The Asian Journal of Technology Management Vol. 8 No. 1 (2015): 68-74

Predicting Success Study Using Students GPA Category

Awan Setiawan* and Kuntjahjo S. L. Margono
Faculty of Engineering, Langlangbuana University, Bandung, Indonesia

Abstract. *Maintaining student graduation rates are the main tasks of a University. High rates of student graduation and the quality of graduates is a success indicator of a university, which will have an impact on public confidence as stakeholders of higher education and the National Accreditation Board as a regulator (government). Making predictions of student graduation and determine the factors that binders will be a valuable input for University. Data mining system facilitates the University to create the segmentation of students' performance and prediction of their graduation. Segmentation of student by their performance can be classified in a quadrant chart is divided into 4 segments based on grade point average and the growth rate of students performance index per semester. Standard methodology in data mining i.e CRISP-DM (Cross Industry Standard Procedure for Data Mining) will be implemented in this research. Making predictions, graduation can be done through the modeling process by utilizing the college database. Some algorithms such as C5, C & R Tree, CHAID, and Logistic Regression tested in order to find the best model. This research utilizes student performance data for several classes. Parameters used in addition to GPA also included the master's students data are expected to build the student profile data. The outcome of the study is the student category based on their study performance and prediction of graduation. Based on this prediction, the university may recommend actions to be taken to improve the student achievement index and graduation rates.*

Keywords: *graduation, segmentation, quadrant GPA, data mining, modeling algorithms*

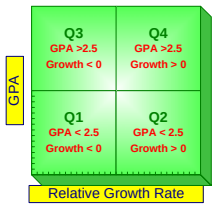
1. Introduction

1.1. Backgrounds

According to Indonesian Council Regulation No. 12/2012 that concern on higher education, the regulations said that one of the University goals is to produce graduates who became master or specialist in branch of science and / or technology to meet national interests and increase the nation's competitiveness. Maintaining student graduation rates are the main tasks of University. High rates of student graduation and the quality of graduates is an indication of success of a university, which will have an impact on public confidence as stakeholders of higher education and the National Accreditation Board as a regulator (government). Accreditation standards require Universities undergraduate study programs should have a focus and a commitment to the quality of the implementation of the academic process (education, research, and service/ community service) in order to provide the competencies required to be a graduate student who is able to compete. This standard also covers how to treat undergraduate study program and provide excellent service to students and graduates to obtain high-quality results. To maintain the quality of education and, as a result, it is necessary to pass the evaluation process. Making predictions of student graduation and determine the factors that hinders will be a valuable input for the University.

Data mining facilitates the University to create the segmentation of students' performance and prediction of their graduation. Segmentation of student learning outcomes can be classified in a quadrant chart formed of two main parameters that GPA and level of progress which will produce 4-quadrant segments, namely: to create a performance index as the y-axis and the rate of progress as the x-axis. For that made restrictions, positive y-axis is GPA above 2.5 and the rate of progress made is calculated in percent cumulatively per semester. The results are as follows:

*Corresponding author. Email: awans2425@gmail.com
Received: May 9, 2015 ; Revised: June 12, 2015, Accepted: July 12, 2015
DOI: http://dx.doi.org/10.12695/ajtm.2015.8.1.7
Print ISSN: 1978-6956; Online ISSN: 2089-791X.



Q1 → GPA < 2.5 and rate of progress <= 0
Q2 → GPA < 2.5 and rate of progress > 0
Q3 → GPA >= 2.5 and rate of progress <= 0
Q4 → GPA >= 2.5 and rate of progress > 0

GPA segmentation will be used in a predictive model of graduation. Parameters used in addition to GPA also included the master's students are expected to form the student profile data.

Making predictions, graduation can be done through the modeling process by utilizing the college database. The more parameters should be used to form a comprehensive model anyway. Some algorithms such as C5, C & R Tree, CHAID, and Logistic Regression tested to look for the best model.

The outcome is student category based on their study performance and prediction of graduation. Based on this prediction, the college may recommend actions to be taken to improve the student achievement index and graduation rates.

