DOI : 10.18843/rwjasc/v7i3/03
DOI URL : http://dx.doi.org/10.18843/rwjasc/v7i3/03

HOW EDUCATIONAL DATA MINING CAN PREDICT STUDENTS' ACADEMIC ACHIEVEMENT

Halim Budi Santoso,

Department of Information System
Duta Wacana Christian University,
Indonesia.

Jong Jek Siang,

Departement of Information System
Duta Wacana Christian University,
Indonesia.

ABSTRACT

One of the important measurement of successful academic achievement is GPA and study period. Department of Information System, Faculty of Information Technology Duta Wacana Christian University has a problem with students GPA and study period. Average of study period in IS Department is 5.37 years. Some students study longer than 5 years. Instead of the study period, some students also fail to continue their study. 20.3% active students are failed within first to fourth year.

Data Mining grows rapidly and is used in all sectors, not only in profit organizations but also nonprofit organizations, such as universities. Data mining algorithm helps the development of machine learning algorithm. In this research, researchers try to use data mining algorithm to predict the successful study. Some data are analyzed to discover the relationship and correlation between first years' GPA to the final GPA.

As a result for this study, researchers discovered that there is correlation between first year GPA (first and second semester in students study period) to the final GPA. It is suggested that students should be taken care during the first year. Faculty member should maintain the students' motivation during the first year of students study period.

Keywords: Educational Data Mining; Data Mining; GPA Prediction; Correlation.

INTRODUCTION:

Duta Wacana Christian University is located in Yogyakarta, Indonesia. This is one of the Christian Higher Education Institutions in Indonesia. Currently, Duta Wacana Christian University inspects problems with student's learning motivation, especially in Information System Departments. Information Systems students have average GPA during 2012 to 2015 is 3.1 (out of 4) and the study period 5.37 years (while the normal study period is 4 years according to the curriculum). It means that motivation among Information System Students is average. Not only GPA, study period is also a problem faced by Information System Department Head. Average study period is 5.37. This number is higher than the targeted study period which is 4 years according to the curriculum of Information System Department.

Bahji, Lefdaoui, & Alami (2013) inferred that there is a decrease found in students' motivation to participate learning activity in the classroom. There are some factors that influence these findings which are as follows

- The development of Information Technology
- The usage of Information Technology
- The usage of internet as a learning media tool

Table 1: Number of Failed Students during first to fourth Year

Year	Number of Students	Number of Students in 2nd Semester 2015 / 2016	% Fail and Stop
1st Year Student	59	53	10%
2nd Year Student	51	40	22%
3rd Year Student	49	36	27%
4th Year Student	78	60	23%
Total	237	189	20.3%

From the table 1, it is observed that around 20% students are failed during their study. For sure, it is a big problem for Information System Department, School of Information Technology. Based on those problems, this research is conducted to study the correlation between academic achievement (as measured by GPA like Najafabadi, Najafabadi, and Farid-Rohani (2013) research did) to predict the percentage of successful students during their study.

Table 1 also supports what Siang and Santoso (2016) study in its paper. Information System Department Students feel demotivated towards their study. Demotivated students can be seen with the high study period and many students are out during their study. Siang and Santoso (2016) found that students have willingness to increase their GPA up to 9.2 out of 10. This willingness is not supported with the eagerness to do the assignments and prepare for the exams. It means that students show less efforts to achieve their ambition to increase its GPA.

Academic Achievement was a research topic for Najafabadi, Najafabadi, and Farid-Rohani (2013) whose study was about academic achievement, measured by GPA, for Shahid Bahesti University in Iran, found that there are some factors influence the academic achievement, such as Teaching – Evaluation, Learner, Environment, Family, Curriculum, and Teaching Knowledge.

Clearly, academic performance is a critical factor to be taken into account. Underachievement is associated with a high dropout rate (Martinez, Karanik, Giovannini, and Pinto, 2015). Educational Data Mining is one of the solution that can predict and anticipate the higher number of student's underachievement and drop out. Martinez, et. al. (2015) found out an alternative solution to characterize the students into some certain groups. Thus, the profiling becomes a strategy of significant value when an action is taken to improve the performance of students.

Following the introduction, literature review in Educational Data Mining is conducted. The next part is research methodology, followed by analysis and discussion. The final part from this study will be conclusion and further research suggestions.

