Pattern Analysis To Prediction of Graduation With CPAR Algorithm

Indah D Lestantri
Information System Department
Polman Astra
Jakarta, Indonesia
e-mail: indahdl2005@yahoo.com

***Abstract* — This paper describes the steps to find the connection between subject on academic achievement. The purpose of this paper for it to see the condition of students. Steps undertaken include domain understanding, collect data, preprocessing, selection of data mining, data processing, evaluation and presentation patterns. Prepocessing includes data cleaning, data integration, data selection and data transformation. In one of the data cleaning is done the equation GPA for students who drop out. Integration done by combining all the data. The selection is done by determining the data attributes that will dianalis. And data transformation is done by doing some math operations. This paper also describes how the student graduation prediction using CPAR algorithm. The data is divided into the positive data and negative data. The next step is to calculate the TWT, forming Gain, and calculate the Laplace accurate and establish rules based on the variation of the gain similarity ratio. The benefits of this research is that we can predict gradution rates students so that early anticipation can be done as aprogram of counseling, special additional duty, or other similar programs.**

Keywords : data mining, CPAR, algorithm, warehouse

I. INTRODUCTION

Topics this paper discusses the prediction of graduation of students at a college in Jakarta. This topic is important to be discussed in an effort to capture the condition of students and help the program of study or other part that is authorized to anticipate the existence of failing students by creating specialized programs in order to reach the vision mission of a college. In each semester, at a college doing exams of the semester, whether it be on the midterm (UTS) and the semester final exams (UAS) as one measure of student success. In general, a week after the semester exams take place the quality of each of the courses will come out. Value is the value achieved quality of students that consists of several components and has a weight that has been set. And at each end of the semester students will obtain achievement index that shows the achievement of academic achievement. If a student scores are eligible, meaning he will graduate and earn the grade point average (GPA) with certain criteria, whereas if a student get a value that does not mean he's qualified to fail or drop out. Drop-out can be prevented if the college counseling to students as early as possible. In order for the counseling process can be done as early as possible a college must have the ability to predict the value to be achieved by the students. This paper discusses the prediction of graduation of students using the data warehouse and data mining. CPAR algorithm used for their solution. To predict the status of GPA use the relationship between the value of each subject obtained and GPA

II. LITERATUR REVIEW

A. Data Mining

Data mining is a process of extracting knowledge from a number of large amounts of data (Han and Kamber, 2001). Data mining can be used to determine whether this information can be used to analyze the data using a particular method. Data mining is a study that develops due to the development of automatic data collection tools and database technologies. Data mining is an integral part of knowledge discovery in databases, a process of converting raw data into useful information (Tan, Steinbach and Kumar, 2006). In general, functions in data mining can be grouped into two, namely prediction and description. Prediction is a data mining using multiple variables to predict unknown values or value in the future of other variables. While the description is a data mining to find patterns that could be interpreted human description existing data. Included in the forecast include classification and regression. While that includes a description of the clustering and association analysis. Association analysis already popularized by Rakesh Agrawal, a researcher at IBM's Almaden Research Centre since 1993 (Prasetyo, 2006). Association analysis is the acquisition of association rules that show attribute value conditions that often occur together in a set of data. Association analysis is useful to obtain interesting relationships hidden in large data set. The basic problem in association analysis is formulated as how to find the rules in the form himpunan1 and himpunan2 (Possas et al, 2000). A popular application of association analysis is market basket analysis that studies shopping

behavior of customers by finding a set of items most frequently purchased together. This application helps retailers plan for the laying of the items of goods sold. In market basket analysis, each item is seen as a boolean variable to represent the presence or absence of these items in the basket. Each shopping cart is represented by a boolean vector indicating the variables. Boolean vectors can be analyzed to see spending patterns that indicate the items most associated or purchased together. This pattern is expressed in the form of association rules. As an example of information that customers who buy computers also tend to purchase financial management software in the form of rules expressed as follows: Computers – Software financial management (support = 2%, confidence = 60%). Support and confidence are two measures attractiveness of rules, each of which demonstrate the utility and the certainty of the acquisition rules. From the example above, the support of 2% means that 2% of all transactions that were analyzed showed that the computer and financial management software purchased together, while the confidence of 60% means 60% of customers who buy computers also purchase financial management software. An association rule is considered strong if it satisfies the second threshold value called min_support and min_confidence.

