e-ISSN : 2745-7818
Oral Presentation
THE 1ST SYEDZA SAINTIKA INTERNATIONAL CONFERENCE ON NURSING, MIDWIFERY, MEDICAL LABORATORY TECHNOLOGY, PUBLIC HEALTH, AND HEALTH INFORMATION MANAGEMENT (SeSICNiMPH)

PERFORMANCE ANALYSIS OF DATA MINING CLASSIFICATION METHODS USING C4.5 ALGORITHM FOR STUDENT GRADUATION PREDICTION (CASE STUDY AT SYEDZA SAINTIKA STIKES)

Nurul Abdillah[1*], Fajrilhuda Yuniko[2]
[1,2] STIKes Syedza Saintika

* Corresponding author : Abdillahadik15@gmail.com

ABSTRACT

The passing rate is one of the parameters of the effectiveness of educational institutions. Decreasing student graduation rate affects higher education accreditation. The college database stores student administrative and academic data, if it is explored properly using data mining techniques, it can be seen patterns or knowledge to make decisions. The C4.5 algorithm is an algorithm used to generate decision trees. This method is popular because it is able to classify as well as show the relationship between attributes. This study used data from students of the 2014 and 2015 class of Public Health study programs. The variables used in this study were: NIM, name, gender, entry status, GPA, area of origin and employment status. Based on the test results by measuring the performance of the method, it is known that C4.5 has a high accuracy value of 97.83%. From the accuracy value, it can be concluded that the C4.5 algorithm has a good performance in predicting the timeliness of student graduation.

Keywords: ***DataMining, Classification, Naïve Bayes, Accuracy of Student Graduation***

INTRODUCTION

Graduating on time is an important thing that needs to be addressed wisely by a university. The passing rate is one of the parameters of the effectiveness of educational institutions. Decreasing student graduation rates will affect tertiary accreditation. Therefore, it is necessary to monitor and evaluate the tendency of students to graduate on time or not. The college database stores administrative and academic data of students, if this data is explored properly using data mining techniques, it can be seen patterns or knowledge to make decisions.

The use of the Data Mining classification method to predict the timeliness of student graduation by using the C4.5 algorithm can provide information about the accuracy of

Previous research examines the performance comparison of several Data Mining classification methods by comparing the Decision Tree and Naive Bayes algorithms. This study aims to predict student dropouts. From the results of testing the accuracy using these two algorithms, the student graduation timeliness. Data Mining is the process of analyzing data and summarizing the results into useful information. Technically, Data Mining is a process to find correlation between many fields in a large dataset [1]. Data Mining has several methods, one of which is the classification method which is a learning technique to classify data items into predetermined class labels. The classification method has several algorithms, one of which is C4.5.

The C4.5 algorithm as a classification algorithm in this study is an algorithm used to produce a decision tree [2]. Thus the use of the Data Mining method will provide the best accuracy results in data classification.

highest accuracy is obtained, namely the Decision Tree algorithm [3].

MATERIAL AND METHODS

1. Classification

Classification is a process of finding a model or function that describes or distinguishes a concept or data class with the aim of estimating the class of an object whose label is unknown. This can also be said as learning (classification) which maps an element (item) of data into one of several predefined classes [5].

$$R_{\alpha}(T(\alpha)) = \min_{T < T_{Max}} R_{\alpha}(T)$$

Classification is a technique by looking at the behavior and attributes of a defined group. This technique can classify new data by manipulating existing data that has been classified and by using the results to provide a number of rules. These rules are used in new data to be classified [5].

2. Algorithm C4.5

The C4.5 algorithm is an improved version of ID3. In ID3, Decision Tree induction can only be done on features of categorical type (nominal or ordinal), while numeric types (interval or ratio) cannot be used. The improvement is that it can not only handle categorical type features, but also can handle features with numeric types, and can also prun (pruning) Decision Tree, and deriving rule sets. The C4.5 algorithm also uses the gain criterion in determining the features that break down the nodes in the induced tree [4].

There are several steps in making a decision tree using the C4.5 algorithm [6], namely:

a. Prepare training data.
   The training data is usually from historical data that has happened before and has been grouped into certain classes.
b. Determine the root of the tree
   The root will be taken from the selected attributes by calculating the gain value of each attribute, the highest gain value will be the first root. Before calculating the gain value of the attribute, first calculate the entropy value, namely:

$$\text{Entropy}(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$

   Information:
   S: case set
   n: number of partitions S
   pi: the proportion of Si to S

c. Then calculate the gain value using the formula:

$$\text{Gain}(S, A) = \text{Entropy}(S, A) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * \text{Entropy}(S_i)$$

   Where:
   S = case set
   A = features
   n = number of partitions attribute A
   $|Si|$ = the proportion of Si to S
   $|S|$ = number of cases in S
d. Repeat step 2 until all records are partitioned.
e. The decision tree partitioning process will stop when:
   1) All records in node N are assigned the same class.
   2) There are no attributes in the partitioned record anymore.
   3) There are no records in the empty branch.

