Atlantis Highlights in Engineering (AHE), volume 2
International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018)

Exploration Analysis of Data Mining Algorithm to Predict Student Graduation Target

Rachmadita Andreswari
Faculty of Industrial and System Engineering
Telkom University
Bandung, Indonesia
andreswari@telkomuniversity.ac.id

Muhammad Azani Hasibuan
Faculty of Industrial and System Engineering
Telkom University
Bandung, Indonesia
muhammadazani@telkomuniversity.ac.id

Dela Youlina Putri
Faculty of Industrial and System Engineering
Telkom University
Bandung, Indonesia
delayoulina.putri@gmail.com

Qalbinuril Setyani
Faculty of Industrial and System Engineering
Telkom University
Bandung, Indonesia
qalbins@gmail.com

***Abstract*— The main objective of a higher education institution is to provide quality education for its students. The most important indicator to measure the quality of higher education performance is the percentage of student graduation on time. However, not all student can successfully have completed their studies during the four years of normal study period where it became problems for academic planners. So, it can affect to the study program accreditation assessment. In this study, C4.5 algorithms and fuzzy AHP are used to predict the number of students graduating on time. An analysis has conducted on how students can graduate on time and plan strategies for groups of students who are likely not to graduate on time. Furthermore, a comparative analysis of the algorithms that have been implemented which will provide more precise and accurate results. Data processing is carried out using the Rapidminer application. The results of the student graduation target analysis were found that the main factor of graduation on time using FAHP was the number of repeating courses (do not pass courses), while in C4.5 it was caused by GPA level 2. Both algorithms had a good level of accuracy, where FAHP and C4.5 were 100% and 82.24% respectively. This research can be used as a reference basis for supporting academic planners in making the right decisions for student groups produced so that all students can graduate on time.**

Keywords—data mining, education, c4.5, fuzzy AHP, decision support system.

I. INTRODUCTION

Time-to-degree compliance is an essential key in assessing the performance of a high degree institution. Observing students with longer time to complete their degrees (> 4 years) are very common. This condition affects the departmental assessment during its periodical accreditation audit. Eventually, it could negatively influence the University-level evaluation. Thus, attention is required to improve the rate of achieving the targeted time-to-degree performance continuously. Data mining algorithm may contribute to this field by proposing the prediction model.

The well-known algorithms in predicting the graduation-on-time rate are ID3, CART, naive Bayes, fuzzy AHP (FAHP) and C4.5 algorithms. The latter algorithm shows better performance than the ID3, CART, naive Bayes [1,2]. The C4.5 algorithm is a decision tree algorithm that recursively visits each decision node, choosing optimal branching, until no more branches can be generated [3]. This algorithm is widely implemented in decision support systems in the field of education including research by [1], [4] and [5]. FAHP offers better performance than the conventional AHP. FAHP capable of covering the weaknesses of the AHP method which has problems with criteria that have more subjective criteria [6]. FAHP is a method used to determine the weight of criteria in making decisions with subjective or natural language perceptions [7]. According to [8], FAHP shows how humans think in using the information to estimate uncertainty in producing decisions.

At this current study, we conform with our previous works [9,10] in developing a prediction model of graduation-on-time rate were conducted and evaluated by two promising and proven methods, they are FAHP and C4.5 algorithm. The work would be performed by describing the construction process of the models and followed by the comparative

analysis of those two models. Students' data from the Information System Department of The School of Industrial Engineering, Telkom University was used to build up the model.

II. Related Works

A. Implementation of FAHP and C4.5 Algorithm

Employing the AHP method in the Information System Department of the School of Industrial Engineering, Telkom University was carried out to support the decision analysis in the streaming process of the seniors [11]. As if AHP, in solving problems FAHP also uses hierarchical structures, decomposition and comparison matrices, decreases inconsistency and produces more important vector. The weakness of applying AHP for such case is the limitation in solving problems that have more subjective criteria [6]. This drawback could be overcome by applying the FAHP. A study by [12] describes the implementation of FAHP as a decision support system for the human resources department use. The test and analysis were conducted with 5 fields as objectives, 7 criterions, and 56 alternatives. Eventually, the FAHP approach succeeded in obtaining an accuracy level of 89.28%. This success story would be referred for the current study of applying FAHP in predicting the time-to-degree for university students.