B. Classification Based Association.

Currently, one of data mining techniques have been developed is to apply the concept of association rule mining in classification problems. There are several methods that can be used, such as Association Rule Clustering System (ARCS) and association clasification (Han & Kamber, 2001). ARCS method perform association rule mining based on clustering kemudiann using the resulting classification rule. ARCS, perform association rule mining is in the form $A_{quant1} \wedge A_{quant2} \rightarrow A_{cat}$, where form and A_{quant1} A_{quant2} is test data whose attributes have a range of values, show the class label for the attribute category. A_{catw} show the class label for the attribute category. Association methods produce *classification rules* in the form *candset -> y,* where candset is a collection of items and y is the class label. Rules in accordance with certain minimum support is called frequent. Rule has support s if s% of data in a data set containing candset and have the class y. Rules in accordance with the minimum confidence is accurate. Rule has confidence c if C% of the sample in the data set containing candset have the class y. If some rule has the same candset the rule with highest confidence is selected as the possible rule. Methods association classification association rule mining algorithm, such as appriori algorithm to generate association rule, then select a group of rules that have high quality and use these rules to improve the data. Association classification is still less efficient because they often produce large amounts of rules (Yin & Han, 2003). Other association-based classification method is CPAR (Classification based on Predictive Association Rule) in generating rules and integrate it with associative association.

C. Classification based on Predictive Association Rule (CPAR)

Effective algorithm to be used in the classification problem is CPAR (Yin & Han, 2003). In this classification algorithm is implemented in three stages, rule generation, rule evaluation, and classification. In the process of rule generation, CPAR build rules by adding literals one by one. At each stage of the process, CPAR calculate Gain of each literal. After each data is processed to get the rule, this data is used again in the calculation of gain but with reduce the weight appropriate decay factor. Weight reduced until it reaches the minimum value calculated by the parameter w is the weight of all the positive data. The weight of the entire data set for the beginning of the process is equal to 1. After the process of rule generation, CPAR evaluate each rule to determine the strength of its predictions. To rule $r = p1 \wedge p2 ,...\wedge pn \rightarrow c$, CPAR define expectations of accuracy as follows: LA = (nc +1) / (ntot + f), where LA is the Laplace accuracy. f is the number of classes, ntot is the total amount of data that meet the rules, nc is the data that meets the class c. Classification of a set of rules for each kelas.CPAR use s the best rule class, chosen by the LA. The basic idea comes from FOIL CPAR which uses greedy algorithm to learn the rules that distinguish positive examples with examples neggatif. FOIL repeatedly looking for the best rules and positive move all covered by the rule until all positive examples in the data set covered. FOIL algorithm first read the data input of the set of positive and negative examples, produces the output in the form of rules that are useful for predicting the class label of positive examples or negative examples. At the beginning of the process r defined set of rules is empty. Formation process repeated for the number of positive samples is greater than 0 and during the process of formation of FOIL algorithm copying examples to positive examples as positive and negative examples to negative examples while, read the attributes of positive examples one by one and add to the rule list if the Gain appropriate with yangg determined. When choosing a literal, FOIL Gain is used to measure the information obtained from the addition of literals to the current rule. Suppose there are | P | examples of positive and | N | negative examples meet the body of rule r. After a literal p is added to r, there is a |P *| examples of positive and |N *| negative examples that meet the body of r the new rules. Gain FOIL p is defined as follows (Yin & Han, 2003). Gain (p) = 2 ((log |P *| / |P *| + |N *|) - (log |P *| / | P | + | N |)), where | P| and | N | is the number of positive examples and the number of negative examples that satisfy the rule r. | P * | and | N * | is the number of positive examples and the number of negative examples that satisfy the rule r new. CPAR algorithm is a development of the RPM algorithm, an algorithm that modifies FOIL to obtain accuracy and better