LITERATURE REVIEW:

ACADEMIC ACHIEVEMENTS:

Academic achievement is a measurement for students' academic performance. The measurement of students' academic performance is Grade Point Average (GPA). GPA can describe the students' performance during their study (Martinez et al., 2015). Bahji, et. al. (2013) showed that there are some factors that influence the learning process. Motivation and engagement are two factors that influence the process and its result. Learning can bring some changes in human knowledge and behavior, including social, psychology, and other kind of changes.

To assess the learning process and how the process is achieved, students should do some assignments and take exams. Through assessment, it can help to bring some valuable information about learning process (Linn and Gronlund, 1995). The assessment can be a score and it is interpreted as an academic achievement. Others think that combination between score and standard achievement process provides two sets of data refers to students achievement.

Chen and Liao (2013) had different perspective towards academic performance. Academic performance is not only a score and measurement of how students perform in academic level. Academic performance is also an effect of strategy during study period. It is a result of implementation for study strategy (Chen and Liao, 2013).

EDUCATIONAL DATA MINING:

Rapid growth in high volume of data mandates the development of data mining algorithm. "*Data mining is an analysis and non-trivia extraction from the data in database to discover new and valuable information, in the form of pattern and rules, from the relationship between elements in data*" (Hirji, 2001). Data mining has some steps before it discovers knowledge from data. Those steps are preparation of data, data transformation, data mining, and finally the knowledge discovery (Suhirman, Zain, and Herawan, 2014)

Suchita and Rajeswari (2013) have similar perspectives with Hirji (2001). Data Mining is a process to analyze data from different perspectives and summarize it into valuable information to identify patterns in big datasets. The main function of data mining is to implement techniques and algorithms with the purpose to detect and extract patterns, artificial intelligence, and visualization techniques (Hand, Mannila, and Smyth, 2001)

The development of data mining algorithms and methods promote the usage of data mining in education sectors as well. Educational Data Mining has been implemented to help education sector to analyze educational data (Faulkner, Davidson, and McPherson, 2010). Educational Data Mining has been implemented in research areas, such as e-learning system, smart tutor application, text mining in some course outline, social media to discover how people learn using social media. Educational Data Mining changes raw data into valuable information (Suhirman et. al., 2014).

Martinez, et. al. (2015) used data mining techniques to analyze students' academic achievement and performance in Algorithm and Data Structure subject. As a result, the study was able to infer the students' profile. The study proved that how data mining can make clusters to categorize students' academic performance. Other researchers use Educational Data Mining to communicate interactively between students and its tutor. It develops smart tutor applications.

Ayan and Garcia (2008) used statistical methods to identify the most suitable approach in terms of goodness of fit and predictive power. The grades awarded in basic scientific courses and demographics variable were entered into the model at the first step. There are four predictors used to predict the academic performance. Those four predictors are Sex (male and female), type of high school (Medicine and Engineering), Type of middle school institution (Public and Private), and Location (Montevideo and Inland). The result for this study found that the students with better grades in the first year of faculty have less risk of curricular lag in the future (Ayan and Garcia, 2008). On the contrary, Bydzovska and Popelinsky (2014) investigated how Educational Data Mining can help predict weak and good students.

Bydzovska and Popelinsky (2014) try to predict weak and good students is an important thing. In this study, the Linear Regression model was used to discover the best fitting machine learning. The result for its prediction helped the teachers with their estimation of students' skills. Linear regression can help to predict more accurate than the other methods (Bydzovska and Popelinsky, 2014).

RESEARCH METHODOLOGY:

In this study, students data and students' grade is being used. Researchers used 362 dataset which is used to determine the correlation between students 1st grade and its final grade. This research used students who enter the university from 2010 to 2015. This study includes odd (August – December) and even semester (January – June). Data is gotten from the Duta Wacana Christian University information center.

PREPROCESSING DATA:

Information Systems students should take 132 credits for mandatory subjects and 12 credits for optional subjects. The first step to conduct this research is to group mandatory subjects into 5 different studies with different weightage:

(1) Programming; (2) Information System Concept; (3) Organization and Management; (4) Logic and Mathematic; (5) Other supporting study. Weight for different study is calculated with equation below:

Weight = Number of Credits / Average of Credits (1)

From the equation (1), it is used to calculate weight for every different study. Table 2 below indicates different weight for different study. It is show that Programming has the highest weight comparing with other studies. After calculating the weight, 1st year weighted GPA is calculated 1st year GPA * Weight for number of credits in 1st semester.