Work on the C4.5 algorithms in predicting the student's performance was conducted by [13]. The study analyzed the entrance examination type, gender, grade of physics, chemistry and mathematics in class XII. However, this study did not consider the extracurricular activities and other vocational programs of the students, which might have a significant impact on the overall performance of students. Another implementation of the C4.5 algorithms in predicting academic performance was performed by [5]. The study was meant to identify students who need special support in their learning process. In this study, data processed by C4.5 algorithms as many as 120 datasets, with 5 attributes namely student class performance, attendance, tasks, laboratory work and student learning performance. Finally, the algorithm indicates that the poor performance of students in the learning process is the main factor of student failure at the final exam. Similarly, [14] develop a model that can be used to predict student success after graduation by using the C4.5 algorithms. The result emphasizes that the GPA-based model is better at predicting student success as compared to the time-to-degree model. Knowing the promising performance of the C4.5 algorithm in predicting student's academic performance, at the current study, the C4.5 algorithms would be utilized to predict the graduation time of the university students.

B. Case Study

Telkom University is a renowned private higher education institution in Indonesia. It comprises of 7 schools with approximately 30,000 students. Though the university was established in 2013, the historical achievements of this institution was dated back to the 1990 when it was named as STT Telkom. The school was a telecommunication-based vocational institution backed by the PT.Telkom Indonesia (the largest telecommunication service provider in Indonesia).

Information System department is part of the School of Industrial Engineering, one of the pioneer schools in the Telkom University. The department receives an A accreditation from the BAN-PT (Indonesian Higher Education Accreditation Body). Yearly, Information System department enrolls for about 400 students. The current study would use the students' information of the Information System Department for constructing the prediction model.

Academic record of the students was collected from the *i-gracias*, an information system created and utilized by the Telkom University. There were 653 students were enrolled in the Information System Department between 2009-2012. Among those enrolled students, there were 99 (15.16%) students were unable to complete the study due to many reasons. Number of recorded graduates were 486 students (74.43%). It comprised of 390 (59.72%) were graduated on-time (4 years) and the remaining graduates (14.70%) were graduated later than 4 years. Refer to Table I for the details information of the graduates.

III. Research Methodology

In conducting the research, we devided into 3 main stages, as described in figure 1. In the initialization phase, we identified problems regarding the target of graduation on time. Then proceed to determine the research objectives and problems limitation that occur and determine the literature study used in this study. In the data processing phase, the first thing to do is collect data, determine attributes related to student graduation and perform data cleansing. After that, data processing is done using a predetermined algorithm, namely C4.5 and FAHP. In the Fuzzy AHP method, the researcher transforms the pairwise comparison matrix scale with the TFN scale against the results of the previous stage. Then determine the fuzzy priority synthesis value which is continued by looking for vector values and defuzzification.

The final step is to normalize the value of the previous vector and determine the weight of the value. After weighting the values, the researcher can determine the factors that influence the graduation on time. Whereas the application of C4.5 algorithm is divided into two parts, namely training set and testing set. The training set begins with calculating the gain ratio for all the attributes that have been determined, then selecting the attribute with the highest gain ratio to be used as a node. After that, the gain ratio calculation process and the node containing the attribute is repeated until all the attributes are used up. At the end of this phase, the performance of each algorithm is measured. Based on these results, an evaluation and conclusions and suggestions are made in accordance with the results of this study.

