DOI : 10.18843/rwjasc/v8i2(1)/08
DOI URL : http://dx.doi.org/10.18843/rwjasc/v8i2(1)/08

LOGISTIC REGRESSION MODEL FOR PREDICTING FIRST SEMESTER STUDENTS GPA CATEGORY BASED ON HIGH SCHOOL ACADEMIC ACHIEVEMENT

R. Gunawan Santosa,

Department of Informatics
Duta Wacana Christian University,
Indonesia.

Antonius Rachmat Chrismanto,

Department of Informatics
Duta Wacana Christian University,
Indonesia.

ABSTRACT

Faculty of Information Technology, Duta Wacana Christian University (DWCU) has two methods of admitting new college students in which the first is on the basis of educational achievements in high school whereas the second is on the basis of regular test entrance examination. This research will seek forms of functional relationships through logistic regression to the first semester GPA category of the student in Faculty of Information Technology, Duta Wacana Christian University. The first semester GPA category is used as the dependent variable and the location of high school, high school class, high school status, and level of English test result are used as independent variables. With regards to the training data required to create a logistic regression model, we used students' admision data from 2008 through 2014, while the students' data of 2015 is taken as the testing data. The accuracy of the model in predicting data, is measured by the percentage of correct predictions in this regard through Crosstab tables between the predicted data and the real observation of the 1st semester GPA from new students of 2015 class. This research found seven models. The highest percentage of correct predictions between the logistic regression model and training data is 79.4% .There is a change in the logistic regression form, three models influence only by level of English test result while the remaining four models are influenced by the level of English test result and high school location.

Keywords: logistic regression, prediction, accuracy model, Crosstab table.

INTRODUCTION:

One of the important stakeholders in university is students. Before entering Duta Wacana Christian University (DWCU), students have had different educational and environment backgrounds. The different environments background means having a secondary school they have taken or take. For example, those coming from senior high school in Java and outside Java. And also they maybe come from ordinary high school or vocational high school. The last, they also maybe come from public high school or private high school high school. After prospective new students are admitted as college students at Faculty of Information Technology, DWCU and after learning process in the first semester, the results of their study are reflected in their first semester Grade Point Average (GPA). The GPA of a student can be regarded as evaluation of learning outcomes in higher education. There are some students who secured high GPA whereas some obtained low GPA. At the first semester, these students can sometimes move to another faculty. In addition, there are some students who have not registered for the next semester, so that they are no longer active as college students or even resign as a student from the Faculty of Information Technology DWCU.

There are many different external factors that may affect the first semester GPA beside existing internal factors such as interests, talents, abilities, diligence, self-motivation. External factors that are used in this research are the location of high school, categories of high school, status of high school, and level of English test result. The location of high school can be categorized into high schools in Java and outside Java and the high schools are classified by usual high school and vocational high school which also thought to be investigated. Similarly, the status of high school is categorized into high school with the status of State and Private status. In addition, the level of English among the students is also a possible factor affecting students' first semester GPA. In this research we only limit on factors as mentioned above. Actually there are many more factors that may affect the GPA categories that is not covered, such as: gender, interest in studies, family economic level, employment background of parents / guardians and others.

In a previous research, Santosa & Setiadi (2015) have investigated about the factors that affect students' GPA significantly, but they have not observed a relationship between GPA and these factors. This research is the next step of previous research which essentially tries to find mathematical models such as logistic regression to predict the categories of students' first semester GPA at the Faculty of Information Technology, DWCU.

LITERATURE REVIEW:

DATA MINING:

Due to the internet development, there is a rise in global data usage. Data is a collection of facts that have been valid and ready to be used for the purposes of seeking information. Information can be obtained in various ways, either manually or automatically. If the amount of data is too large, then the information search becomes increasingly difficult, especially if it is performed manually. A search for information from a large number of datasets can be done automatically using Data Mining. Data Mining (Knowledge Discovery from Data) is a discipline in Information Technology that is interdisciplinary and is able to perform mining of data in large numbers in order to obtain various kinds of information in the form of certain patterns that are hard to find by searching manually (Rud, 2010).