1.2. Contribution and Research Output

The results of this research will generate quadrant student performance that will allow University to act help to the success of student learning, especially for quadrant 1, 2 and 3. For sharper handling required additional data that will result in the group with specific characteristics that require different handling tips.

1.3. Purpose of Research

The purpose of this study was to analyze the performance of the progress of student learning outcomes and prevent the failure of student learning to improve graduation index University. As the research data is limited to certain departments of several forces.

1.4. Research Methodology

Standard methodology in data mining i.e. CRISP-DM (*Cross Industry Standard Procedure for Data Mining*) will be implemented in this research.

- *Business Understanding,* the target is to understand the real problem which is faced by University including *business context* that influence the problem.
- *Data Understanding,* the target is to determine the relevant parameters, variables and attributes related to the problem.
- *Data Preparation* is an activity to manipulate data so that it's ultimately valid. This valid data will then be inputs to the modeling process.
- *Modeling,* is a process of determining the best model that will be used to solve the problems.
- *Evaluation* is a group of activities intended to evaluate in detail the accuracy of the output of the chosen model inclusive of changing business context that may alter the chosen model. The model has to be able to solve the defined problem.
- *Deployment* the target is to use the output of the chosen model as a business tool to solve the problems.

2. Research Outcomes

2.1. Business Understanding

Once a prospective student accepted at universities and begin the process of learning, the student data will be entered into the database University. And then based on the progress of learning outcomes shown by GPA per semester, the student performance can be classified into four quadrants. By grouping based on this quadrant, the university can more easily monitor the progress of learning outcomes and immediately act to anticipate setbacks student achievement index by holding a mentoring program so that the expected progress is achieved. Target to be achieved is to understand the problems facing the following business context that influence.

2.2. Data Understanding

The target is to determine the parameters, variables and attributes that affect the problem at hand.

Collect Initial Data

Some datasets are collected from University database comprising of:

- Students Master Dataset
- Students Supporting Data
- Students Performance Dataset (per semester)

All datasets are compiled for per semester for five years to get GPA.

Describe Data

Every column in all datasets are described so that everybody can have correct understanding of the data.

Explore Data

Compiling datasets per period that will be needed in the modelling process and/ or forecasting process as below.

All datasets above and the same datasets one year before are compiled.

Verify Data Quality.

Verifying data quality is done in order to avoid an error that may cause while extracting and processing data from the server .

2.3. Data Preparation

In this process, we determined which datasets that would be used as well as taking them out from the servers' databases. Then, we choose and separated columns and data content that would be used further for modeling and forecasting. Data which was null and undefined was cleaned so that it could be used for modeling process. The process continued with constructing data according to the required data structure so that it can be used for the modeling as well as integrating some files to make a new-more-complete file that will used for further processes. Included in this process was reformatting data.

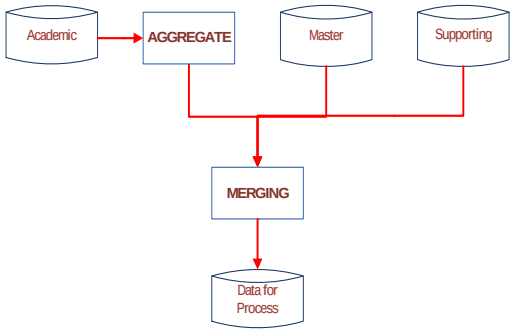


Figure 1. DFD Prediction

Figure 1 describe the flow of making main data that will be used in segmentation process and prediction process. We use aggregate of student academic performance dataset, that merge with student master dataset and student support dataset