This research are implemented using the RapidMiner to process data and perform algorithm performance evaluations. RapidMiner is an application that helps in the process of making artificial intelligence that can be used by companies through a data science platform. This application is built for the analysis team, RapidMiner is an open source software platform for analyzing data mining, text mining and predictive analysis that can be operated on a variety of operating systems [16]

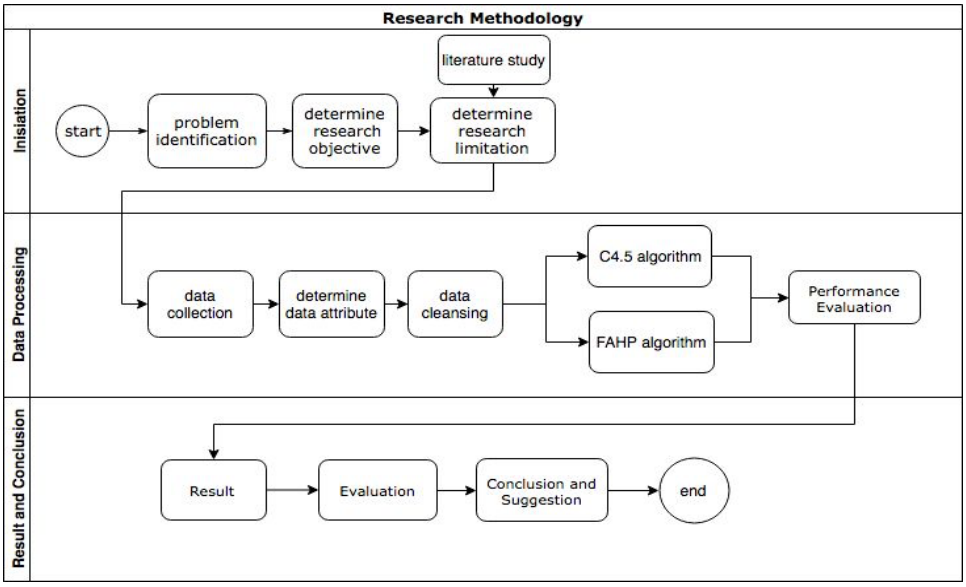

Fig. 1. Research Methodology

IV. Constructing Prediction of Graduation Time

Student academic record data is obtained from academic information systems. The data contains student data between 2009-2012 totaling 653 students. The following is a graduation table of the student (Table I).

The number of graduates from the 2009-2012 class year there were 486 students with 390 students graduated on time and 96 students graduated not on time. Furthermore, some attributes which support the graduation of students including Grade-Point Average (GPA) of level 2 (C1), Number of Re-take class (C2), High School type (C3), Parents' Job (C4), Parents' Income (C5), Entrance Path (C6), and Student Academic Transcripts (SAT) grade (C7). The next step is to do a data cleansing process by using Pentaho Data Integration (PDI). At this stage, the data will be transformed into a format that matches the research input. This process includes deleting incomplete data, completing data according to justification, adjusting category, and converting attribute data. This process can be seen in Figure 3.

TABLE I. STUDENT GRADUATION STATISTICS

Academic year	Total student	Graduate			Non-retained student
		On time (4 years)	Late (more than 4 years)	Total	
2009	153	77	43	120	33
2010	127	82	23	105	16
2011	166	104	24	128	21
2012	207	127	6	133	29
Total	653	390	96	486	99

A. *Fuzzy-AHP*

The data processing was then carried out by using the fuzzy AHP (FAHP) method. This method is in accordance with the step of AHP method, which requires a criteria hierarchy structure from the top level to the lowest level. Where the top level is the goal (goals), then followed by the criteria used for decision making. In this case, the level of criteria used was only 1 level with no sub criteria. To weight the criteria above, use pairwise comparison matrix in the AHP method. This matrix is a comparison of n x n (calculation among criteria) to see the relationship between each criterion. In addition, it also sees the consistency of criteria based on subjective values that have been given.

Because the input matrix questionnaire has 3 questionnaires, it is merged into one and divided according to the number of questionnaires. The next step was to normalize matrix using average and calculate the eigenvalues, in accordance with previous research [10].