Stages of data mining in knowledge discovery process is as follows (Han & Kamber, 2011):

1. Data collection phase (either from a database or other data source): In this research, the data retrieved from the achievements of high school students who come from the path of achievement in the Faculty of Information Technology UKDW act as the data source.
2. The pre-processing stage is the stage of data cleaning and data processing early, in order to be completely ready for use.
3. Phase transformation is the data transformation from one phase to another so that it becomes suitable for the data to process to next stage in numeric form i.e., ready to be calculated. This stage also features selection i.e., selection stage of all the right attributes as the distinguishing feature data.
4. Data mining is the stage when the data mining algorithms are applied. Few examples in this stage are clustering, regression, or classification algorithm. In this research, we use logistic regression algorithm analysis.
5. Phase evaluation is the stage to check the validity, accuracy, precision stages of data mining whether it has been done before, so that the data mining result phase can be used well for a variety of data other than testing.

Several studies using data mining to analyze the concept of academic ability in college were conducted earlier, such as:

- (Aziz, Ismail, & Ahmad, 2013) conduct research by building a prototype system to predict Students Academic Performance (SAP) in Higher Education Institution (HLI) using Naive Bayes algorithm based on Weka. This system can provide early warning of potential weaknesses in the students' learning. Parameters that are used in this research were gender (male / female), race (Malaysia, China, or India), origin (from town or rural), how to get into university (STPM, diploma, or matriculation), parents income, and the first semester GPA (high, low, and medium)
- (Ahmad, Ismail, & Aziz, 2015) conduct research related to data mining to predict the IP category of the first half using Decision Tree algorithm, Naive Bayes, and Rule Based Classification. Based on these studies showed that the Rule Based Classification is the best method among the three methods above, with the accuracy reached 71.3%. In this study they divided GPA into 3 categories: low, medium and high level.
- (Thakar, Mehta, & Manisha, 2015) presented a paper which contains extensive and exhaustive survey on rise of data mining to education and its direction in the future.
- The use of Data Mining is also used to predict the likelihood of students dropped out (Jadrić, Garača, & Ćukušić, 2010) and the possibility of student's failed (Vadivu & Bharathi, 2014).

GRADE POINT AVERAGE (GPA):

Every student in the university has to be evaluated at the end of every semester to evaluate their learning outcomes. According to Indonesian Technology Research Higher Education Department (RISTEK DIKTI) (RISET, TEKNOLOGI, & DAN PENDIDIKAN, 2015), Grade Point Average (GPA) is the result of assessment outcomes of learning for university students at the end of semester. It is expressed in value calculated by summing the multiplication of the course grade point taken and course credits. Then this result is divided by sum of all gradable credits courses have been taken, as express in formula (1) below:

$$GPA = \frac{\sum Course\ Grade\ Point * Course\ Credits}{\sum Gradable\ Credits} \ldots (1)$$

GPA is important because it relates to:

1. The number of courses that can be taken in the following semester.
2. At the end of the graduation in university, GPA is considered as an important determination to apply for jobs.

In a previous study (Santosa & Setiadi, 2015), first semester GPA affects on another next semester's GPA. So there is a strong correlation between first semester GPA and GPA of the next semester. Therefore first semester GPA is the main focus of this research.

LOGISTIC REGRESSION:

Logistic regression was developed and applied in epidemiological research. The common question to this research is "what is the relationship between one or more independent variables (exposure) on the condition of the illness". The condition of the illness is worth a dichotomous variable (binary) i.e 0 represents "no sick" and 1 represents "the sick". (Kleinbaum & Klein, 2002) do research about classification of Coronary Heart Disease (CHD) and result the dichotomy are the status of subjects classified as 0 ("no CHD") and 1 ("CHD"). In this research, the independent variables used are x_1 as age, x_2 as nationality and x_3 as gender. Figure 1 illustrates the schematic model of logistic regression.