Table 2: Different Weight for Different Study Group

Study Group	Number of Credits	Weight
Programming	36	1.25
Information System Concept	32	1.11
Organization and Management	27	0.94
Logic and Mathematic	25	0.87
Other supporting study	24	0.83
Average	**28.8**	

TRANSFORMING AND CLEANING DATA:

Following the preprocessing data, researchers transform the letter grades into numeric grades. Data is also cleaned and deleted students data whose entry year is before 2010. In this process, researchers got 5 groups of data : (1) graduate profile; (2) 1st year GPA; (3) 1st semester and 2nd semester GPA; (4) Numeric grades for every subjects taken in 1st and 2nd semester; (5) 1st semester and 2nd semester weighted GPA.

ANALYSIS AND DISCUSSION:

REGRESSION MODEL:

After transformation and cleaning, researchers split the data into training and testing data. 80% of data becomes training data and the rest of data becomes testing data. Training data consists of students who entered the university in 2010 (65%), 2011 (28%), and 2012 (7%). This training data is used to find correlation and form Linear Regression between Graduates GPA and 1st semester GPA, 2nd semester GPA, numeric grades for every taken subjects and weighted GPA. As a dependent variable is GPA and others are independent variable.

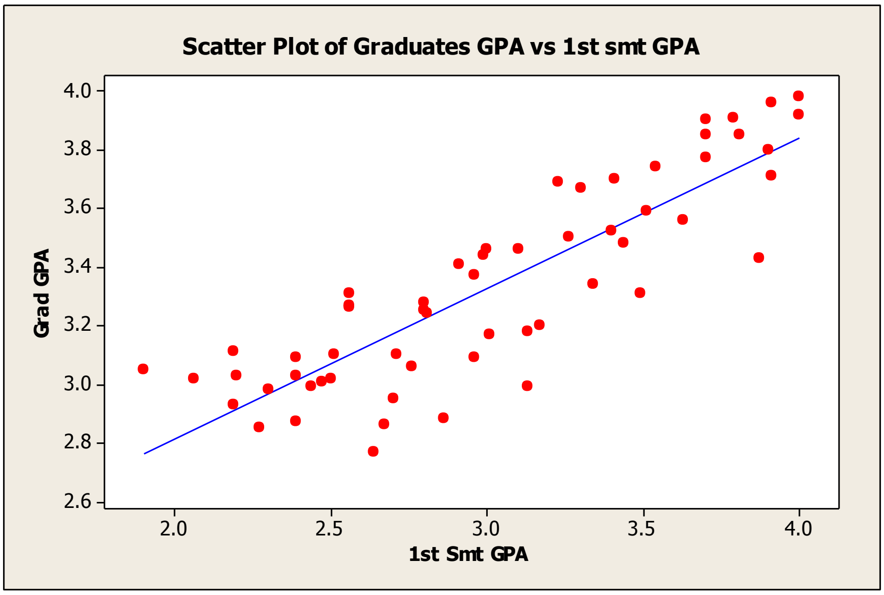

Figure 1: Relationship between Graduates GPA and 1st Semester GPA

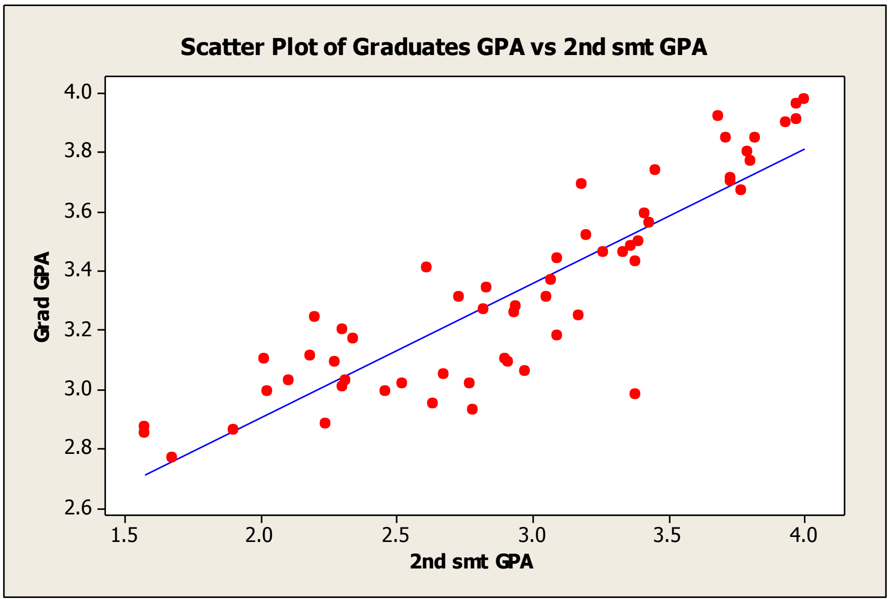

Figure 2: Relationship between Graduates GPA and 2nd Semester GPA

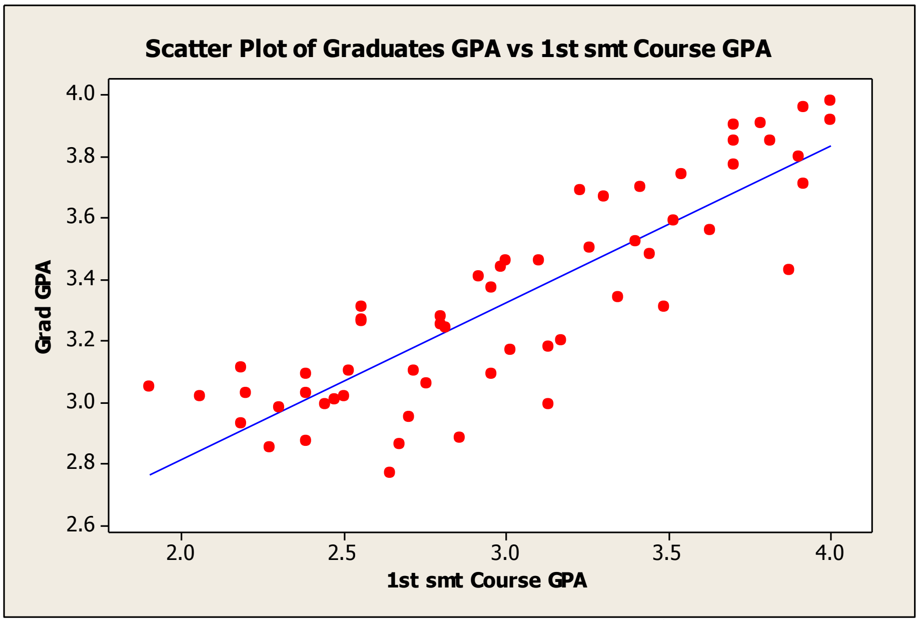

Figure 3: Relationship between Graduates GPA and 1st semester course GPA

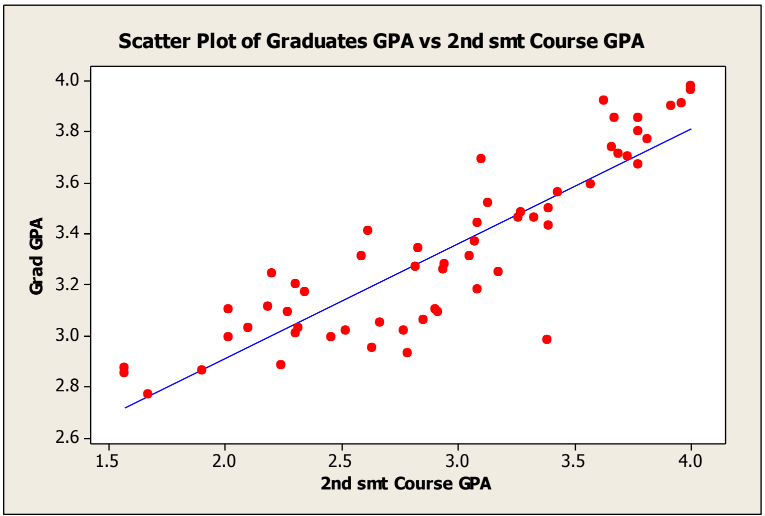

Figure 4: Relationship between Graduates GPA and 2nd Semester course GPA

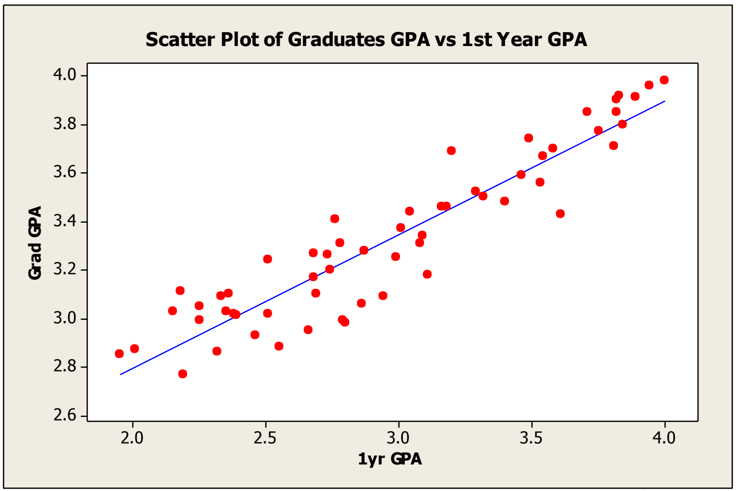

Figure 5: Relationship between Graduates GPA and 1st year GPA

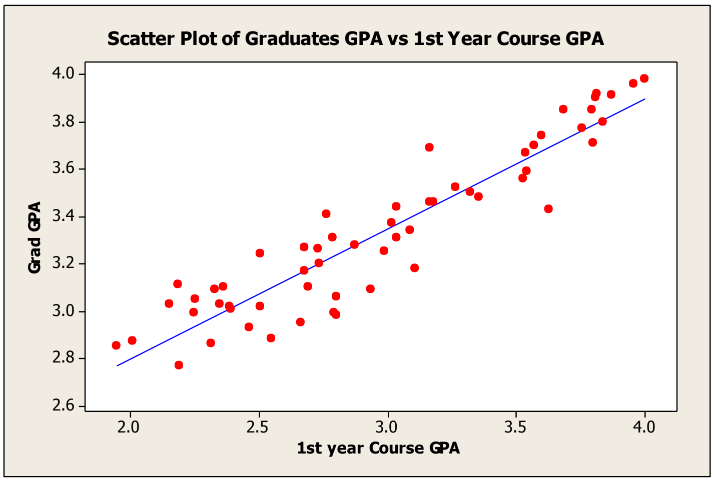

Figure 6: Relationship between Graduates GPA and 1st year course GPA