The maximum lamda value obtained was 7.443. Furthermore, we calculated the value of Consistency Index (CI), Random Index (RI), and Consistency Ratio (CR). The calculation results of CI, RI, and CR were 0.0738, 1.32, and 0.0559. Because the CR value obtained (0.0559) is less than 10%, the subjective view of the comparison on the criteria of students graduating on time can be accepted, and the matrix is stated to be consistent.The next stage was fuzzy AHP performed by calculating the value of the Triangular Fuzzy Number (TFN) scale, calculating the value of Fuzzy priority synthesis, calculating the minimum value, normalizing and calculating priority weights. The minimum value of each criteria can be seen in Table II. After obtaining the minimum value, it was normalized on the fuzzy value so that the priority weight can be calculated. The followings are the results of the calculations with priority weights (Table III).

TABLE II. MINIMUM VALUE OF FUZZY

Criteria	**Minimum**
C1	0,9052
C2	1,0000
C3	0,6331
C4	0,0276
C5	0,0323
C6	0,0000
C7	0,2656
Total	**2,8637**

TABLE III. NORMALIZATION AND PRIORITY WEIGHT

Criteria	**Priority Weight**	**Rank**
C1	0,3161	2
C2	0,3492	1
C3	0,2211	3
C4	0,0096	6
C5	0,0113	5
C6	0,0000	7
C7	0,0927	4

Based on Table III, the results of the calculation using the FAHP method found that the number of re-take class was in rank 1 with a priority weight of 0.3492. The second rank was GPA of Level 2 with a priority weight of 0.3161, and the third was High School type with a priority weight of 0.2211. The rating shows the criteria that influence the time-to degree compliance.

The accuracy rank test was done by comparing the results of ranking using AHP and FAHP. The results on the test of accuracy level performed by using formula (2) as shown in Table IV below:

TABLE IV. ACCURACY MEASUREMENT

Comparison	**RANK**						
AHP	C2	C1	C3	C7	C4	C5	C6
FAHP	C2	C1	C3	C7	C4	C5	C6
	1	1	1	1	1	1	1

Based on the tests above, the accuracy level of the two methods is identical. The last row shows the result of accuracy for each criteria. The similarity of the two methods is 100**%**.

$$Accuracy\ (\%) = \frac{tp + tn}{tp + tn + fp + fn} x\ 100\% \quad \ldots (1)$$

Description:

tp = true positive, positive proportion in the data set classified positive

fn = false negative, positive proportion in the data set classified negative

fp = false positive, negative proportion in a data set that is classified as positive

tn = true negative, negative proportion in the data set that is classified as negative

B. C4.5 Algorithm

The C4.5 algorithm implementation began with calculating the total number of cases and the number of cases for late time-to-degree along with the entropy values for each attribute used. Then we calculated the gain ratio for each attribute.

Furthermore, it can be seen that the attribute with the highest gain ratio is the GPA of Level 2 with a value of 0.129. Thus, this attribute can be a root node. There are five attribute values at the GPA of Level 2, namely A, AB, B, BC, and C. Based on the five attributes, attribute A has classified the case, namely the decision to graduate on time while attribute C classifies the case into a late graduation decision so that no further calculation is necessary.

The attributes that have been classified in a particular decision form a rule, but the three other attribute values, namely AB, B, and BC, still need to be calculated again because the case is still not classified. Then we need to do the same thing repeatedly to find the next node by calculating the gain ratio value for each remaining attribute until all attributes are used up and all cases are classified according to the class, as was done in previous research [9]. The overall classification results can be seen in Figure 2.