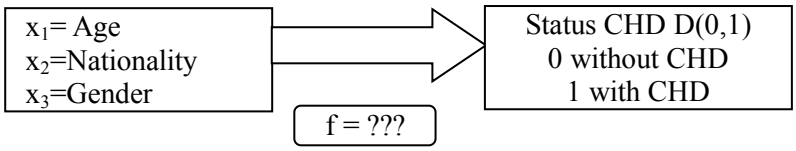

Figure 1: The schematic model of logistic regression

It can also be said that logistic regression model in Data Mining can be used to describe the relationship of several independent variables X on the dependent dichotomy value variable D (0,1) (Kantardzic, 2003) . For epidemiology researchers, this logistic regression is the most popular procedure, especially when the data is about dichotomous disease conditions. Logistic regression is based on a logistic function which can be seen in the formula [2]:

$$f(z) = \frac{1}{1+e^{z}} \text{ dengan } -\infty < z < \infty \ldots (2)$$

The form of logistic function is described in figure 2 below (Hosmer & Lemeshow, 2003)

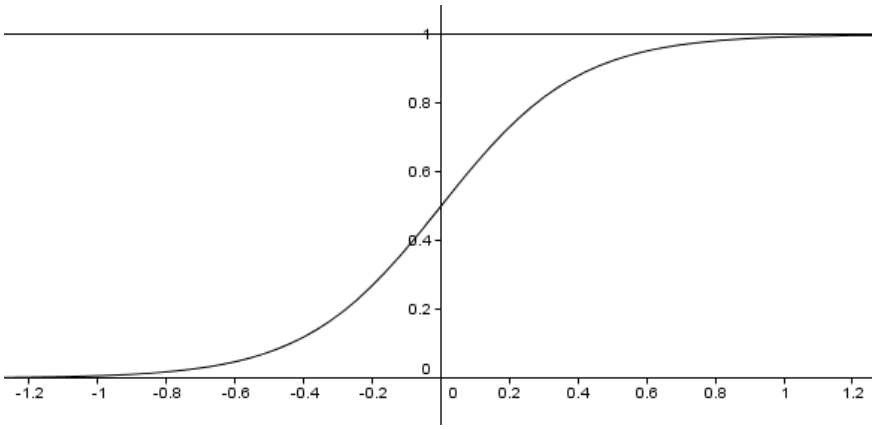

Figure 2: Graph of logistic function

The Figure 2 shows the range the value of f (z) lies from 0 to 1 regardless of the value of z. For z → ∞, then f (∞-) = 0 and for z → ∞ , then f (∞) = 1. Since the value of f (z) lies between 0 and 1, the logistic regression is popular. This model is designed to describe the probability that always lies between 0 and 1. In epidemiological terms, probabilities like these give the value of an individual's risk of getting certain diseases (Hosmer & Lemeshow, 2003).

Another reason why logistic model became popular is because of the shape of the logistic function. As observed in Figure 2, if we start from z = ∞ move to the right, then the value f (z) rose sharply curved approaches z = 0 for a while, then it gets declined dramatically close to 1 and eventually will approach around 1. These results illustrate the shape of the curve S.

Application of the logistic regression was performed by (Hajarisman, 1998) to be implemented in order to determine the pattern graduation TPB-IPB. In this study, logistic regression was compared with beta-binomial regression method since beta-nomial method is easier for both calculating the probability function decline and interpreting the parameters that are inside the model. In this research, it has been shown that beta-binomial regression model is a tool that is able to address overdispersi in clustered binary observational data compared to the usual binary logistic regression model. This research is the logistic regression model f and the main focus is z can be written as a linear combination of x_1, x_2, x_3 ..., x_k, which can be seen in the formula (3) and (4):

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \ldots (3)$$

so

$$y = \hat{f}(z) = \frac{1}{1+e^{-(\alpha+\sum \beta_i x_i)}} \ldots (4)$$

The values α and β_i are unknown parameters and will be estimated from the data set in the form of samples using Maximum Likelihood Estimator (MLE). The parameters α and β_i can be estimated by $\hat{\alpha}$ and $\widehat{\beta_i}$.