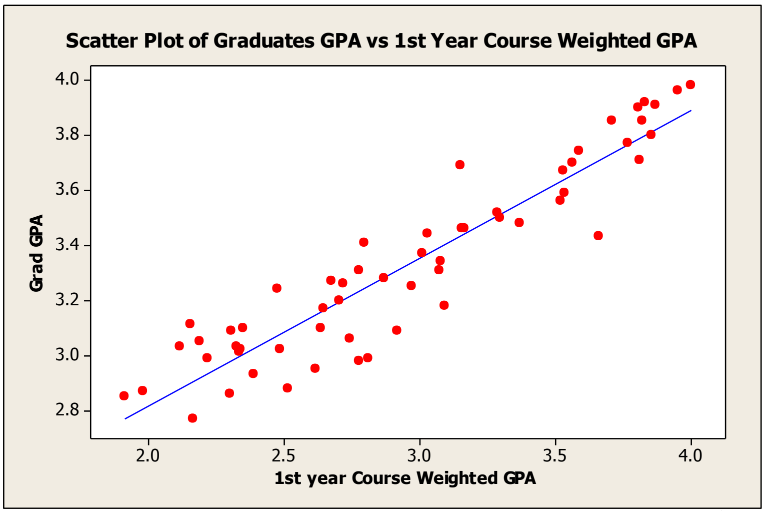

Figure 7: Relationship between Graduates GPA and weighted 1st year GPA

From figure 1 to 7, there is slightly different between graduates GPA with other independent variables. It means that there is no significance difference between using GPA (either using 1st year or semesters in 1st year), course GPA (either 1st or 2nd semester course GPA), and weighted GPA. From the figure 1 to 7, it is also found that there is a positive linear relationship exists between two variables. Next step, this linear regression is tested to training data. Three models is used:

1. Linear regression between graduates GPA and 1st year, 1st semester, and 2nd semester as independent variable.
2. Linear regression between graduates GPA and course GPA (1st year, 1st semester, and 2nd semester) as independent variable.
3. Linear regression between graduates GPA and weighted course GPA (1st year, 1st semester, and 2nd semester).

