12–19, 2015.
[4] F. Yang, "Decision Tree Algorithm Based University Graduate Employment Trend Prediction," *Informatica*, vol. 43, no. 4, pp. 573–579, 2019.
[5] S. Liu, "Knowledge Discovery of the Orientation of University Graduates in Beijing Area Based on Decision Tree," *Comput. Sci. Eng.*, pp. 282–286, 2018.
[6] A. C. Lagman, L. P. Alfonso, M. L. I. Goh, J. P. Lalata, J. P. H. Magcuyao, and H. N. Vicente, "Classification Algorithm Accuracy Improvement for Student Graduation Prediction Using Ensemble Model," *Int. J. Inf. Educ. Technol.*, vol. 10, no. 10, pp. 723–727, 2020.
[7] A. E. Tatar and D. Dü, "Prediction of Academic Performance at Undergraduate Graduation : Course Grades or Grade Point Average?," *Appl. Sci.*, vol. 10, no. 14, pp. 1–15, 2020.
[8] W. Y. Chin, C. Keong, and J. M. Jamil, "A Study of Graduate on Time (GOT) for Ph . D Students using Decision Tree Model," *AIP Conf. Proc.*, vol. 2138, no. 1, pp. 1–6, 2019.
[9] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Procedia Comput. Sci.*, vol. 72, pp. 414–422, 2015.
[10] B. Khan, M. S. Hayat, and M. Daud, "Final Grade Prediction of Secondary School Student using Decision Tree," *Int. J. Comput. Appl.*, vol. 115, no. 21, pp. 32–36, 2015.
[11] A. A. Saa, "Educational Data Mining & Students' Performance Prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 5, pp. 212–220, 2016.
[12] M. C. G. Fernando *et al.*, "Development of a Predictive Decision Support System for Student Graduation using a Decision Tree Algorithm," *Int. J. Simul. Syst. Sci. Technol.*, vol. 20, no. S2, p. 27.1-27.6, 2019.
[13] E. G. Dada, J. S. Bassi, E. G. Dada, A. A. Hamidu, and M. D. Elijah, "Students Graduation on Time Prediction Model Using Artificial Neural Network," *IOSR J. Comput. Eng.*, vol. 21, no. 3, pp. 28–35, 2019.
[14] F. Zhou, L. Xue, Z. Yan, and Y. Wen, "Research on college graduates employment prediction model based on C4 . 5 algorithm," *J. Phys. Conf. Ser.*, pp. 1–6, 2020.
[15] S. Wiyono and T. Abidin, "Comparative Study of Machine Learning KNN, SVM, and Decision Tree Algorithm to Predict Student's Performance," *Int. J. Res. -GRANTHAALAYAH*, vol. 7, no. 1, pp. 190–196, 2019.
[16] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised data mining approach forpredicting student performance," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 16, no. 3, pp. 1584–1592, 2019.
[17] H. A. Gameng, B. D. Gerardo, and R. P. Medina, "A Modified Adaptive Synthetic Smote Approach in Graduation Success Rate Classification," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 6, pp. 3053–3057, 2019.
[18] T. M. Ogwoka, W. Cheruiyot, and G. Okeyo, "A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms," *Int. J. Comput. Appl. Technol. Res.*, vol. 4, no. 9,pp. 693–697, 2015.
[19] F. Jauhari and A. A. Supianto, "Building student's performance decision tree classifier using boosting algorithm," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1298–1304, 2019.
[20] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 2, p. 26, 2018.
[21] P. J. M. Estrera, P. E. Natan, B. G. T. Rivera, and F. B. Colarte, "Student Performance Analysis for Academic Ranking Using Decision Tree Approach in University of Science and Technology of Southern Philippines Senior High School," *Int. J. Eng. Tech.*, vol. 3, no. 5, pp. 147–154, 2017.