International Research Journal of Advanced Engineering and Science

ISSN (Online): 2455-9024

C4.5 Algorithm Application for Student Graduation Data Analysis of Darma Persada University

Sheila Pramita Hervianti[1], Emy Haryatmi[2]

[1]Department Computer Science, Gunadarma University, Jakarta, Indonesia, 16424

[2]Department Computer Science, Gunadarma University, Jakarta, Indonesia, 16424

Abstract— *Every time entering a new academic year the quota of students and student enthusiasts who register at Darma Persada University increases, but not all students can graduate on time according to the period of study they are taking, resulting in an increasing number of students. If processed using certain methods, this large amount of data will provide new information that can assist higher education institutions in making decisions or policies, such as the classification method using the C4.5 algorithm. This study aims to determine the application of the classification method with the C4.5 algorithm and to analyze the results of the C4.5 algorithm on student graduation data. The variables analyzed were Major, Age, Gender, Semester 1 to 4 Semester Achievement Index (IPS), Cumulative Achievement Index (GPA), Semester Credit Units (SKS), and Graduation Status. The analysis was carried out using WEKA software. The stages of the research method used CRIPS-DM. From the three testing data, training and testing data, high accuracy values are obtained at 90% of training data and 10% of testing data. The results of the application of the algorithm have an accuracy of 93.75% and an AUC value of 0.965 which is categorized as very good classification. Input variables that significantly influence student performance prediction among other credits, GPA, and IP Semester 1 Semester 2.*

Keywords—*Classification Method, C4.5, Student Graduation, WEKA.*

I. Introduction

In the world of education, students are an important resource and asset in educational institutions. Public assessment is usually based on the accuracy of graduating from students or students of an educational institution so that it affects the level of credibility and existence of the institution in accordance with the regulations presented in book II of standards and procedures regarding the accreditation of higher education institutions by BAN-PT (National Accreditation Board for Higher Education) in 2011 which states that one of the aspects of accreditation assessment is students and graduates [BAN-PT, 2011].

Darma Persada University is one of the private universities in Indonesia which has 15 study programs [DIKTI forlap, 2017]. From the information obtained, it is known that every time entering a new academic year the quota of students and student enthusiasts who register increases, but not all students can graduate on time according to the study period taken, resulting in the number of students increasing. If processed using certain methods, this large amount of data will provide new information that can assist higher education institutions in making decisions or policies, such as the classification method using the C4.5 algorithm. This study aims to determine the application of the classification method with the C4.5 algorithm and to analyze the results of the C4.5 algorithm on student graduation data. The variables analyzed were major, age, gender, Semester 1 to 4 semester Achievement Index (IPS), Cumulative Achievement Index (GPA), Semester Credit Units (SKS), and graduation status. Analysis was carried out using WEKA software.

II. Methodology

A. Data Mining

Data mining has long roots in fields of science such as artificial intelligence, machine learning, statistics, databases and also information retrieval [Marselina et al., 2010].

B. Data Mining Stages with CRISP-DM

CRISP-DM provides a standard process data mining as a general problem-solving strategy of business or research units [Larose, 2006], namely the Business Understanding Phase, the Data Understanding Phase, the Data Preparation Phase, the Modeling Phase, the Evaluation Phase, and the Deployment Phase.

C. Classification Method Classification

Classification is one of the algorithms data mining, using data with a target in the form of a nominal value. The classification is based on four fundamental components [Gorunescu, 2011], namely Class, Predictor, training dataset, Dataset Testing.

D. Algorithm C4.5

Algorithm C4.5 is an algorithm that is included in data mining data classification type using decision tree techniques as a tool for decision making. In this study, the C4.5 algorithm is used in building a decision tree.

E. Confusion Matrix

Evaluation using a confusion *matrix* yields values *accuracy*, *precision*, and *recall*. The Value *accuracy* is the percentage of the number of data records classified correctly by an algorithm. The Value *precision* or also known as *confidence* is the proportion of the number of positive predicted cases that are also true positives in the actual data. Meanwhile, the *recall* or *sensitivity value* is the proportion of the number of positive cases that are actually predicted to be positive correctly. According to Ian H. Witten, it can also be explained that *confusion matrix* is a method for evaluation that uses a table *matrix* as in table 1 [Ian H. Witten, 2005].

TABLE 1. Confusion Matrix

Correct Classification	Classification as	
	+	-
+	True Positives	False Negatives
-	False Positives	True Negatives

F. AUC Value

ROC curves are used in machine learning and data mining research to assess predictive results. The ROC curve is divided into two dimensions, where the TP level is plotted on the Y axis and the FP level is plotted on the X axis. Meanwhile, to represent a graphic which determines which classification is better, a method that calculates the area under the ROC curve is used called AUC (Area Under the ROC Curve). For data mining classification, the AUC value can be divided into several groups [Gorunescu, 2011].

TABLE 2. Standard AUC Value

Range AUC	Quality AUC
0.90 - 1.00	Excellent
0.80 - 0.90	Good
0.70 - 0.80	Fair
0.60 - 0.70	Poor
0.50 - 0.60	Failure

III. Research Results

A. Implementation of the C4.5 Algorithm

In table 3 the comparison of the results of the three testing algorithms of c4.5 which has the greatest testing data accuracy value is in the percentage of training data and testing data with an accuracy of 95.8%, namely a comparison of 90% for training data and 10 % for testing data. Furthermore, the value *precision* explains the level of accuracy between the information requested by the user and the answer given by the system as well as in the third test, which is 96.5%, the value *recall* which explains the success rate of the system in finding back information is 95.9% and *f-measure recall* which is the weight of the harmonic *mean* from *recall* and precision, the value is 96.1%.