From the training model, the result shows in table 3 below:

Table 3: Regression Test and Correlation between Graduates GPA and Independent Variables

		Independent Data								
		Only 1^{st} smt data		**Only 2^{nd} smt data**		**1^{st} Year data**		**1^{st} smt (X1), 2^{nd} smt (X2)**		
		Const	**X1**	**Const**	**X2**	**Const**	**X1**	**Const**	**X1**	**X2**
Based on 1^{st} year GPA	**Regression Coef**	1.79	0.513	2	0.454	1.69	0.551	1.7	0.27	0.28
	T Value	14.4	12.68	20.46	14	19.52	19.24	18.74	6.03	7.16
	F Anova Value	F = 160.76		F = 195.94		F = 369,99		F = 178.14		
	R^2 (adj) Value	R^2 (adj) = 73.7 %		R^2 (adj) = 77.4 %		R^2 (adj) = 86,6 %		R^2 (adj) = 86.1 %		
	No of unusual obs	4		2		2		2		
	Correlation	r = 0.861		r = 0.882		r = 0,932		na		
Based on Grade of 1^{st} year Courses Taken	**Regression Coef**	1.79	0.512	2.01	0.452	1.69	0.551	1.7	0.272	0.276
	T Value	14.43	12.67	20.56	13.94	19.62	19.34	18.88	6.2	7.28
	F Anova Value	F = 160.47		F = 194.32		F = 373.85		F = 181.22		
	R^2 (adj) Value	R^2 (adj) = 73.7 %		R^2 (adj) = 77.2		R^2 (adj) = 86.7 %		R^2 (adj) = 86.3		
	No of unusual obs	4		2		3		3		
	Correlation	r = 0.861		r = 0.881		r = 0.933		na		
Based on 1^{st} year Weighted GPA	**Regression Coef**	1.82	0.504	2.06	0.436	1.74	0.536	1.76	0.266	0.266
	T Value	15.14	12.81	22.35	14.17	21.02	19.5	20.25	6.16	7.31
	F Anova Value	F = 164.11		F = 200.71		F = 380.13		F = 185.52		
	R^2 (adj) Value	R^2 (adj) = 74.1 %		R^2 (adj) = 77.8 %		R^2 (adj) = 86.9 %		R^2 (adj) = 86.6 %		
	No of unusual obs	4		1		4		4		
	Correlation	r = 0.863		r = 0.884		r = 0.934		na		

From the table 3, it is shown that for every test, p-value = 0.0. It means that there is linear strong relationship Graduates GPA with GPA in the specific year / semester, Course GPA in specific year / semester, or weighted GPA in specific year / semester. The result of R2 is also supported this finding. It means that graduates GPA is determined by academic achievement in the first year.

From the table 3, it is also found that there is strong correlation between Graduates GPA and academic achievement in first year. The correlation is positive. It means that higher academic achievement in first year, graduates GPA will be higher.

TESTING THE REGRESSION MODEL:

Regression model in the table 3 is tested using 20% of training data. Standard error level is measured using Mean Absolute Deviation using following equation:

$$MAD = \frac{\sum |y - \hat{y}|}{n} \qquad (2)$$

$$\%MAD = \frac{MAD}{\bar{y}} *100\%$$

With:
y = GPA of graduates data test
$\hat{y}$ = Predicted GPA of graduates data test
$\bar{y}$ = Average of GPA of graduates' data test
n = number of data test (in this case = 14).

As the result, it is shown in table 4 below:

Table 4: Testing Result to Regression Model in Table 3

		Only 1st smt data	Only 2nd smt data	1st Year data	1st smt (X1), 2nd smt (X2) data
Based on 1st year GPA	MAD	0.0923	0.1244	0.085	0.0848
	% MAD	2.8	3.78	2.58	2.57
Based on Grade of 1st year Courses Taken	MAD	0.0929	0.125	0.086	0.0861
	% MAD	2.82	3.8	2.61	2.61
Based on 1st year Weighted GPA	MAD	0.094	0.1216	0.085	0.0857
	% MAD	2.85	3.69	2.58	2.6

From the table 4 shows that Mean absolute difference for every single variable testing is less than 0.2 and % MAD is below 4%. It means that the regression model is fit to the training data. And it is supported the finding that the graduates GPA has strong relationship with first year academic achievement.

