Journal of
Information Systems Engineering
and Business Intelligence

Vol.7, No.1, April 2021
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Predicting Students Graduate on Time Using C4.5 Algorithm

Herman Yuliansyah[1)], **Rahmasari Adi Putri Imaniati**[2)],
Anggit Wirasto[3)], **Merlinda Wibowo**[4)*]

[1,2)]*Department of Informatics, Universitas Ahmad Dahlan, Indonesia*
Jl. Ringroad Selatan, Kragilan, Tamanan, Kec. Banguntapan, Bantul, Yogyakarta
[1)]herman.yuliansyah@tif.uad.ac.id, [2)]rahmasariadiputri@gmail.com

[3)]*Department of Information Technology, Universitas Harapan Bangsa, Indonesia*
Jl. Raden Patah No. 100, Ledug, Kec. Kembaran, Banyumas, Jawa Tengah
[3)]anggitwirasto@uhb.ac.id

[4)]*Department of Informatics, Institut Teknologi Telkom Purwokerto, Indonesia*
Jl. D.I Panjaitan No. 128, Kec. Purwokerto, Banyumas, Jawa Tengah
[4)]merlinda@ittelkom-pwt.ac.id

Abstract

Background: Facilitating an effective learning process is the goal of higher education institutions. Despite improvement in curriculum and resources, many students cannot graduate on time. Mostly, the number of students who graduate on time is lower than the number of new students enrolling to universities. This could dilute the chance for students to learn effectively as the ratio between faculty members and students becomes non-ideal.
Objective: This study aims to present a prediction model for students' on-time graduation using the C4.5 algorithm by considering four features, namely the department, GPA, English score, and age.
Methods: This research was completed in three stages: data pre-processing, data processing and performance measurement. This predicting scheme make the prediction based on the department of study, age, GPA and English proficiency.
Results: The results of this study have successfully predicted students' graduation. This result is based on the data of students who graduated in 2008-2014. The prediction performance result achieved 90% of accuracy using 300 testing data.
Conclusion: The finding is expected to be useful for universities in administering their teaching and learning process.

Keywords: Classification, C4.5 Algorithm, Data Mining, Prediction, Student Academic Performance, Student Graduation

Article history: Received 3 February 2021, first decision 71 February 2021, accepted 21 March 2021, available online 28 April 2021

I. Introduction

Whether or not students can experience optimum learning and then graduate on time depends on, among others, quality education at the university, the degree program, and the quality of facilities and human resources [1]. Late graduation is likely to cause extra workload for faculty members because they have to, for example, supervise more students at a time. Therefore, universities usually have a strategy to improve and maintain the on-time graduation rate[2] [3].

Data mining can extract educational data to improve the education process quality [4] and identify strategies for improving the students' performance [5]. 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 [4] or to design intervention to prevent failure [6]. Data mining has three main functions, which are clustering data [7][8], classifying data [9][10], and identifying association rules patterns [11]. The current student performance prediction study shows that student performance prediction is challenging due to educational data variants [12][13]. A framework of an intelligent recommender system based on background factors was designed by Goga *et al.* [14] to recommend necessary actions for improvement. Ashraf *et al.* [15] also develop an intelligent prediction system based on ensemble

[*] Corresponding author

doi: http://dx.doi.org/10.20473/jisebi.7.1.67-73

and filtering approaches. Yang and Li [16] proposed a function to identify student potential by evaluating the achievement. The results show that the proposed tools achieve more accurate results. Hamsa *et al.* [17] also develop student's academic performance prediction models using two selected classification methods, and the practical implication is to help students improve their performance. Furthermore, Sahiri *et al.* [13] evaluated five prediction students' performance techniques effectiveness and determined that the decision tree technique is the second-best technique that outperforms other classification techniques. Helal *et al.* [18] create several classification models for student performance prediction, and the experiments concluded that no particular method shows superior performance.

Meanwhile, Kurniawan *et al.* [19] proposed a graduation prediction system by using the C4.5 algorithm and a small dataset [21][22]. This study could be improved by increasing the dataset. In previous studies, an association rules pattern analysis was created by using several variables: duration of study, the amount of time to complete the thesis, GPA, and English proficiency [21]. The result showed that English proficiency score, GPA, and study duration have a significant impact [21]. Meanwhile, the student performance characteristic can be classified into excellent performance, standard performance, and underperformance [22].