efficiency. CPAR and the RPM difference is other than simply choosing an attribute that has the best gain in each iteration, CPAR can select a number of attributes that gain almost the same value. Selection of attributes is done by calculating and applying the gain similarity ratio. All attributes with a value greater than the gain similarity ratio will be selected and processed further. In the method CPAR process begins by reading the data in the form of a set number of two-dimensional Array each column given attribute A and the last attribute shows the class. The next input data are grouped into positive examples P and N negative examples according to the class. The weight of positive examples | P | and the weight of negative examples | N | each attribute is added to form a PN array, a two-dimensional array containing a list of all the attributes, weight and the weight of positive examples negative examples. Total Weight Threshold (TWT) is calculated by multiplying the number of positive weights with a constant that the project was set 0.05. The process of rule formation is repeated until the total weight of positive examples is smaller than the TWT. In every process done copying P, N, A and PN to P ', N', A' and PN'. Calculating the gain and insert rules into the rule list. In this project the minimum constant gain is 0.7 and the decay factor is 1 / 3. CPAR create a rule s by adding literals one by one. After finding the best literal, the other literal Gain like to search.

D. Build Prediction Model

The basic process of building predictive models is the same, regardless of data mining techniques are used. Success in building a model more dependent on the process, and the process is highly dependent on the data used in producing a model where the main challenge is to gather sufficient preliminary data.

III. METHODOLOGI

The steps undertaken in this project work are as follows:

Step 1 : Domain Understanding
Want to explore whether there is an association between subjects with the quality of the obtained values. The value of students drawn from all subjects values obtained by students in semester 1.

Step 2 : Collecting research data
Student data to be studied is limited only to the student information management courses at a private college in the first four years (with the armed forces until 2002 / 2003 1999 / 2000) of approximately 280 people. The data captured includes student data, course data, and data values. From the combined data attribute data obtained in the form of denim, name, gender, class year, the value of each course grade point (IP), and grade point average (GPA).

Step 3 : Data preprocessing
In the process of data preprocessing performed includes several steps: data cleaning, data selection, data transformation.

1) Data cleansing, carried out to identify, modify, clean up data that is inconsistent or inaccurate. After the identification process the data collected to check the completeness of the data. The process of change is done by giving the file naming standard for ease of management data. The process of data cleansing do remember to clean up data that is inaccurate. There are some students who have a blank value before entering semester 6, because due to drop out. So for an empty data subjects are given the value of E. GPA used is considered the same as the last score obtained.
2) Data integration, data from various sources combined. After cleaning the data, performed the integration of data from all the students so that merged into one fact table, a summary GPA student who has the attributes of nim, name, class year, course, the value of each course, the value of GPA.
3) The selection of data, selected relevant data for analysis. Attribute data collected from the selection of attributes to be analyzed and decided the attribute to be analyzed is nim, the courses, the courses, and GPA. For GPA status under the provisions of the institution.
4) Transformation of data, at this stage the data are transformed or consolidated into a form suitable for mining. At this stage we do some calculations such as the number of students, and some other operations.

Step 4 : Selection of mining
Because in this study wanted to see the association between subjects, grades, to predict their academic achievement and graduation of students then selected data mining is association analysis.

Step 5 : Selection of algorithm
To predict the graduation of students, used the CPAR algorithm used, since the algorithm is effectively used for classification problems (Yin & Han, 2003).

Step 6 : Data mining
The data is processed to obtain the association rules and processed by the algorithm CPAR to see the accuracy of predictions.

Step 7 : Pattern evaluation
Identifying the patterns really interesting that describe knowledge based on specific measurements. At this stage after performing the data processing, we evaluate the attributes of nim connectedness, the courses, the value of each course and GPA to predict graduation.

Step 8 : Knowledge presentation

Present knowledge, where the techniques of presentation and representation of knowledge used for the presentation of the knowledge generated from mining to the user. Once we identify status_GPA relationship with sex, the value of subjects with status_GPA we present the knowledge in the form of conclusions. The conclusion that this presented a new and useful knowledge for the user to make decisions.