RESEARCH METHOD

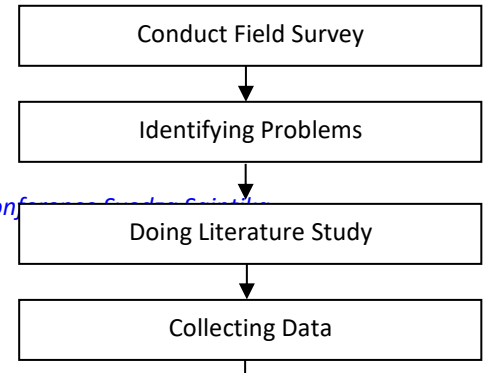

Figure 1. Framework

Based on the framework in Figure 3.1, each of the steps can be described as follows:

1. Conduct Field Survey
   Before starting the research, first conducted a field survey to get a qualitative picture of the accuracy of student graduation at STIKes Syedza Saintika.
2. Identifying Problems
3. The problem identification stage is the stage where the object of research formulates the problem.
4. Doing Literature Study
5. To achieve the goals to be determined, it is necessary to study some of the literature used.
6. Collecting Data
7. The data collection was carried out in several ways, namely:
   a. Direct observation method
   b. Interview method
   c. Literature study method
   d. Browsing method
8. Processing and Data Transformation
   In the Data Processing and Transformation stage, raw data will be converted and combined into a format suitable for processing into Data Mining.
9. Implementing the Method
   After the analysis process, the testing stage is then carried out. In testing computer hardware and software are needed. At this stage the previously proposed method will be implemented, which will be tested using the Rapid Miner Software
10. Calculating Accuracy and Error
    At this stage the Accuracy and Error values of the C4.5 algorithm will be calculated to evaluate the Accuracy and Error values of the measurement against the actual value or the value considered true.
11. Creating Results and Discussion
    Results and discussion aims to provide an overview and the results obtained from this study.

RESULTS

1. Data Mining Analysis
   It is a series of processes that include collecting, using historical data to find regularities, patterns or relationships in large data sets.
2. Data Collection
   In this study, the data used were data from students of the STIKes Syedza Saintika Public Health Study Program in 2014 and 2015. The data used were 46 records.4.2.1
   a. Variable Selection
      From the student data, the decision variable was taken to pass on time and late. Meanwhile, the determining variables in decision making are gender, entry status, GPA, area of origin and work status.
   b. Pre-Process
      After selecting the variables, the data format will be transformed based on the variables that have been selected.
3. Classification Method

The classification results obtained can provide information, regarding the level of accuracy and error on the timing of graduation of STIKes Syedza Saintika students. The use of the C4.5 Algorithm is carried out in several steps to obtain the desired information.

a. Classification Process Using the C4.5 Algorithm

Decision Tree

Table 1. Calculation of Node 1

Node		Jumlah Kasus(S)		Tepat Waktu (S1)	Terlambat (S2)	*Entropy*	*Gain*
1	Total		46	25	21	0.99403	
	*Jenis Kelamin	L	21	8	13	0,9587	0,0653
		P	25	17	8	0,9043	
	*Status Masuk	PM	6	6	0	0	0,1755
		SN	8	4	4	1	
		BM	6	4	2	0,9182	
		JP	8	5	3	0,9544	
		RM	18	6	12	0,9182	
	*IPK	1	7	4	0	0	**0,9945**
		2	10	10	0	0	
		3	18	11	7	0,964	
		4	10	0	10	0	
		5	2	0	2	0	
		6	2	0	2	0	
	*Daerah Asal	D	26	11	15	0,9828	0,0558
		L	20	14	6	0,8812	
	*Status Pekerjaan	BB	36	20	16	0,991	0,0015
		B	10	5	5	1	

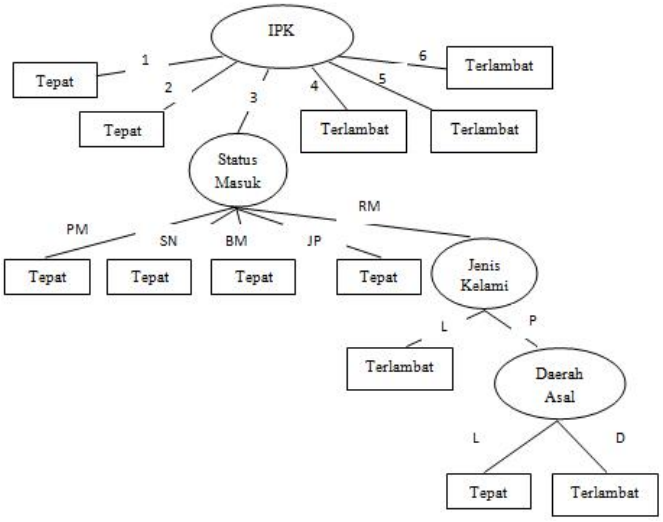

Figure 2. Decision Tree

DISCUSSION

Implementation and results will explain Implementation or testing to find out the results of manual calculations with the results using the supporting software C4.5 algorithm. This aims to see the data that is analyzed and processed is correct or not. The software used is Rapidminer Studio 7.5.3. Rapidminer Studio is an open source Data Mining application. In the case of predicting the timeliness of this student's graduation, the data that will be used on Rapidminer is 92 records.

