Graduation Prediction of S1 Industrial Engineering Students IST AKPRIND by Using Data Mining Method

1) Lecturer of Industrial Engineering, Technology Industri Departement, Institut Sains & Technology AKPRIND, Yogyakarta, Indonesia

Correponding email [1]:
bagushind@akprind.ac.id

Agus Hindarto Wibowo [1]

Abstract. There is data of students who experience Drop Out which raises the curiosity in IST AKPRIND's industrial engineering study program on students' graduation patterns. It is necessary to have research on how to classify the data held by industrial engineering study programs in order to obtain students' graduation patterns as evaluation material in the administration of study programs. This study also produced a design to set the goals of Educational Data Mining, this case as a student modeling that would be achieved by predicting using the Decision Tree method. The final results showed a mismatch between the general information data passed and the drop out of the rule obtained using the decision tree algorithm in the Rapidminer software which is shown by an accuracy of 95.83%. This value indicates that there is a match between the prediction of student identity data with the rule obtained using the decision tree algorithm.

1. INTRODUCTION

Students at the tertiary level of education whose rights are regulated in Law no. 20 of 2003 Chapter V concerning Students. universities should have students' data stored in information systems. The data consists of students' registration data, students' academic data for each semester to students' graduation data. After students graduate, these data tend not to be used optimally. Therefore, these data need to be utilized to obtain deep information. However, it is not easy to make predictions by utilizing various raw data held by the institute, so Educational Data Mining techniques are necessary to help transform raw data from the system into information that has the potential to have a positive impact on education [1]. Data Mining Education is a sub-area of Data Mining Domain. This new area has great potential to mine various aspects for improving the student's quality as well as in decision making by educational institution authorities [2]. "Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to understand the students better, and the settings which they learn in." [3]

Every educational institution would want to contribute to enhancing the education world, especially in S1 Industrial Engineering Study Program, Faculty of Industrial Technology IST AKPRIND Yogyakarta. The normal time of graduation for a Bachelor's degree is a maximum of 7 years, but not a few students take education by graduating more than 7 years. Hence, in order to improve its quality, study programs issued a policy regarding the maximum graduation limit, since the level of graduation delays can be reduced, but there are students who must Drop Out because they cannot complete their studies. The proportion of students who experience late graduation, until the emergence of students who drop out, encourages research to predict student graduation patterns related to the study period and student performance as evaluation material.

Educational Data Mining has a variety of purposes. Some studies use EDM to predict academic patterns by reviewing the accuracy of the study period of students in their education such as research conducted by [4], [5], [2], [6], [7], [8], [9]. In addition there are studies with the aim of predicting academic patterns by reviewing student performance such as research conducted by [10], [11], [12], [13]. In addition, there are also studies that aim to predict students 'academic patterns based on students' performance on the experiments or scenarios provided as research conducted [14] such as research to predict academic patterns by conducting 3 experiments.

Similar research was also conducted by [15] to predict students' performance by using blended learning scenarios.

Based on the description above, this study aims to examine the data held by the study program to predict students' graduation as an evaluation material for S1 Industrial Engineering Study Program in IST AKPRIND.

2. METHODS
2.1 Data Mining
Data mining is a data collection technique obtained from various sources which is then transformed into very useful information using various predetermined methods. Data mining is a field of several scientific fields that unites techniques from machine learning, pattern recognition, statistics, databases, and visualizations for handling the problem of retrieving information from large databases [16]. In general, data mining can be grouped into 2 main categories, such as descriptive mining and predictive [17]. In doing predictive, data processing techniques that already exist are needed, in order to be collected and processed. The most popular technique used is Data Mining [18]. Several methods in data mining techniques have been used to predict patterns, in this case predict student graduation. According to [18], data mining is the most popular technique used in the last ten years from 2000 to 2011. information in data mining has a different type [19].

2.2 Classification
Data mining classification is the placement of objects into one of several predetermined categories. Classification is widely used to predict classes on a particular label, namely by classifying data (building models) based on training sets and values (class labels) in classifying certain attributes and using them in classifying new data (testing sets) [20].