This study's main objective is to present a prediction model by using C4.5 algorithm to predict students' graduation by considering four features, namely the department, GPA, English score, and age. The decision tree result is expected to be useful to inform university's staff members to prevent student failures and design intervention [23]. This paper is organised into four sections: Section 2 reviews the literature; Section 3 explains the methodology of research; Section 4 presents the result and discussion, and the last section concludes the study.

II. Research Methodology

The prediction model contains three phases, namely data pre-processing, data processing and performance measurement, as shown in Fig. 1.

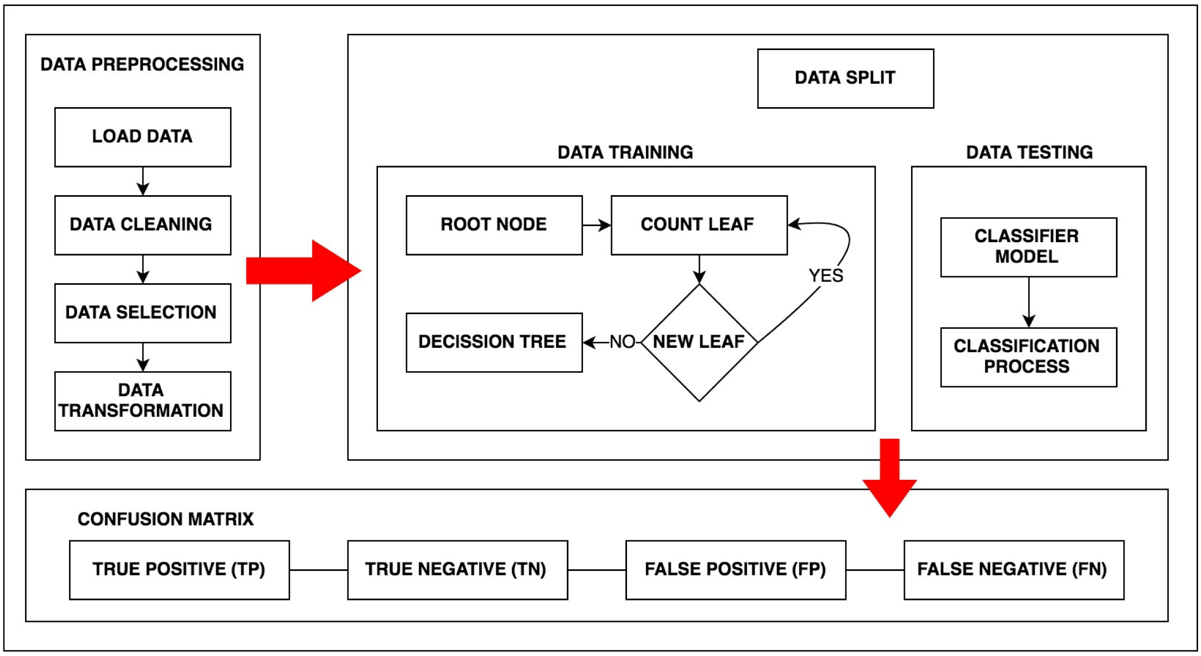

Fig. 1 Prediction for on-time Graduation Model

A. Data Pre-processing

Data pre-processing is needed to enhance the quality of data. The data source has potential problems due to human error. For example, the data from the administration unit's internal format may not be reliable and inconsistent. Therefore, data pre-processing needs to be done before the data processing phase. It can investigate and identify useful data attributes. The data attribute is required for the processing of the data phase. The data phase processing consists of four steps: data loading, data cleaning, data selection, and data transformation, as shown in Fig. 1. The process of loading data is the initial process in this step. The data collected in the form of an excel file is then loaded into the program and then processed. Data cleaning is conducted by removing data that is not necessary or will not be used for prediction. These data are data that are inconsistent as the same data or wrong.