IV. DISCUSSION AND RESULT

A. Predicting the status students of GPA

This section discusses GPA predict with some initial values and assumptions used in data analysis as follows:

- The code describes the course subjects. To facilitate the search, the project name of the course is used. The code name of the course explains the following subjects: KU121 = english, SI111 = basic math, SI141 = programming, SI131 = orkom, SI121 = introduction of informatics, SI101 = industrial management, KU101 = Pancasila, KU102 = religion..
- decay factor = 0.3
- gain minimum threshold = 0.7
- gain similarity ratio = 0.6
- TWT from positive data = 12.55
- TWT from negative data = 1.2

class based on the institution pursuant GPA so there are 5 classes, namely:

1) if (GPA < 2.00)
2) if (2.00 < GPA < 2.50)
3) if (2.50 < GPA< 3.00)
4) if (3.00 < GPA_ 3.50)
5) if (GPA > 3.50)

- categories based on the value of subjects and obtained 40 attributes. Table 1 describes the categories used to predict the value of student graduation.

- Processing of data for prediction GPA status subjects taken from the value obtained by students in semester 1. Data sets used include ID_tran, sex, values in each subject, and GPA. Figure 1 describes Value data structure .

TABLE I CATEGORIES

No	Categories	Description
1	1	English = A
2	2	English = B
3	3	English = C
4	4	English = D
5	5	English = E
6	6	Basic Math = A
7	7	Basic Math = B
8	8	Basic Math = C
9	9	Basic Math = D
10	10	Basic Math = E
11	11	Programming = A
12	12	Programming = B
13	13	Programming = C
14	14	Programming = D
15	15	Programming = E
16	16	Orkom = A
17	17	Orkom = B
18	18	Orkom = C
19	19	Orkom = D
20	20	Orkom = E
21	21	Introduction to Informatics = A
22	22	Introduction to Informatics = B
23	23	Introduction to Informatics = C
23	23	Introduction to Informatics = D
25	25	Introduction to Informatics = E
26	26	Industral management = A
27	27	Industral management = B
28	28	Industral management = C
29	29	Industral management = D
30	30	Industral management = E
31	31	Pancasila = A
32	32	Pancasila = B
33	33	Pancasila = C
34	34	Pancasila = D
35	35	Pancasila = E
36	36	Religion = A
37	37	Religion = B
38	38	Religion = C
39	39	Religion = D
40	40	Religion = E

ID_trans	sex	co TH	KU121	SI111	SI141	SI131	SI121	SI101	KU101	KU102	Status_IPK
1	L	T1	A	A	A	A	B	A	A	A	4
2	L	T1	B	B	B	B	B	C	B	B	3
3	L	T1	C	B	B	B	A	B	B	B	3
4	L	T1	C	B	B	C	B	C	B	B	3
5	L	T1	B	B	B	B	B	B	B	B	4
6	L	T1	C	C	C	C	B	C	C	B	3
7	P	T1	B	B	C	C	B	A	B	A	3
8	L	T1	B	C	C	B	B	C	C	A	3
9	L	T1	C	B	A	A	B	C	D	B	3
10	L	T1	B	B	B	A	A	B	C	B	3
11	P	T1	B	B	C	B	A	A	B	B	3
12	P	T1	B	B	A	B	B	B	B	A	4
13	L	T1	B	B	B	C	C	B	C	B	3
14	L	T1	B	B	A	C	B	A	B	B	3
15	L	T1	C	C	B	B	B	B	A	A	3
16	L	T1	C	C	B	C	B	B	C	B	3

Figure 1 Value data

- From the data value, then formed the positive data and negative data. Negative data is data that pass negative (positive did not pass), the data obtained from the data class 1. Positive data is passed positive data, the data obtained from the data class 2, 3, 4, and 5. The following image is a data created based on the positive class. The last column contains the total weight of each record from positive data. This applies also to the negative data.

ID_transaksi	sex	co TH	KU121	SI111	SI141	SI131	SI121	SI101	KU101	KU102	Kelas	Bobot
1	L	T1	A	A	A	A	B	A	A	A	4	1
2	L	T1	B	B	B	B	B	C	B	B	3	1
3	L	T1	C	B	B	B	A	B	B	B	3	1
4	L	T1	C	B	B	C	B	C	B	B	3	1
5	L	T1	B	B	B	B	B	B	B	B	4	1
6	L	T1	C	C	C	C	B	C	C	B	3	1
7	P	T1	B	B	C	C	B	A	B	A	3	1
8	L	T1	D	C	C	D	D	C	C	A	3	1
9	L	T1	C	D	A	A	D	C	D	D	3	1
10	L	T1	B	B	B	A	A	B	C	B	3	1
11	P	T1	B	B	C	B	A	A	B	B	3	1
12	P	T1	B	B	A	B	B	B	B	A	4	1
13	L	T1	B	B	B	C	C	B	C	B	3	1
14	L	T1	B	B	A	C	B	A	B	B	3	1
15	L	T1	C	C	B	B	B	B	A	A	3	1
16	L	T1	C	C	B	C	B	B	C	B	3	1