The data used as input is 360 identity data of S1 IST AKPRIND industrial engineering students regarding 5 variables that are considered to have an influence on student graduation. From 360 criminal report data will be divided into two types of data, such as training set and testing set. Many previous studies have explained the ratio used in determining training sets and testing sets. As much as 60% of all data is used for training data which will result in a rule. The remaining 40% is used as testing data [20].

2.3 Algoritma Decision Tree
Decision Tree (decision tree) is a tree that is in the analysis of problem solving, mapping about alternative solutions to the problem that can be taken from the problem. Decision trees can also be called one of the most popular classification algorithms because they are easy to interpret. The concept of a decision tree is to convert data into decision trees and decision rules. Decision tree is suitable for cases where the output is a discrete value. The main benefit of using a decision tree is its ability to examine and describe complex decision in order to be simpler and easier to interpret the problem solving.

Decision trees are usually used to obtain information for making a decision. The decision tree starts with a root node (starting point) used by the user to take action. Based on this root node, the user solves leaf nodes according to the decision tree algorithm. The final result of composing the root node and leaf node is a decision tree with each branch showing possible scenarios of the decision taken and the results. The concept of a decision tree is to change data into a decision tree (decision tree) and decision rules (rule).

2.4 Research Object
This research was conducted by reviewing and classifying attributes or variables that could affect student graduation rates. The method used in this research is data mining classification techniques with the decision tree algorithm method. The results of the research to be achieved is the rule or the rules of graduate or drop out based on the attributes or variables. Data processing is done with the help of Ms. software. Excel, and Rapidminer. In this preliminary study there were 360 data on the identity of S1 IST AKPRIND industrial engineering students in 2006-2015 which were used as objects in this study.

2.5 Collect Data Method
Quantitative research is research that is intended to obtain data in the form of numbers or qualitative data that has been framed. In this preliminary study, the data obtained are attribute or variable data that might influence student graduation, such as the home province, gender, entry point, GPA, and education background.

2.6 Data Type
Secondary data is data obtained or collected by people who conduct research from existing sources. In this preliminary study, secondary data used are 360 attribute or variable data from student's identity data, such as the home province, gender, entry point, GPA, and education background Other secondary data in this study are previous research related to the problems, such as data mining classification techniques, decision tree algorithm methods with problems related to EDM.