This study uses student graduation data from the faculty of engineering of a private university with Grade A in 2008-2014. Student information that does not graduate on time from the data will not be used. The data selection step is to determine the attributes used in the data mining process. Attributes to be used include age when enrolling, department, GPA, English proficiency score, and whether or not they graduate on time. The purpose of the selection is to get accurate results from the data used. The data transformation step is to change the predictor attribute data from the alphabet form into numeric form. The results of data collection for 2008-2014 graduation obtained 740 student data. The following is a set of student graduation data that can be seen in Table 1.

TABLE 1
SAMPLE OF STUDENT DATA FOR ACADEMIC YEAR 2008-2014

No.	Student Number	Department	Date Reg.	Date Grad.	BOD	GPA	English Score
1	xxxxxx	Informatics	1 September 2014	24 Feburary 2018	14 Januari 1996	3,81	433
2	xxxxxx	Informatics	1 September 2014	28 June 2018	2 January 1997	3,64	503
3	xxxxxx	Informatics	1 September 2014	28 June 2018	30 July 1996	3,62	456
4	xxxxxx	Informatics	1 September 2014	28 June 2018	13 April 1995	3,54	420
5	xxxxxx	Informatics	1 September 2014	23 July 2018	25 May 1996	3,69	436
....	....	....	....	....	....	....	....
740	xxxxxx	Chemical Engineering	8 September 2008	14 October 2013	5 December 1989	3,06	400

The data in Table 1 has the following components:

1) *Student Number*
   This is a unique ID for each student—this ID created by the university to distinguish between each student's data.
2) *Date Registration*
   The date when students when entering or registering as a student at the university.
3) *Date Graduation*
   The date when students graduate from the university.
4) *Department*
   This attribute consists of study program in Faculty of Industrial Engineering: Informatics, Chemical Engineering, Industrial Engineering, and Electrical Engineering.
5) *Birth of Date*
   This attribute contains the date of birth of students. This variable is to find out the student's age.
6) *GPA*
   The GPA data variable is based on the GPA data of students who have graduated. The number of GPA varies between 2.48 - 3.87. Information on the graduation predicate according to academic regulations, is as follows: GPA score < 2.76 given the title Pass, GPA score 2.76 to 3.00 awarded the title Satisfactory, GPA score of 3.01 to 3.50 given the title Very Satisfactory and a GPA score of 3.51 to 4.00 are given the title of Distinction (*Cumlaude*).
7) *English Proficiency Score*
   English proficiency score (TOEFL) for graduation in the engineering faculty is a minimum of 400. The value is used to be a thesis examination requirement for each student. The range of the English proficiency score are: 400, 410 - 420, 421 - 430, 431 - 440, 441- 450, 451-460 461-470, 471-480, 481-490, 491-500, 510-520, 521-530, 531-540, 541-550, 551-560, 561-570, 571-580, 581-590, 591-600. The range grouped into 10. This range transformation can affect the pre-processing time becomes more efficient.

B. Data Processing (Prediction Modelling)

The data processing phase is conducted by splitting the data from data pre-processing into data training and data testing, as shown in Fig. 1. The data training used to get knowledge from the data, training data modelling the C4.5 algorithm to find the node or root until the last branch cannot be counted anymore. The root node step process is the most important step because the data that has been transformed will be filtered using the C4.5 algorithm. The root node's first step is to determine the root node that will be used for branching. Furthermore, making a new leaf or node will be created if the previous node can still be calculated further. The new leaf or node calculated until the final step finds a decision. The decision tree results from the analysis of problem-solving decisions, depending on the likelihood or probability of the decision. The decision tree results are obtained from calculating a leaf or node, and each leaf node marks the class label. The decision tree process changes the data tables' shape into a tree model and then transforms the tree model into rules.

C. Performance Measurement

The confusion matrix is used to measure the performance of the proposed prediction model [24]. There are four terms representing the classification process results, as shown in Fig. 1 and 2. The four terms are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). True Negative (TN) value is the amount of negative data detected correctly, while False Positive (FP) is negative data but identified as positive data. Meanwhile, True Positive (TP) is positive data that is detected correctly. False Negative (FN) is the opposite of True Positive, so the data is positive but identified as negative data.