Figure 2 Positive data

- At the beginning of each process each record has a weight of 1. From 274 data records, there are 251 records positive data at the initial weight of 1, so the total weight of positive data at the beginning of the process 251. Total Weight Threshold (TWT) are calculated based on the formula TWT = Total weight of the positive sample x 0.05. TWT then obtained 12.5. All categories hereinafter processed one by one until the total weight is less than TWT.
- The next stage is to establish Gain. From the formula Gain, Gain obtained in each category shown in Table 6.
- Next, the formation of the gain, where gain is obtained by the formula: Gain (p) = 2 ((log | P * | / | P * | + | N *|)-(log | P * | / | P | + | N |)). Let us take the example formation Gain in category 2. Then Gain (2) = 2 ((log | 133 | / | 133 | + | 9 |) - (log | 133 | / | 154 | + | 9 |)). Gain values in category 2 will be obtained by 0:12

KATEGORI	BOBOT DT POSITIF	BOBOT DT NEGATIF	GAIN
1	21	0	0.00
2	133	9	0.12
3	95	7	0.83
4	2	3	3.46
5	0	4	0.00
6	50	0	1.62
7	84	2	1.36
8	114	11	1.26
9	2	6	3.66
10	1	4	4.08
11	27	0	2.66
12	88	0	1.75
13	121	4	1.60
14	15	15	2.87
15	0	4	0.00
16	14	0	3.55
17	141	0	1.68
18	94	18	1.98
19	2	1	5.12
20	0	4	0.00
21	13	0	3.86
22	111	0	2.08
23	122	10	2.02
24	4	9	4.04
25	1	4	4.88
26	25	0	3.49
27	116	0	2.23
28	107	11	2.28
29	3	1	4.35
30	0	4	0.00
31	36	0	3.34
32	136	7	2.21
33	76	11	2.68
34	3	1	5.36
35	0	4	0.00
36	49	0	3.21
37	139	7	2.32
38	63	9	2.96
39	0	0	0.00
40	0	7	0.00

Figure 3 Categories and Gain value

From Figure 3 the biggest gains are owned by the category 34 with a value of 5.68. The next step to calculate LGT (Local Gain Threshold). LGT is obtained from the largest gain_similarty_ratio. So LGT = 5.68 - 0.6. LGT obtained was 3:22. From the table there were 10 category formation gain. Gain whose value is above the LGT, the categories 19, 25, 29, 10, 21, 9, 16, 26, 4, and 31. The ten categories are processed one by one either on the weight of positive data and negative data weights.

Iteration 1:

1) In the 19 categories are processed one by one, by inserting rule 19 → 41 to rule the list temporary, positive and copy data negative data by removing the line that contains no attributes on category 19. Category 41 is an empty category, so it is considered a temporary storage to accommodate data that do not meet the category 19. From the overall data, the data that meets the requirements shown in Figure 4.

ID_trans	sex	co TH	KU121	SI111	SI141	SI131	SI121	SI101	KU101	KU102	Status IPK	Bobot	Kelas
43	L	T1	C	D	D	D	C	C	B	E	1	1	negatif
195	P	T3	A	E	D	D	E	C	A	C	2	1	positif
248	L	T4	C	C	C	D	C	C	C	B	2	1	positif

Figure 4 . Categories of 19 for positive value and negative value

2) The next step is to calculate the LA for each of the positive data and negative data. LA is calculated using the formula (nc +1) / (ntot + f). Where to positive data from this iteration is known nc = 2, (ie ID_195 and ID_248), ntot = 2, (ie ID_195 and ID_248). For the negative data from iteration is known to nc = 1, (ie ID_43) and ntot = 1, (ie ID_43). f = 5, ie 1,2,3,4,5 class. So LA to positive data is 0375 and the LA to the negative data is 0.25. Bobot positive data and negative data subsequently revised by using the decay factor, so that the data generated positive and negative data are new. Figure 5 is the result of positive and negative data only shown just a few lines.