2.7 Flowchart
Figure 1 shows a flowchart of the entire research stage.

```
                          ┌──────────────┐
                          │    Start     │
                          └──────────────┘
                                 │
              ┌──────────────────────────────────┐
             /  1. Student identity data          /
            /   sourced from IST AKPRIND          /
           /    portal                           /
          └──────────────────────────────────┘
                                 │
         ┌─────────────────────────────────────────┐
         │ 1. Province origin, 2. Gender, 3. kind of pathway, 4. GPA, 5. │
         │      Kind of school origin              │
         └─────────────────────────────────────────┘
                                 │
              ┌────────────────────────────────┐
              │   *Preprocessing data*          │
              │   *1. Data Cleaning*            │
              │   *2. Transformasi data*        │
              └────────────────────────────────┘
                                 │
              ┌────────────────────────────────┐
              │ **generate a rule of *decision tree*** │
              │ The resulting rule will describe the results of │
              │ combinations between variables  │
              └────────────────────────────────┘
                                 │
           ┌──────────────────────────────────────┐
           │      **Accuracy calculation**         │
           │ The accuracy value is used to find out the difference between │
           │ the existing report data and the results obtained │
           └──────────────────────────────────────┘
                                 │
                          ┌──────────────┐
                          │ analysis of  │
                          │   results    │
                          └──────────────┘
                                 │
                          ┌──────────────┐
                          │ conclusions  │
                          │     and      │
                          │ suggestions  │
                          └──────────────┘
                                 │
                          ┌──────────────┐
                          │    Finish    │
                          └──────────────┘
```
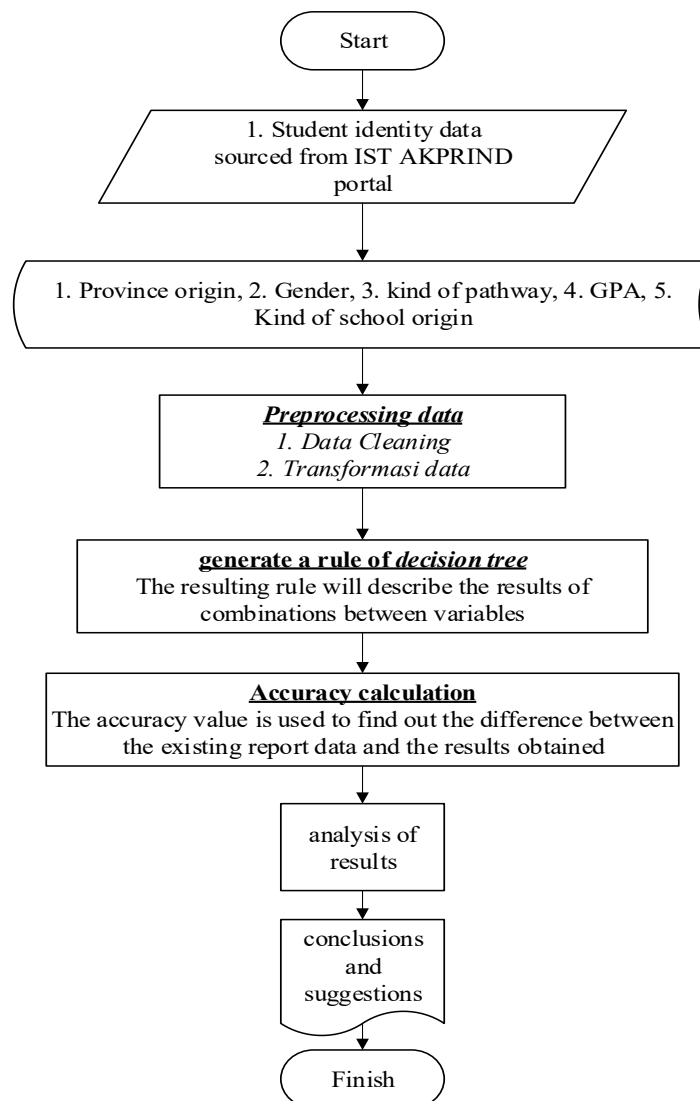
Figure 1. Flowchart diagram

3. RESULTS AND DISCUSSION

The first step was collecting data, such as administrative data of S1 Industrial Engineering students from 2006-2015, from the official website of the IST AKPRIND portal. After the data were obtained, it is necessary to do a selection and transformation of the data. Data selection was condducted to make it more efficient during the classification process, while data transformation was conducted in order to change the shape of the data to be more suitable and can be processed. However, in this initial study, not all data was obtained, therefore the data selection stage could not be carried out yet. If it is assumed that all data needs to be entered into the mining process, then the data needs to be transformed. The following table 1 explains the transformation rules of the temporary data that has been obtained.