CROSSTAB TABLE :

To measure the accuracy of prediction, logistic regression model can be used Crosstab (Cross tabulation) between the prediction categorical data and the observation categorical data value as seen at Table 1:

Table 1: Crosstab between Observation Data and Prediction Data

		Prediction Data		Total Row
		Category 1	Category 2	
Observation Data	Category 1	n_{11}	n_{12}	b_1
	Category 2	n_{21}	n_{22}	b_2
Total Column		k_1	k_2	n

The accuracy of the prediction results measured by Formula (5) below (Bhattacharyya & Johnson R.A., 1977):

Equality Percentage EP = (n_{11} + n_{22}) / n …(5)

with

n_{11} = number of training data which is in Category 1 in Prediction and actually this data is include in Category 1 in Observation

n_{22} = number of training data which is in Category 2 in Prediction and actually this data is include in Category 2 in Observation
n = number of all training (test) data.

METHODOLOGY:

Several stages of the research is as follows;

1. The students' data is retrieved from Faculty of Information Technology from PUSPINDIKA which consists of Information System Department and Information Technology Department. A total of 8 consecutive academic years of student data was retrieved from 2008 to 2015. The data retrieved from PUSPINDIKA includes high school category, high school status, high school location and first GPA students.
2. The English skills data of prospective students from PPBA (Pusat Pelatihan Bahasa Asing) was also retrieved.
3. The researchers chose the data from Faculty of Information Technology students that received through achievement. In the process of new student admissions in DWUC, there are two ways which are path of high school achievement line and non-high school achievement lines (regular admission through entrance test). The difference between these two paths are in the process of acceptance in case of path to achievement is not to take the test academic skills (ability of Numerical, Verbal, Spatial and Analogies tests) while the non-achievement path should follow the four tests of academic ability.
4. Then, after cleaning process, the data is used for logistic regression.
   In logistic regression model, the main focus is z which can be written as a linear combination of x_1, x_2, x_3, ..., x_k, and can be seen in the Formula (3) and (4) above.
   x_1, x_2, x_3, ..., x_k are value from training data. We use x_1, x_2, x_3, ..., x_k to get parameter value of α and β_i which is unknown parameters. Then we estimate α and β_i by using Maximum Likelihood Estimator (MLE) symbolized as $\hat{\alpha}$ and $\hat{\beta}_i$. Furthermore $\hat{\alpha}$ and $\hat{\beta}_i$ will be used for predicting y as seen in Formula (4).
5. Making 7 logistic regression models for new Information Technology Faculty students through high school achievement line. The seven models are as follows,
   - Model A is the logistic regression model using data from student class of 2008
   - Model B is the logistic regression model using student data from 2008 up to 2009 class
   - Model C is the logistic regression model using data from 2008 students to the class of 2010
   - Model D is the logistic regression models using data from 2008 students to the class of 2011
   - Model E is the logistic regression models using data from 2008 students to the class of 2012
   - Model F is the logistic regression models using data from 2008 students to the class of 2013
   - Model G is the logistic regression models using data from 2008 students to the class of 2014
6. With the help of these 7 logistic regression models, the first semester GPA category class of 2015 is predicted.
7. Measuring accuracy of the results of logistic regression prediction of seven such models to actual observed category class of 2015 data by using Crosstab table (cross tabulation).

FINDINGS AND DISCUSSION:

After the data is collected, it is cleaned in order to obtain the cumulative number of students at Faculty of Information Technology in order to run logistic regression. The data is tabulated as Table 2 below.