ID_trans	sex	co TH	KU121	SI111	SI141	SI131	SI121	SI101	KU101	KU102	Kelas	Bobot
42	L	T1	C	C	D	C	D	C	C	C	1	0.3333
43	L	T1	C	D	D	D	C	C	B	B	1	1
44	L	T1	B	C	C	C	D	C	D	B	1	0.3333
124	L	T2	C	C	D	C	D	D	B	B	1	0.3333
126	L	T2	C	C	D	C	C	C	C	C	1	0.3333
126	L	T2	B	B	B	B	B	B	B	B	1	0.3333
127	L	T2	B	B	B	B	B	B	B	B	1	0.3333
212	L	T3	B	B	B	B	B	B	B	B	1	0.3333
213	L	T3	B	B	C	C	C	C	B	B	1	0.3333
214	L	T3	B	C	C	C	C	C	B	C	1	0.3333
216	L	T3	B	C	C	C	C	C	B	C	1	0.3333
216	L	T3	B	B	B	B	B	B	B	B	1	0.3333
264	L	T4	B	D	D	C	C	C	C	C	1	0.3333
266	P	T4	C	C	D	C	D	D	C	B	1	0.3333
267	L	T4	B	C	D	C	C	D	C	C	1	0.3333
269	L	T4	C	B	D	C	C	C	C	C	1	0.3333
270	P	T4	D	D	D	C	D	C	C	D	1	0.3333
271	L	T4	C	D	D	C	C	C	B	B	1	0.3333
272	L	T4	D	C	D	C	D	D	C	C	1	0.3333
274	P	T4	B	D	D	C	C	D	C	B	1	0.3333
32	L	T1	B	C	C	C	B	B	B	B	2	0.3333
33	L	T1	B	C	C	B	C	B	B	B	2	0.3333
34	L	T1	B	B	B	C	B	B	B	B	2	0.3333

Figure 5. Categories of 19 and formed new wight

3) Calculating the value of weight on each positive data and negative data where the value obtained from the total weight of each weight to the positive data and negative data are new. Weight values obtained new positive data was 85 and the new value of the negative weight is 8667.

4) The weight of positive data is now 85, still higher than 12.5, so the next category of 25 which has a gain value over the LGT is processed by inserting the 25 → to the list temporary rule.

Iteration 2:

1) Insert rule 25 → 41 to rule list temporary, copy positive data and negative data by removing the line that contains no attributes on category 25. From the overall data, the data that meets the requirements shown in Figure 6.

ID_trans	sex	co TH	KU121	SI111	SI141	SI131	SI121	SI101	KU101	KU102	Kelas	Bobot
126	L	T2	e	e	e	e	e	e	e	e	1	1
127	L	T2	e	e	e	e	e	e	e	e	1	1
212	L	T3	e	e	e	e	e	e	e	e	1	1
216	L	T3	e	e	e	e	e	e	e	e	1	1
195	P	T3	A	E	D	D	E	C	A	C	2	1

Figure 6 Categories of 25 for Positive value and negative value

2) The next step is to calculate the LA for each of the positive data and negative data. LA is calculated using the formula (nc +1) / (ntot + f). Where to positive data from this iteration is known nc = 1, (ie ID_195), ntot = 4, (ie ID_195). For the negative data from iteration is known to nc = 4, (ie ID_126, ID_127, ID_212, and ID_216) and ntot = 4, (ie ID_126, ID_127, ID_212, and ID_216). f = 5, ie 1,2,3,4,5 class. So LA is 0.5 for positive data and negative data is LA for 0.2. The weight of the positive data and negative data was subsequently revised by using the decay factor, so that the data generated positive and negative data are new. Figure 7 is the result of positive and negative data only shown just a few lines.