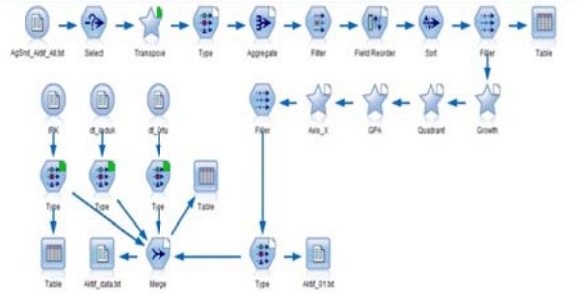
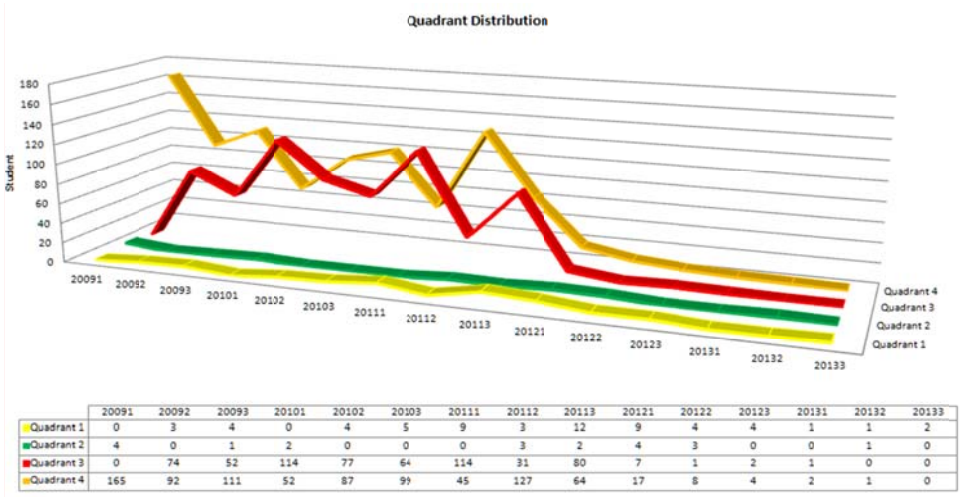


Figure 2. Data Preparation Using IBM SPSS Modeler

Figure 2 is a group of nodes, known as the stream, created using IBM SPSS Modeler that perform data preparation process. This stream represent nodes for data input, process and output with the several algorithm.

2.4. Data Analysis

After merging process of three databases above, we finally have a single completed database which is ready to be used for analysis and prediction. This main data process can be analyzed e and presented in graphic mode as follow



	20091	20092	20093	20101	20102	20103	20111	20112	20113	20121	20122	20123	20131	20132	20133
Quadrant 1	0	3	4	0	4	5	9	3	12	9	4	4	1	1	2
Quadrant 2	4	0	1	2	0	0	0	3	2	4	3	0	0	1	0
Quadrant 3	0	74	52	114	77	64	114	31	80	7	1	2	1	0	0
Quadrant 4	165	92	111	52	87	99	45	127	64	17	8	4	2	1	0

Figure 3. Student Quadrant Distribution

Having known the profile of student performance category per quadrant, we need to know the historical movement from Q4 to other quadrants, which can be described below.

Table 1. Student Quadrant Movement From Q4 To Other Quadrants

	SEMESTER													
	20092	20093	20101	20102	20103	20111	20112	20113	20121	20122	20123	20131	20132	20133
Q4 to Q4	90	53	12	22	37	17	30	45	6	1	0	0	1	0
Q4 to Q3	74	38	99	26	48	78	11	74	4	1	2	1	0	0
Q4 to Q2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q4 to Q1	1	1	0	4	2	4	1	5	2	0	1	1	0	1

2.5. Modelling & Its Output

Having prepared data needed for modeling, we then must choose the best model from some available models in IBM SPSS modeler by using auto-classifier node that can be powerful technique that estimates and compares a number of different modeling methods, ranking them in order of overall accuracy. The chosen model is C.5 which is then be used for predicting successful completion of the curriculum in a timely manner.