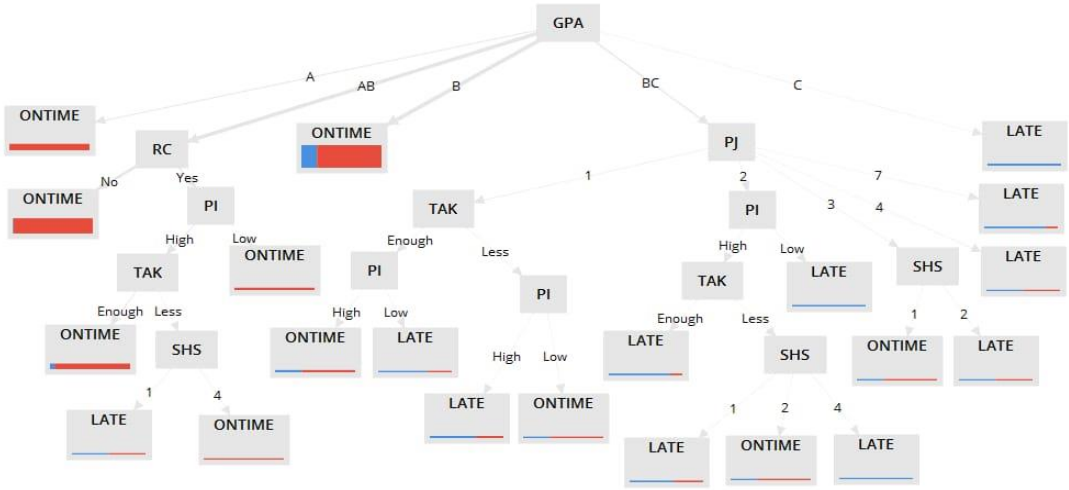

Fig. 2. Decision Tree of Graduation on Time [9]

Furthermore, the implementation process of the prediction model produced from the previous stage on the student data from 2013 to 2014 was 458 records. In this study, the method used to evaluate the performance of data classification done previously was confusion matrix model. Classification performance evaluation process was obtained from the testing phase. This testing stage will perform testing set of 20% which has been obtained from 532 records by using the classification rule model that has been obtained, determining the output class label based on the existing rules, and then comparing the classification results label obtained with the actual data class label. Test results conducted on testing sets are presented in the following Table V which contains the actual class labels and prediction class labels.

TABLE V. PERFORMANCE EVALUATION RESULT

	True Ontime	True Late	Class Precision
Prediction Ontime	81	12	87.10 %
Prediction Late	7	7	50.00 %
Class Recall	92.05 %	36.84 %	

According to Table V, it can be calculated that the accuracy value of the prediction model obtained by using formula (2) is 82.24%. Based on these calculations, it can be known that the accuracy of classification results model in predicting data classes was indicated by the precision value of each class. Then the success rate of the classification result model in classifying data correctly was also known, which was indicated by the recall value

C. Comparison Analysis

The results obtained from data processing to predict graduation are on predictions using fuzzy AHP. The three main factors determining the graduation on time are the number of Re-take class, GPA of Level 2, and high school type while the prediction using the C4.5 decision tree shows that the three main factors are GPA of level 2, number of re-take class, and high school type. These results indicate a difference for the main factors that influence, where Fuzzy AHP showed re-take class, while C4.5 showed GPA level 2. Based on previous step shows that fuzzy AHP approach considering the results of questionnaires from expert judgment has significant accuracy compared to the classification of decision trees. This provides an illustration that time-to-graduate predictions using justification from stakeholders directly involved in operations can be used as a reference in decision support.

V. CONCLUSION

Fuzzy Analytical Hierarchy Process (FAHP) method can be implemented in determining the factor of students' timely graduation. The results of ranking accuracy testing based on the calculation comparison by using the AHP method and by using the FAHP method are at 100%. Therefore it can be ascertained that the test of ranking results between the two methods has the same results while the prediction of student graduation used the decision tree classification method with C4.5 algorithm. The classification model formed has an accuracy value of 82.24% and the precision value in determining the class on time of 87.10%, so the C4.5 algorithm is good enough to be implemented to predict students' timely graduation.