Based on the value of True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP), values of accuracy, precision and recall can be obtained [25]. Accuracy scores describe how accurately the system can classify data correctly. In other words, the accuracy score is a comparison between correctly classified data and the whole data. Equation (1) can obtain accuracy scores. Precision scores describe the number of positive categorised data correctly divided by the total positive classified data. Equation (2) can get precision. Meanwhile, recall shows what percentage of positive category data is correctly classified by the system. Equation (3) obtains the recall value, and F1-Score is shown in (4).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \qquad (1)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \qquad (2)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \qquad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (4)$$

III. RESULTS

In the data pre-processing phase, student data is uploaded separately based on the majors: informatics, industrial engineering, chemical engineering and electrical engineering study program in 2008-2014 academic year. The format is spreadsheet file. Then, the selection process is conducted to retrieve the components used in this study. The data selected and managed through data cleaning that eliminates empty data or incomplete data. It consists of the registration date, graduation date, GPA, and English proficiency score. A total of 640 rows of data were obtained from the cleaning process. After the cleaning process, the data is transformed into other forms to suit the data mining process. The result of the pre-processing data is illustrated in Table 2.

TABLE 2
DATA PRE-PROCESSING RESULT

No.	Student Number	Department	GPA	English Score	Age	Graduate Information
1	xxxxx	Informatics	Distinction (Cumlaude)	< 440	19	GoT
2	xxxxx	Informatics	Distinction (Cumlaude)	< 510	18	GoT
3	xxxxx	Informatics	Distinction (Cumlaude)	< 460	19	GoT
4	xxxxx	Informatics	Distinction (Cumlaude)	< 430	20	GoT
5	xxxxx	Chemical Engineering	Very Satisfactory	< 410	22	Not GoT
….	….	….	….	….	….	….
640	xxxxx	Chemical Engineering	Satisfactory	< 420	20	Not GoT

The pre-processing results' main features are GPA, English proficiency score, age and graduation predicate. The graduation predicate is addressed as Pass, Satisfactory, Very satisfactory, and Distinction (*Cumlaude*). Then, the phrase GoT for students who graduated on time and Not GoT for students who did not graduate on time were used to label the graduation information. The graduate information is calculated by the graduate date minus the registration date.

The data pre-processing result is used as training and testing data. Training data is used to form a classifier model. The results of the C4.5 implementation used for training data are 640 data. Meanwhile, the testing data is used to test the performance and correctness (of correlations) in the relevant model. In the data testing section, accuracy testing is performed using the confusion matrix. The prediction results are in the form of a new label for training data obtained from a classification using the C4.5 algorithm, as shown in Table 3. There is a new column named prediction.

TABLE 3
PREDICTION STUDENT GRADUATE ON TIME RESULT

No.	Student Number	Department	GPA	English Score	Age	Prediction
1	xxxxx	Informatics	Praise	< 400	17	Not GoT
2	xxxxx	Informatics	Praise	< 400	17	Not GoT
3	xxxxx	Informatics	Praise	< 400	17	Not GoT
4	xxxxx	Informatics	Very Satisfactory	< 410	18	Not GoT
5	xxxxx	Informatics	Very Satisfactory	< 410	18	Not GoT
….	….	….	….	….	….	
640	xxxxx	Informatics	Very Satisfactory	< 400	19	Not GoT

Fig. 2 shows the rule that if the informatics study program student's GPA is very satisfactory, English proficiency score is < 400, the age when entering is 18 years old, this student is included in graduating on time. Furthermore, suppose students of the Informatics study program, GPA are very satisfactory. In that case, English proficiency score < 400, age at entry is 18 years old, this student is considered to have graduated not on time. Moreover, if the Informatics study program student, GPA is very satisfactory, English proficiency score < 400, age at entry is 18 years old, this student is considered to have graduated not on time.