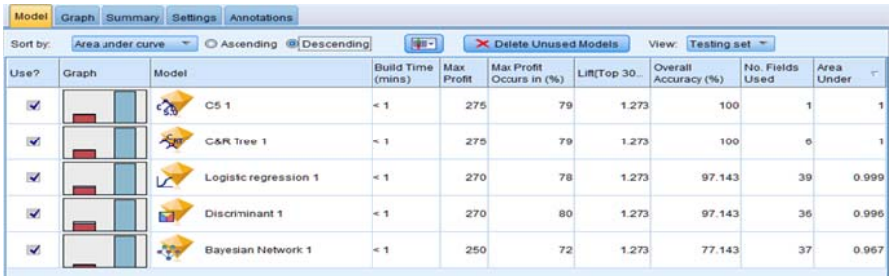


Figure 4. Choosing Method for Modeling Analysis

Data modeling using the data of one class period. The focus of the analysis is the accuracy of students completing the curriculum. Modeling process generates the main parameters that influence, as follows:

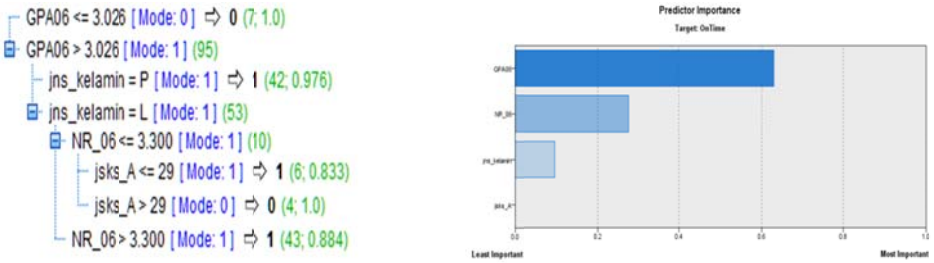


Figure 5. Predictor Importance

The predictor importance for modeling process is GPA 6th semester, student gender, average score 6th semester and *'A score'* amounts as shown in figure 5.

Modeling process has the following algorithm:

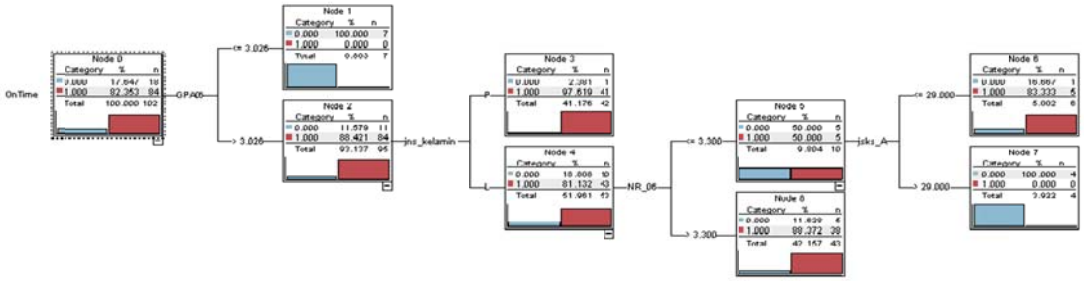


Figure 6. Algorithm of Prediction Model

OnTime

$C-CnTime		0	1	Total
0	Count	19	7	26
	Row %	73.077	26.923	100
	Column %	51.351	5.303	15.385
1	Count	18	125	143
	Row %	12.587	87.413	100
	Column %	48.649	94.697	84.615
Total	Count	37	132	169
	Row %	21.893	78.107	100
	Column %	100	100	100

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 47.074, df = 1, probability = 0