TABLE 3. Comparison of Results 3 Testing Algorithm C4.5

Percentage of Data	Accuracy Data Training	Accuracy Data Testing	*Precision* Data Testing	*Recall* Data Testing	*F-measure* Data Testing
70:30	92.4%	92.3%	92.4%	92.4%	92, 4%
80:20	91.7%	93.8%	94.7%	93.8%	94.2%
90:10	91.7%	95.8%	96.5%	95.9%	96.1%

The results of the comparison of the three testing values of training and different testing values can be concluded that the C4.5 algorithm has high accuracy at the percentage of 90% training data and 10% testing data. That way, the visual *decision tree* used for prediction is in test III as in Figure 1.

B. Accuracy Value and AUC Value C4.5

TABLE 4. Accuracy and AUC values of C4.5

Algorithm	Accuracy Value	AUC value of
C4.5	93.75%	0.965

Table 4 shows the Accuracy and AUC values of the C.45 algorithm and the average value of the calculation results from training data and testing data is taken. It can be seen that the C4.5 Algorithm method has an accuracy value of 93.75% and an AUC value of 0.965 so that the application of the C4.5 algorithm on graduation data is classified as very good, because it has an AUC value between 0.90 - 1.00.

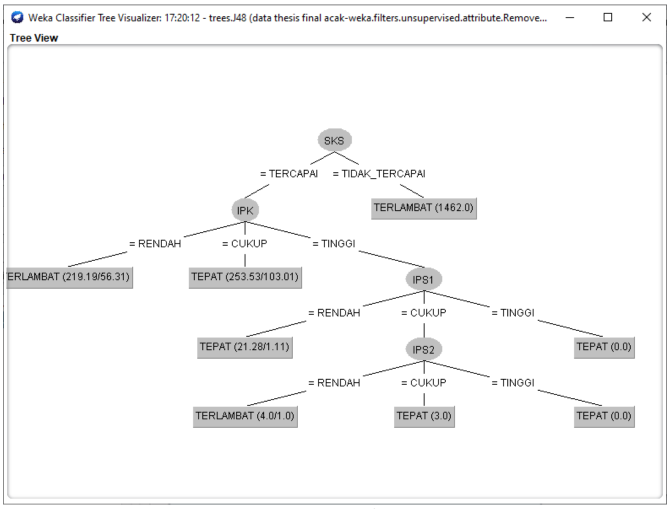

Fig. 1. Decision Tree Pattern Accuracy of Student Graduation Using Data Testing

IV. Conclusion

Based on the results of research conducted by testing three times on the six attributes used in the classification process sequentially, namely education, training, experience, expertise, skills and prediction results, the following conclusions can be drawn:

1. After 3 tests were carried out on the C4.5 algorithm obtained the percentage results for determining training data and testing data with high accuracy values, testing III has high accuracy values for each algorithm. Testing III has 90% percent of training data and 10% training data.
2. The results of the C4.5 algorithm produce an accuracy value of 93.75%, a precision of 96.5% and a *recall* of 95.9%.
3. In prediction accuracy, there are more student data who have prediction results false positive, and only a few students who have negative prediction errors (false negative) and the AUC value of the C4.5 algorithm is 0.965 so that the C4.5 algorithm included in the very good classification, because it has an AUC value between 0.90 - 1.00 including the very good classification category. Therefore, the application of the C4.5 algorithm is very suitable to be used to assess the accuracy of student datasets, which means that the system built using the C4.5 method is good enough to provide early detection as a warning to study program managers regarding student performance.
4. Input variables that significantly influence student performance prediction among other credits, GPA, IP Semester 1, IP Semester 2. The variables Sex, age, IP Semester 3, IP Semester 4 no significant effect on the response variable.

V. SUGGESTIONS

This research cannot be said to be perfect, there are still many things that can be explored deeper from this research, including:

1. Future research development needs to be done with a relatively large dataset so that the level of accuracy of the model's performance is guaranteed considering that this research only uses student data from the year 2009 to 2016 and the latest data.
2. It is necessary to do further research on the method of selecting predictor variables in more detail and adding more parameter attributes such as student entry paths, grades for each semester, and the number of credits taken each semester.
3. Determination of training data and testing data on the WEKA application can be done by testing using the *n-fold cross validation* technique or split.

ACKNOWLEDGMENT

I want to say thank you to Dr. Emy Haryatmi S.Kom., MEngSc., MT, as The Thesis Supervisor who patiently gave direction guidance, motivation sincerely took the time to help the author in compiling this research.

I also want to say thank you to my family and friends who have always provided support and encouragement to me in completing this research.

REFERENCES

[1] B. A. N. Indonesia, (2017). “Peraturan Badan Akreditasi Nasional PT No 4/2017,”.

[2] Forlap DIKTI (2017). Pangkalan Data Pendidikan Tinggi Kementrian Riset, Teknologi dan Pendidikan Tinggi. [Online]. Available at: https://forlap.ristekdikti.go.id/, [Accessed 2 Januari 2018].

[3] Gorunescu, F. (2011). Data Mining Concept Model Technique. Romania: Springer.

[4] Ian H. Witten, Eibe Frank dan Mark A. Hall. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan kaufman, Burlington, MA, 3 edition.

[5] Larose, Daniel (2006). Discovering Knowledge in Data: An Introduction to Data mining,: Jhon Wiley & Son, Inc., USA.

[6] Marselina, Rosy, (2010). Evaluasi Kuantitatif Efektifitas Hasil Pencarian Dokumen Dengan Menggunakan Jaccard Coeficient : Suatu Studi Kasus Penerapan Stemmed Term Vektor Model Untuk Representasi Dokumen Yang Menggunakan Bobot Jumlah Satu Term Didalam Satu Dokumen pada Mesin Pencari Berbasis HTML.