1. Level of Accuracy and Error C4.5 Algorithm
   a. Accuracy Level C4.5
      In the calculation of the accuracy of C4.5, an accuracy of 97.83% is obtained because it produces 90 data that are classified correctly.

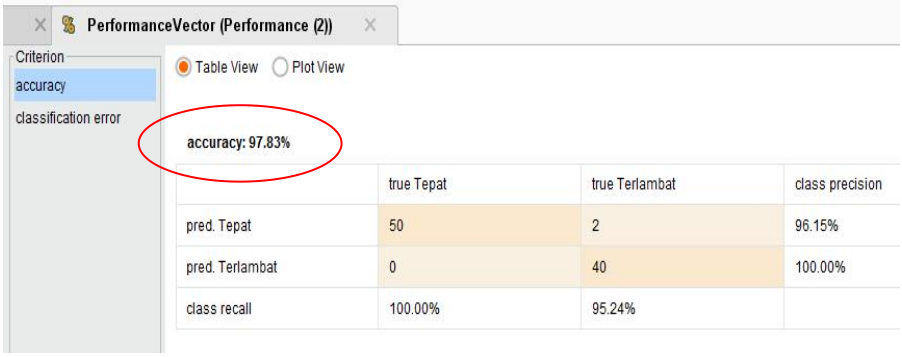

Figure 3. Accuracy of C4.5

   b. Algorithm Error Rate C4.5
      In the calculation of error C4.5, the error value is 2.17% because it produces 2 data that are classified incorrectly.

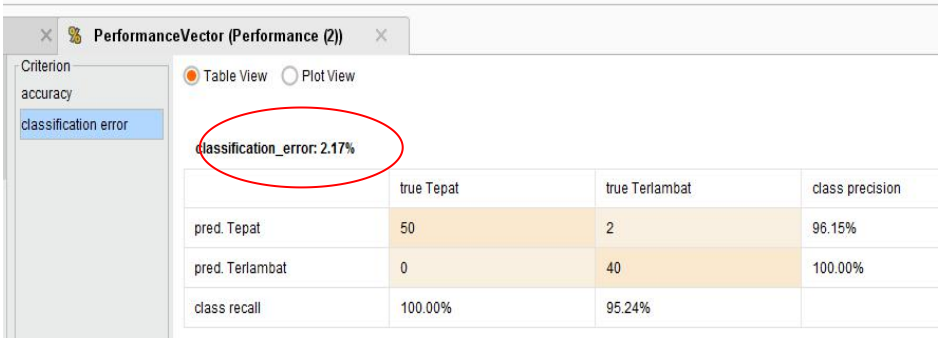

Figure 4. Error C4.5

CONCLUSION

1. Measurement of the accuracy level of the C4.5 classification method produces an accuracy value of 97.83%.
2. From the results of the tests that have been carried out, the C4.5 Algorithm has good performance because the C4.5 Algorithm has a high accuracy value, the higher the accuracy value, the closer to the correct data classification. The C4.5 algorithm also has a low error value, the lower the error value, the closer to the correct classification.

REFERENCES

1. Sumathi K., Kannan S. dan Nagaraan K. 2016, *"Data Mining: Analysis of student database using Classification Techniques"*, International Journal of Computer Applications, Vol. 141, No. 8, Hal. 22-27.
2. Agrawal Gaurav dan Gupta Hitesh 2013. *"Optimization of C4.5 Decision Tree Algorithm for Data Mining Application"*. International Journal of Emerging Technology and Advanced Engineering, Vol.3, Hal. 341-345.
3. S. Ghadeer, Oda Abu dan M. El-Halees Alaa 2015. "*Data Mining In Higher education: University Student Dropout Case Study"*, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, Hal. 15-27.
4. Miftahul Chair, Yuki Novia Nasution dan Nanda Arista Rizki 2017. *"Aplikasi Klasifikasi Algoritma C4.5 (Studi Kasus Masa Studi Mahasiswa Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Mulawarman Angkatan 2008)"*, Jurnal Informatika Mulawarman, Vol. 12, No. 1, Hal. 51-55.
5. Cipta Riang Sari 2016. *"Teknik Data Mining Menggunakan Classification Dalam Sistem Penunjang Keputusan Peminatan SMA Negeri 1 Polewali"*. Indonesian Journal on Networking and Security. 2016; Volume 5: Hal 48-54.
6. Siska Haryati, Aji Sudarsono dan Eko Suryana 2015, "*Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)",* Jurnal Media Infotama Vol. 11 No. 2, Hal. 130-138.