Figure 7. Accuration of Prediction Model

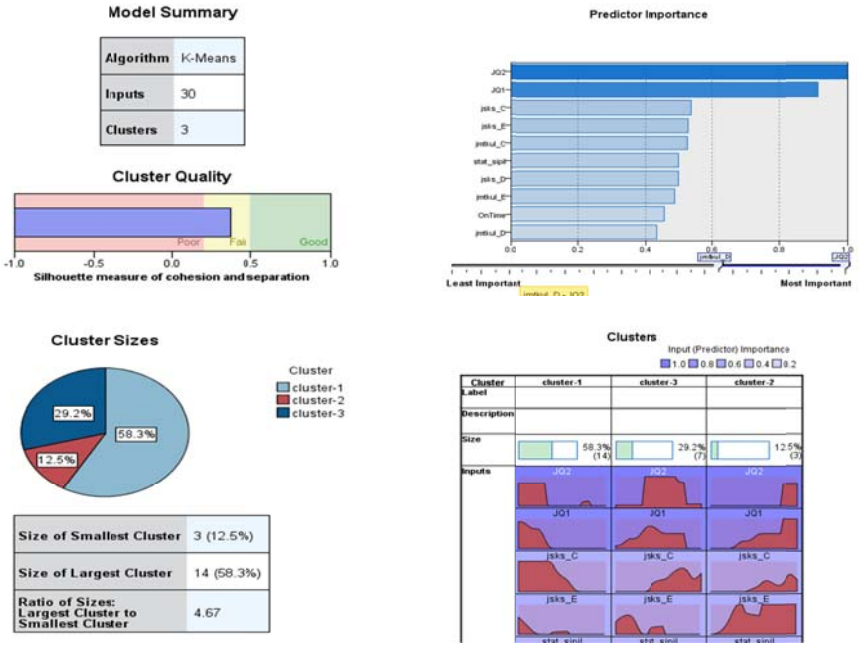


Figure 8. Model Clustering

2.6. Evaluation

The prediction model is applied to the data of student for next period. In order to evaluate the accuracy of prediction, we compare between the prediction with the actual data which is shown below.

Table 2. Accuracy Data, Predicted versus Actual

Matrix | Appearance | Annotations

OnTime

$C-OnTime		0	1	Total
0	Count	20	23	43
	Row %	46.512	53.488	100
	Column %	58.824	18.254	26.875
1	Count	14	103	117
	Row %	11.966	88.034	100
	Column %	41.176	81.746	73.125
Total	Count	34	126	160
	Row %	21.250	78.750	100
	Column %	100	100	100

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 22.424, df = 1, probability = 0

The above table shows that the accuracy of predicted vs actual success of the students that completed the study program curriculum in a timely manner is 88.034 %. We are quite confident with the results.

2.7. Deployment

Considering the accuracy of a predicted, we have to deploy the model as a business tool to solve the problem i.e. understanding student performance behavior which is reflected through its quadrant. With that knowledge, University will be able to exert its best effort to maintain the students to always be in Q4. In order to avoid a shift from Q4 to other quadrants, a special treatment must be implemented.

3. Research Constraints

Complete supporting data will be greatly helpful to track the characteristics of the learning process of students inhibitor category. Unfortunately, such data cannot be obtained easily because it is in many cases very confidential. However with the available data, this study should be able to make the positive contribution.

4. Conclusion and Recommendation

The accuracy of the predictions against the actual data completion of the curriculum is from 87 % to 88.034 %. This proves that the resulting model can be used for the data set used. Model results from IBM SPSS Modeler streams will produce different accuracy when directly applied to the different data groups, for example to other universities, because the resulting parameter values will be different. Nevertheless forming method applied models have been tested, so that the same method, and different data sets should be re-modeling, to form the corresponding model. The prediction results and recommendations can be used by the counselor in a motivating student to intensify efforts to achieve learning success. For prediction accuracy, external information like activities, academic activities of students and family background data, should be included as a part of student demographic data.

References

Bramer, M. (2007). *Principles of Data Mining*. Springer, London.

Giudici, P. (2003). *Applied Data Mining Statistical Methods for Business and Industry*, Wiley.

Han, J. and Kamber, M. (2006). *Data Mining Concepts and Techniques Second Edition*. Morgan Kauffman, San Francisco.

Pyle, D. (2003). *Business Modeling and Data Mining*. Morgan Kaufmann Publishers.

Rud, O. P. (2000). *Data Mining Cook Book,* John Wiley and Sons Nov.

Susanto, S., and Suryadi, D. (2010). *Pengantar Data Mining, Menggali Pengetahuan dari Bongkahan Data*, Andi Yogyakarta

Santosa, B. (2007). *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Graha Ilmu.