ID_trans	sex	co TH	KU121	SI111	SI141	SI131	SI121	SI101	KU101	KU102	Kelas	Bobot
42	L	T1	C	C	D	C	D	C	C	C	1	0.1111
43	L	T1	C	D	D	D	C	C	B	e	1	0.1111
44	L	T1	B	C	C	C	D	C	D	e	1	0.1111
124	L	T2	C	C	D	C	D	D	B	B	1	0.1111
126	L	T2	C	C	D	C	C	C	C	C	1	0.1111
126	L	T2	e	e	e	e	e	e	e	e	1	1
127	L	T2	e	e	e	e	e	e	e	e	1	1
212	L	T3	e	e	e	e	e	e	e	e	1	1
213	L	T3	B	B	C	C	C	C	B	e	1	0.1111
214	L	T3	B	C	C	C	C	C	B	C	1	0.1111
216	L	T3	B	C	C	C	C	C	B	C	1	0.1111
216	L	T3	e	e	e	e	e	e	e	e	1	1
264	L	T4	B	D	D	C	C	C	C	C	1	0.1111
266	L	T4	B	C	D	C	D	D	B	C	1	0.1111
266	P	T4	C	C	D	C	D	D	C	B	1	0.1111
267	L	T4	B	C	D	C	C	D	C	C	1	0.1111
269	L	T4	C	B	D	C	C	C	C	C	1	0.1111
269	P	T4	D	D	D	C	D	C	C	B	1	0.1111
270	P	T4	D	C	D	C	D	D	C	B	1	0.1111
271	L	T4	C	D	D	C	C	C	B	B	1	0.1111
272	L	T4	D	C	D	C	D	D	C	C	1	0.1111
273	L	T4	B	D	D	C	D	D	C	B	1	0.1111
274	P	T4	B	D	D	C	C	D	C	B	1	0.1111
32	L	T1	B	C	C	C	B	B	B	B	2	0.1111
33	L	T1	B	C	C	B	C	B	B	B	2	0.1111

Figure 7. Categories of 25 and formed new weight

3) Calculating the value of weight on each positive data and negative data where the value obtained from the total weight of each weight to the positive data and negative data are new. Weight values obtained positive data is new is 28.77 and the new value of the negative weight is 6:22.

4) Value weighting positive data is now 28.77, still higher than 12.5, so the next category of 29 that has a value above LGT Gain processed by inserting the 25 → to the list

temporary rule. Iteration is done in the ten categories. Iteration is stopped when the weight value of positive data is smaller than 12.5 or all categories Gain greater value has been processed. In this project the iteration is stopped on the third iteration for the positive data with weighted values positive 12.18 data and on the tenth iteration for negative data with a negative weight value data 3:00.

5) Once the data is processed, the next step to calculate the weight gain that has been adjusted. Gain calculation results shown in Figure 8.

KATEGORI	BOBOT DATA POSITIF	BOBOT DATA NEGATIF	GAIN
1	1	0	0.00
2	6	1	0.11
3	5	1	0.82
4	0	0	2.95
5	0	1	0.00
6	2.427	0	1.72
7	4.078	0.261	1.41
8	5.534	1.439	1.24
9	0.097	0.785	3.06
10	0.049	0.523	3.45
11	1.311	0	2.77
12	4.272	0	1.85
13	5.874	0.523	1.64
14	0.728	1.962	2.45
15	0	0.523	0.00
16	0.68	0	3.67
17	6.845	0	1.78
18	4.563	2.355	1.88
19	0.097	0.130	4.84
20	0	0.523	0.00
21	0.68	0	3.91
22	6.845	0.392	1.95
23	4.563	1.439	2.19
24	0.097	1.177	3.55
25	0	0.523	0.00
26	1.214	0	3.61
27	5.631	0.915	2.34
28	5.194	1.57	2.26
29	0.148	0	3.77
30	0	0.523	0.00
31	1.748	0	3.46
32	6.602	0.915	2.26
33	73.838	1.57	2.59
34	0	0	0.00
35	0	0.523	0.00
36	2.379	0	3.32
37	6.748	0.915	2.37
38	3.058	1.177	2.91
39	0	0	0.00
40	0	0.915	0.00

Figure 8 New gain that meets TWT

6) The ten categories in iteration 1 has a value greater than the TWT gain is not included in the next process is shown in Figure 8, so the value of the gain on the ten categories are ignored. Maximum gain is now 2.91 which is smaller than the global mean minimum 0.7, so the process is stopped

7) The next stage is to calculate the accuracy with using LA. on this project from all iteration we got the result shown in Table 2.