FINDING DISCUSSION:

From the findings in table 3 and 4, some there are two discussion:

1. **Regression Model Accuracy:**
   From the result in table 3 and 4, it is shown that regression model is satisfied and fit with training data. It can be shown from training result, where MAD is less than 3.8%. This number indicates that linear regression model using first year academic achievement is accurate. This regression model can predict the graduates GPA. Regression model using two variables, as it is shown in table 4 column 6 does not generate escalation significant result of the model.
2. **Data Usage:**
   The most accurate to predict the graduates GPA is using academic achievement during 1st year (1st semester and 2nd semester). While the less accurate to predict the graduates GPA is using academic achievement during 2nd semester (2nd semester GPA). The using of 1st year academic achievement can be more accurate 0.2% than others model. On the contrary, weighted predictor data is slightly significant difference
3. **The Regression Model:**
   GPA, as a measurement of academic achievement is a valid measurement (Najafabadi, A. T., Najafabadi, M. O., & Farid-Rohani, M. R., 2013). By using GPA, it is clearly understood how students can perform in their academic activities. Regression model as a tool to predict the graduates GPA using 1st year academic achievement is also a valid prediction. It can be measure by the model tested for training data.

CONCLUSION:

1. Linear regression model using 1st year GPA is the most accurate model to predict graduates GPA. In the training data, R2 value is 86.9% with correlation level 0.934. In the testing model, Mean Absolute Deviation value is 0.085 (2.57%).
2. The most optimal regression model use either 1st year Academic achievement or 1st semester academic achievement as a predictor with mean absolute deviation 2.58% and 2.8%. The usage of more complex regression model (using 2 independent variables) does not produce significant escalation. It also happens if weighted predictor data is used

REFERENCES:

[1] Ayan, M. N., & Garcia, M. T. (2008). Prediction of University Students' Academic Achievement by Linear and Logistic Model. *The Spanish Journal of Psychology, 11*(1), 275 - 288.

[2] Bahji, S. E., Lefdaoui, Y., & Alami, J. E. (2013). Enhancing Motivation and Engagement: A Top Down Approach for the Design of a Learning Experience According to the S2P-LM. *International Journal of Emerging Technologies in Learning, 8*(6).

[3] Bydzovska, H., & Popelinsky, L. (2013). Weak Students Identification: How Technology can Help. *Proceedings of the European Conferene on e-Learning*, (pp. 89-97).

[4] Chen, M.-h., & Liao, J.-L. (2013). Correlations among Learning Motivation, Life Stress, Learning Satisfaction, and Self-Efficacy for Ph.D Students. *The Journal of International Management Studies, 8*(1), 157-162.

[5] Faulkner, R., Davidson, J. W., & McPherson, G. E. (2010). The value of data mining in music education research and some findings from its application to a study of instrumental learning during childhood. *International Journal of Music Education, 28*(3), 212-230.

[6] Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining.* Cambridge: MIT Press.

[7] Hirji, K. K. (2001). Exploring data mining implementation. *Communications of the ACM, 44*(7), 87-93.

[8] Linn, R. L., & Gronlund, N. E. (1995). *Measurement and Evaluation in Teaching, 7th edition. Englewood Cliffs.* New Jersey: Prentice Hall.

[9] Martinez, D. L., Karanik, M., Giovannini, M., & Pinto, N. (2015). Academic Performance Profiles: A Descriptive Model Based on Data Mining. *European Scientific Journal, 11*(9), 17-38.

[10] Najafabadi, A. T., Najafabadi, M. O., & Farid-Rohani, M. R. (2013). Factors contributing to academic achievement: a Bayesian Structure Equation Modelling Study. *International Journal of Mathematical Education in Science and Technology, 44*(4), 490-500.

[11] Siang, J. J., & Santoso, H. B. (2016). Learning Motivation and Study Engagement: Do They Correlate with GPA? An Evidence From Indonesian University. *Researchers World, VII*(1), 111-118.

[12] Suchita, B., & Rajeswari, K. (2013). Predicting students academic performance using education data mining. *International Journal of Computer Science and Mobile Computing, 2*, 273-279.

[13] Suhirman, Zain, J. M., & Herawan, T. (2014). Data Mining for Education Decision Support: A Review. *International Journal of Emerging Technologies in Learning, 9*(6), 4-19.