The results of the calculation among criteria produce the criteria for the Number of Re-take class with a weight of 0.3492 as the highest assessment weight. Thus the weight reaches the goal of the student graduation factor which means that it can influence student graduation. At the same time, the C4.5 algorithm predictor attributes used in the study have an influence in predicting student's timely graduation where GPA of Level 2 are the most influencing factor because these attributes are the root nodes of the decision tree produced.

REFERENCES

[1] Yadav, S., & Pal, S. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. World of Computer Science and Information Technology Journal (WCSIT), 51-56.

[2] Risqiati, & Ismanto, B. (2017). Analisis Komparasi Algoritma Naive Bayes Dan C4-5 Untuk Waktu Kelulusan Mahasiswa. IC-Tech Jurnal STMIK WP, XII, 33-38

[3] Rahmayuni, I. (2014). Perbandingan Performansi Algoritma C4.5 dan CART Dalam Klasifiksi Data Nilai Mahasiswa Prodi Teknik Komputer Politeknik Negeri Padang. Jurnal TEKNOIF, 2, 40-46.

[4] Kamagi, D., & Hansun, S. (2014). Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa. Ultimatics, VI, 15-20

[5] Guleria, P., Thakur, N., & Sood, M. (2014). Predicting Student Performance Using Decision Tree Classifiers and Information Gain. International Conference on Parallel, Distributed and Grid Computing, 126-129.

[6] Jasril, E. Haerani, I. Afrianty. Sistem Pendukung Keputusan (SPK) Pemilihan Karyawan Terbaik Menggunakan Metode Fuzzy AHP (F-AHP), Seminar Nasional Aplikasi Teknologi Informasi, 2011.

[7] Y. C. Chen, Vivien & Hui-Pang, Lien & Liu, Chui-Hua & Liou, James & Tzeng, Gwo-Hshiung & Yang, Lung-Shih. (2011). Fuzzy MCDM approach for selecting the best environment-watershed plan. Appl. Soft Comput.11. 265-275.

[8] Kahraman, Cengiz, Ufuk Cebeci, and Da Ruan. 2004. "Multi-Attribute Comparison of Catering Service Companies Using Fuzzy AHP: The Case of Turkey." International Journal of Production Economics 87: 171–84.

[9] Putri, D.Y, Andreswari, R & Hasibuan, M. (2018). Analysis of Students Graduation Target Based On Academic Data Record Using C4.5 Algorithm Case Study: Information Systems Students of Telkom University. The 6th International Conference on Information Technology for Cyber and IT Service Management (CITSM 2018)

[10] Setyani, Q., Andreswari, R., & Hasibuan, M. Target Analysis of Students Based On Academic Data Record Using Method Fuzzy Analytical Hierarchy Process (FAHP) Case Study: Study Program Information Systems Telkom University. The 6th International Conference on Information Technology for Cyber and IT Service Management (CITSM 2018)

[11] Febryanti, A.C, Darmawan, I, and Andreswari, R. 2017. "Pembobotan Kriteria Sistem Pendukung Keputusan Pemilihan Bidang Peminatan Menggunakan Metode Analytic Hierarchy Process (Studi Kasus : Program Studi Sistem Informasi Universitas Telkom)." 3: 7–15.

[12] Norhikmah, Rumini and Henderi. 2013. Metode Fuzzy Ahp Dan Ahp Dalam Penerapan Sistem Pendukung Keputusan. Seminar Nasional Teknologi Informasi dan Multimedia. 09-32.

[13] Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting Students' Performance Using ID3 And C4.5 Classification Algorithms. International Journal of Data Mining & Knowledge Management Process (IJDKP), 39-52.

[14] Dragičević, M., Bach, M., & Šimičević, V. (2014). Improving University Operations with Data Mining: Predicting Student Performance. International Journal of Economics and Management Engineering, 1101-1106.

[15] RapidMiner Inc. (2017). RapidMiner Platform. Retrieved December 5, 2017, dari Data Science Platform: https://rapidminer.com/products//