TABLE 2 GAIN OF THE POSITIVE CLASS

No	RULE	Aturan	LA
1	19 -> 1	Jika Orkom = D Then Negatif Lulus	0.25
2	19 -> 2	Jika Orkom = D Then Positif Lulus	0.375
3	25 -> 1	Jika Pengantar Informatika = E Then Negatif Lulus	0.2
4	25 -> 2	Jika Pengantar Informatika = E Then Positif Lulus	0.5
5	29 -> 1	Jika Manajemen Industri = E Then Negatif Lulus	0.14
6	29 -> 2	Jika Manajemen Industri = E Then Positif Lulus	0.375
7	29 -> 3	Jika Manajemen Industri = E Then Positif Lulus dengan status_IPK 3	0.25
8	10 -> 1	Jika Matematika Dasar = E Then Negatif Lulus	0.5
9	24 -> 1	Jika Pengantar Informatika = D Then Negatif Lulus	0.7
10	29 -> 1	Jika Manajemen Industri = D Then Negatif Lulus	0.67
11	4 -> 1	Jika Bahasa Inggris = D Then Negatif Lulus	0.5

8) From Table 2, shows that the rules if Introduction to Informatics = D Then Passed Negative has LA's most high on negative class pass is 70%. This means that students with the course Introduction to Informatics = D have a negative chance of graduating (not pass) of 70%. Rules if Orkom = D then Positive has the highest graduation LA on 37.5% positive class. This means that students with the course Orkom D have the opportunity pass by 37.5%.

9) Based on Table 2 shown also that the value of currency introductory lecture informatics, industrial management, language English, basic mathematics and orkom be a determinant principal to determine whether the student is not positive pass or not pass negative. By using the gain similarity ratio to produce rules that are combination of the above subjects determinants. Rules can include 2 combinations, 3 combined, 4 and 5 combined combination.

V. CONCLUSION

After doing the research can be concluded that:

1. By using a combination, be obtained if a student who scored Introduction to Informatics = D then has a chance of 70% Negative Pass.
2. By using a combination obtained if one student gets the value Orkom = D then mrmiliki opportunities Passed A total of 37.5% Positive 41.97% of all data, if the female has a 2:00 _nilai GPA GPA <2.50, with a degree of confidence of 45 , 57%
3. To improve the quality of the presentation of knowledge can be used up to 5 combinations.
4. In the discussion of the predictive value of students, it can be concluded that determining the predictive value of students does not depend on the formation of the gain

and the techniques used. Predictive value depends on the amount of student data and data processing.

REFERENCES

[1] Srikant, Ramakhrishnan, Agrawal, Rakesh, (1996). Mining Quantitative Association Rules in Large Relatio Tables, Proceedings of the 1996, ACM SIGMOD, International Conference on Management of Data

[2] Possas, Bruno, Meira Wagner Jr., Carvalho Marcio and Resende Rodolfo, 2000, Using Quantitative Information for Efficient association Rule Generation, ACM SIGMOD, Vol 29 , 19-25

[3] Tsai, Pauray S.M. and Chen Chien Ming, (2001) Mining Quantitative Association Rules in a Large Database of Transactrions, Journal of Information Science and Engineering 17, 667-681.

[4] Yin & Han (2003) journal CPAR Classification based on Predictive Association Rule.

[5] Han, Jiawei, Kamber, Micheline, (2001). Data Mining Concepts and Technique, (Academic Press, San Diego)

[6] Tan, Pang-Ning, Steinbach, Michael dan Kumar, Vipin, (2006) Introduction to Data mining, Pearson Education, Inc.

[7] Mehmed Kantardzic, Data Mining - concept, models, methods, algorithms, (2001) Wiley Interscience.

[8] Mohd Shariff. (1995). Steam Regeneration of A Fixed Bed Adsorption System, (Ph.D. Thesis, Leeds University, United Kingdom), 88-90.