Table 1. Transformation Rule

Atribut	Information
Province origin	1-30 (Babel, Bali, Banten, Bengkulu, DIY, Jabar, Jambi, Jateng, Jatim, Kalbar, Kaltara, Kalteng, Kaltim, Kepri, Lampung, Maluku, NTB, NTT, Papua, Riau, Sulsel, Sulteng, Sumbar, Sumsel, Sumut, Timor Leste, Sulbar, NAD, Gorontalo, Jakarta)
Gender	1 dan 2 (Male (M) dan Female (F))
Kind of pathway	1-8 (Transfer path, Scholarship, Camabapres, Transfer, PMB, Readmisi, Regular, Without test)
GPA	1-8 (0-0.5; 0.6-1; 1.1-1.5; 1.6-2; 2.1-2.5; 2.6-3; 3.1-3.5; 3.6-4)
Kind of school origin	1 dan 2 (Public school dan Privat school)
Status	1 dan 2 (Graduate dan Drop out)
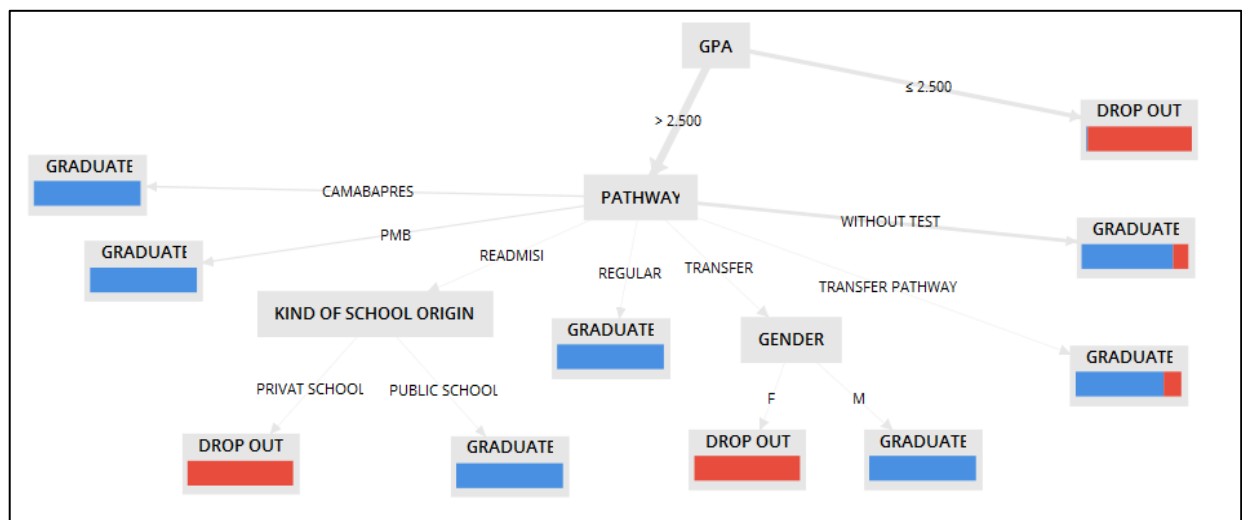
Figure 2. decision tree result

The decision tree above shows that the GPA attribute is the root node since according to manual calculations, the information gain value of the highest GPA attribute from others. The overall picture of the decision tree above can be presented in the form of rules, as follows:

```
GPA > 2.500
|  PATHWAY =  CAMABAPRES: GRADUATE {GRADUATE=30, DROP OUT=0}
|  PATHWAY =  PMB: GRADUATE {GRADUATE=28, DROP OUT=0}
|  PATHWAY =  READMISI
| |   KIND OF SCHOOL ORIGIN = PRIVAT SCHOOL: DROP OUT {GRADUATE=0, DROP OUT=2}
| |   KIND OF SCHOOL ORIGIN = PUBLIC SCHOOL: GRADUATE {GRADUATE=2, DROP OUT=0}
|  PATHWAY =  REGULAR: GRADUATE {GRADUATE=2, DROP OUT=0}
|  PATHWAY =  TRANSFER
| |   GENDER = F: DROP OUT {GRADUATE=0, DROP OUT=2}
| |   GENDER = M: GRADUATE {GRADUATE=5, DROP OUT=0}
|  PATHWAY =  TRANSFER PATHWAY: GRADUATE {GRADUATE=5, DROP OUT=1}
|  PATHWAY =  WITHOUT TEST: GRADUATE {GRADUATE=54, DROP OUT=9}
GPA ≤ 2.500: DROP OUT {GRADUATE=1, DROP OUT=75}
```

The above rule explains one root of the decision tree obtained by GPA, to make it easier to read the rule, for example if the GPA is more than 2.5, the entry point taken is transfer path, the gender is male, so what happens is students' graduate, if the sex is female, what happens is the student drop out (DO).

Table 2. Confusion matrix result

	True graduate	True drop out	Class Precision
Prediction graduate	88	0	100%
Prediction drop out	6	50	89.29%
Class Recall	93.62%	100%	

Accuracy: 95.83%

Based on the table above, it can be seen that the discrepancy between the general information data graduate and drop out to the rules obtained using the decision tree algorithm in the Rapidminer software which is shown by an accuracy of 95.83%. This value indicates that as much as 95.83% of the testing data set there is a match between the prediction of students' identity data with the rule obtained using the decision tree algorithm. Class prediction for graduate prediction of 100% shows the classification prediction obtained from the calculation of the decision tree algorithm classification, then it turns out that the system can perform class recall for the graduate prediction of 93.62%. Class prediction for drop out prediction is 89.29% indicating the classification prediction obtained from the calculation of the decision tree algorithm classification, then it turns out the system can do a class recall for drop out by 100%. The prediction results from the decession tree and the previous rule, can be used to predict the graduation of S1 IST AKPRIND Industrial Engineering students by looking at the results that match the prediction rules and testing sets that already exist.