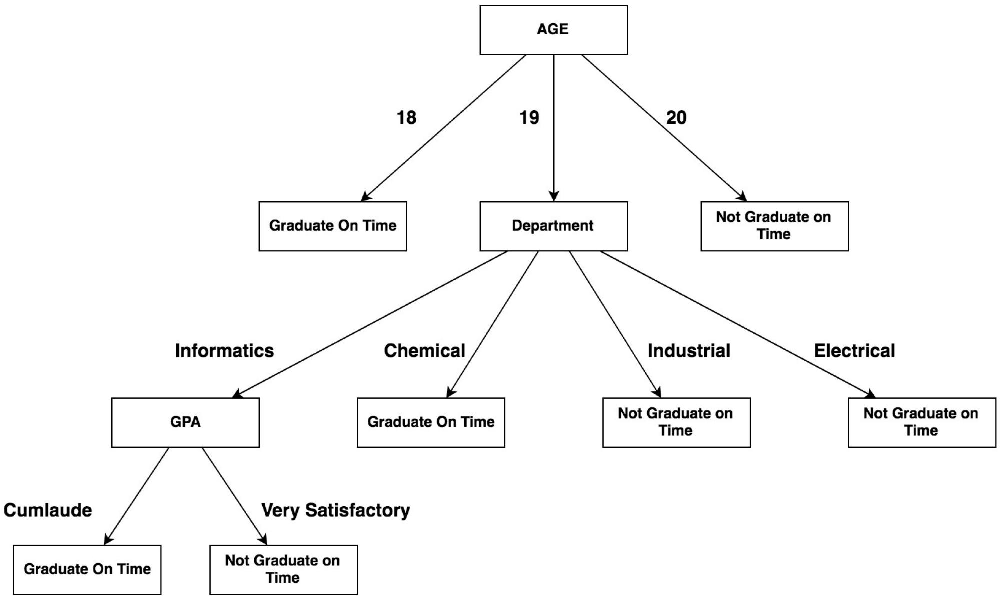

Fig. 2 Decision Tree of Student On-time Graduation

Based on the results of the decision tree shown in Fig. 2, prediction using the C4.5 algorithm obtained a pattern or prediction rule on time as follows:

1) If the age of 18 years, it is right on time.
2) If the age of 19 years, informatics study program, and GPA predicate is with distinction, then it is on time. If the age of 19 years, Informatics Engineering study program, and the GPA of the category is very satisfactory, then it is not on time.
3) If the age is 19 years old and from Chemical Engineering, then it is not on time.
4) If the age is 19 years old and from Industrial Engineering, then it is not on time.
5) If the age is 19 years old and from Electrical Engineering, then it is not on time.
6) If the age is 20 years, it is not graduate on time.

The experiment was conducted in five iterations by testing 100, 200, 300, 400 and 500 test data, as shown in Table 4. Based on this experiment, measurement using a confusion matrix is conducted to obtain the value of accuracy, precision, recall and F1 score. The best value achieved for accuracy is 90.00% with 300 test data, while the best precision and F1 score achieved are 61.90% and 57.77%, respectively, with 100 data testing. Furthermore, the best recall value is 54.76% for 400 data testing. The performance of each measurement result varies and the stability of the advantages of the measurement results also varies. This can be caused by the prediction model that requires improvement in the pre-processing phase to obtain a more stable training data input by the C4.5 algorithm. On average, based on these four iterations, the C4.5 algorithm succeeded in obtaining values for accuracy, precision, recall and F1 score of 87.44%, 52.84%, 50.68% and 51.73% respectively.

TABLE 4
EXPERIMENT RESULT

Number of testing data	Precision	Recall	Accuracy	F1-Score
100	**61.90 %**	54.16 %	81.0 %	**57.77 %**
200	51.85 %	53.84 %	87.5%	52.82 %
300	46.42%	46.42%	**90.0 %**	46.42 %
400	51.11 %	**54.76 %**	89.75 %	52.87 %
500	51.92 %	47.368%	89.0 %	49.53 %
Average	52.84%	50.68%	87.44%	51.73 %

IV. DISCUSSIONS

This study presents a prediction model for students' on-time graduation using the C4.5 algorithm. The data is collected from the faculty of engineering from a private university in the academic year 2018-2014. The data consist of several features: student number, department, registration date, graduation date, the date of birth, GPA, and English proficiency score. The data were processed using pre-processing stage in cleaning, selection, and transformation. Four features were considered as prediction features, namely department, GPA, English score, and age. Age is calculated based on the date of birth and date of registration to the university. The evaluation of C4.5 algorithm performance is conducted in five iterations to obtain the value of accuracy, precision, recall, and F1 score.

The C4.5 algorithm started the prediction by determining the root node. Based on the C4.5 algorithm, age is determined as the root node. Furthermore, the department and GPA are determined as the branch. Graduating on time and not graduating on time are the end of the branch. Six rules are generated from the decision tree as the learning phase of the dataset. The rules can be used by the university to improve the quality of education and prevent student failures in the education process and achieve on-time graduation.

There are at least two limitations of this study for evaluation performance improvement. First, to avoid the imbalance of data transformation for English scores, the English score needs to be classified as equal to the number of classifications in age, department, and GPA. Second, to avoid bias sampling, the experiments are conducted using cross-validation. This study used a split test for the sampling approach. This sampling allows for the potential for sampling bias, although in this study, no impact was seen. Because this study uses iterations in which the amount of testing data determines the iteration. In future research, other data variables can also be considered based on data availability.

This study still achieved a higher accuracy even though it uses a large test data. However, the classification technique for prediction used in this study still needs to be optimised using feature selection techniques to achieve the best result of all performance measurements [26]. Furthermore, sampling techniques can also be improved using cross-fold validation to split the test data fairly [27].

V. CONCLUSIONS

In this study, we present a prediction model for students' on-time graduation using the C4.5 algorithm. The dataset was cleaned, selected, and transformed in pre-processing stage and resulted in four features that are considered for prediction. The overall data analysis of students in the academic year 2008-2014 using the C4.5 algorithm produced the highest accuracy of 90% for 300 data testing. Furthermore, the best average classification performance reaches a a precision value of 61.90%, recall value 54.76%, and F1 score 57,77%. The C4.5 algorithm determines the age for the root node. Later, department and GPA determined the branch. However, English score data features were not considered in the tree. This is caused by the data transformation process that divides on a small scale so that it is eliminated.

Author Contributions: *Herman Yuliansyah*: Conceptualization, Methodology, Validation, Resources, Data Curation, Writing - Original Draft, Supervision, Project administration, Funding acquisition. *Rahmasari Adi Putri Imaniati*: Conceptualization, Methodology, Software, Validation, Investigation, Data Curation. *Anggit Wirasto*: Validation, Writing - Review & Editing, Visualization. *Merlinda Wibowo*: Validation, Writing - Review & Editing, Visualization

Funding: This research supported by Universitas Ahmad Dahlan under research scheme Competitive Research Grant (Penelitian Hibah Bersaing/PHB) with grants No: PHB-073/SP3/LPPM-UAD/IV/2019 on 08 April 2019.

Acknowledgments: The authors would like to thank the editors and the anonymous reviewers for their useful comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

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

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] M. Goga, S. Kuyoro, and N. Goga, “A Recommender for Improving the Student Academic Performance,” *Procedia - Soc. Behav. Sci.*, vol. 180, pp. 1481–1488, May 2015.

[15] M. Ashraf, M. Zaman, and M. Ahmed, “An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches,” *Procedia Comput. Sci.*, vol. 167, pp. 1471–1483, 2020.

[16] 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, pp. 97–108, Aug. 2018.

[17] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, “Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm,” *Procedia Technol.*, vol. 25, pp. 326–332, 2016.

[18] S. Helal *et al.*, “Predicting academic performance by considering student heterogeneity,” *Knowledge-Based Syst.*, vol. 161, pp. 134–146, Dec. 2018.

[19] D. Kurniawan, A. Anggrawan, and H. Hairani, “Graduation Prediction System On Students Using C4.5 Algorithm,” *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 19, no. 2, pp. 358–365, 2020.

[20] D. H. Kamagi and S. Hansun, “Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa,” *J. Ultim.*, vol. 6, no. 1, pp. 15–20, 2014.

[21] 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.

[22] K. Sya’iyah, H. Yuliansyah, and I. Arfiani, “Clustering Student Data Based On K-Means Algorithms,” *Int. J. Sci. Technol. Res.*, vol. 8, no. 8, pp. 1014–1018, 2019.

[23] 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.

[24] 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, pp. 1108–1119, 2019.

[25] 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, pp. 216–231, Jul. 2019.

[26] M. DASH and H. LIU, “Feature selection for classification,” *Intell. Data Anal.*, vol. 1, no. 1–4, pp. 131–156, 1997.

[27] T. Fushiki, “Estimation of prediction error by using K-fold cross-validation,” *Stat. Comput.*, vol. 21, no. 2, pp. 137–146, Apr. 2011.