4. CONCLUSION

Based on the results of research, it can be concluded that the decision tree method can find patterns of students' graduation, and can be used as useful information for the Institute and industrial engineering study programs. The generated rules can be new information that is useful in predicting the graduation of IST AKPRIND industrial engineering students. The use of rapidminer software can present data in the form of a tree and the level of accuracy of the suitability between the report data and the rules obtained using the decision tree algorithm. After obtaining the rule result as above, it is hoped that it can be used as a basis for the Institute and industrial engineering study programs in improving the quality of education.

5. ACKNOWLEDGEMENT

6. REFERENCES

[1]   G. Siemens and R. S. J. d. Baker, "Learning analytics and educational data mining," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, 2012, p. 252.

[2]   S. Hussain, R. Atallah, A. Kamsin, and J. Hazarika, "Classification, Clustering and Association Rule Mining in Educational Datasets Using Data Mining Tools: A Case Study," Springer, Cham, 2019, pp. 196–211.

[3]   R. S. J. d. Baker and K. Yacef, *Journal of Educational Data Mining JEDM.*, vol. 1, no. 1. International Educational Data Mining Soc, 2009.

[4]   G. S. Abu-Oda and A. M. El-Halees, "Data mining in higher education: university student dropout case study," *Data Min. High. Educ. Univ. student dropout case study*, vol. Volume: 5, Number: 1, no. Volume: 5, Number: 1, 2015.

[5]   D. Kusumawati, W. W. Winarno, and M. R. Arief, "Prediksi Kegraduatean Mahasiswa Menggunakan Metode Neural Network Dan Particle Swarm Optimization," *Semnasteknomedia Online*, vol. 3, no. 1, pp. 3-8–37, Feb. 2015.

[6]   R. Ahuja, A. Jha, R. Maurya, and R. Srivastava, "Analysis of Educational Data Mining," Springer, Singapore, 2019, pp. 897–907.

[7]   S. Ameri, M. J. Fard, R. B. Chinnam, and C. K. Reddy, "Survival Analysis based Framework for Early Prediction of Student Dropouts," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16*, 2016, pp. 903–912.

[8]   T. Ojha, G. L. Heileman, M. Martinez-Ramon, and A. Slim, "Prediction of graduation delay based on student performance," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 3454–3460.

[9]   L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting Student Dropout in Higher Education," Jun. 2016.

[10]   A. B. E. D. Ahmed and I. S. Elaraby, "Data Mining: A prediction for Student's Performance Using

Classification Method," *World J. Comput. Appl. Technol.*, vol. 2, no. 2, pp. 43–47, 2014.

[11] 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, Oct. 2017.

[12] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory," *Comput. Human Behav.*, vol. 47, pp. 168–181, Jun. 2015.

[13] G. Thakur, M. Olama, … A. M.-… 20th A. S., and undefined 2014, "Towards adaptive educational assessments: predicting student performance using temporal stability and data analytics in learning management systems," *aspiringminds.com*.

[14] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, "Early dropout prediction using data mining: a case study with high school students," *Expert Syst.*, vol. 33, no. 1, pp. 107–124, Feb. 2016.

[15] N. Z.-I. J. of A. I. and and undefined 2016, "Predicting student academic performance in blended learning using Artificial Neural Networks," *aircconline.com*.

[16] D. T. Larose and C. D. Larose, *Discovering knowledge in data : an introduction to data mining*. .

[17] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Addison Wesley, 2005.

[18] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11303–11311, Sep. 2012.

[19] P. Kaur, M. Singh, and G. S. Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector," *Procedia Comput. Sci.*, vol. 57, pp. 500–508, Jan. 2015.

[20] A. Wibowo and A. Wijaya, "Kajian Awal Profiling Kejahatan Dan Strategi Dalam Usaha Mencegah Terjadinya Tindak Kriminalitas Di Kabupaten Sleman," *idec.ft.uns.ac.id*. 2018.