Table 2: Number of students of the Faculty of Information Technology from 2008 until 2015

No	Class	Total	High School Achievement	Regular Admission
1	2008	368	63	305
2	2008 until 2009	628	72	556
3	2008 until 2010	892	128	764
4	2008 until 2011	1143	273	870
5	2008 until 2012	1387	398	989
6	2008 until 2013	1592	523	1069
7	2008 until 2014	1763	613	1150
8	2015	254	193	61

Because the dependent variable should dichotomy of data D (0,1), then the first GPA is transformed into binary data, namely:
0 < GPA_1 < average coded into "0" and (Average +0.0001) < GPA_1 <4.000 encoded into "1", in this case the average is taken based on the entire sample data.
The next step is *data transformation* (data is transformed or converted into a form suitable for processing or analysis). Data category of SMA ("1" for high school and "2" for vocational high school), the status of SMA ("1" to the State high school and "2" for the Private high school), the location of SMA ("1" to high school from Java and "2" for high school from outside Java) and level English (level_1 encoded into "1", level_2 encoded into "2", level_3 encoded into "3" and ESP encoded into "4") transformed from alpha numeric data into numerical data because logistic regression data requires a numeric type data. So that the model has only four independent variables: the status of high school (x1), the location of SMA (x2), category SMA (x3), the level of English (x4). As in the case of ordinary regression of existing regression model can be simplified to the significance test for the beta coefficient (β). Here is a form of models of computation results before and after simplified (Santosa & Chrismanto, 2016).
Model A is the logistic regression model using high school achievement data from class of 2008 is;

$$P(\mathbf{x}) = 1/(1+\exp\{-[-22{,}567 - 0{,}490(x_1) - 0{,}015(x_2) + 20{,}124(x_3) + 1{,}523(x_4)]\})$$

and after a simplified model becomes:

$$P(\mathbf{x}) = 1/(1+\exp\{-[-1{,}460 + 1{,}420(x_4)]\})$$

Model B is the logistic regression model using high school achievement data from class of 2008 until class 2009 is;

$$P(\mathbf{x}) = 1/(1+\exp\{-[-22{,}060 + 0{,}587(x_1) - 0{,}540(x_2) + 21{,}206(x_3) + 0{,}802(x_4)]\})$$

and after a simplified model becomes:

$$P(\mathbf{x}) = 1/(1+\exp\{-[-0{,}381 + 0{,}698(x_4)]\})$$

Model C is the logistic regression model using high school achievement data from class of 2008 until class 2010 is;

$$P(\mathbf{x}) = 1/(1+\exp\{-[-0.565 + 0{,}050(x_1) - 0{,}648(x_2) + 0{,}664(x_3) + 0{,}768(x_4)]\})$$

and after a simplified model becomes:

$$P(\mathbf{x}) = 1/(1+\exp\{-[-0{,}793 + 0{,}833(x_4)]\})$$

Model D is the logistic regression model using high school achievement data from class of 2008 until class 2011 is;

$$P(\mathbf{x}) = 1/(1+\exp\{-[1{,}895 - 0{,}181(x_1) - 1{,}438(x_2) - 0{,}125(x_3) + 0{,}573(x_4)]\})$$

and after a simplified model becomes:

$$P(\mathbf{x}) = 1/(1+\exp\{-[1{,}409 - 1{,}404(x_2) + 0{,}582(x_4)]\})$$

Model E is the logistic regression model using high school achievement data from class of 2008 until class 2012 is;

$$P(\mathbf{x}) = 1/(1+\exp\{-[1{,}509 - 0{,}236(x_1) - 1{,}160(x_2) + 0{,}320(x_3) + 0{,}465(x_4)]\})$$

and after a simplified model becomes:

$$P(\mathbf{x}) = 1/(1+\exp\{-[1{,}350 - 1{,}057(x_2) + 0{,}456(x_4)]\})$$

Model F is the logistic regression model using high school achievement data from class of 2008 until class 2013 is;

$$P(\mathbf{x}) = 1/(1+\exp\{-[1{,}625 - 0{,}251(x_1) - 1{,}082(x_2) - 0{,}322(x_3) + 0{,}646(x_4)]\})$$

and after a simplified model becomes:

$$P(\mathbf{x}) = 1/(1+\exp\{-[0{,}809 - 1{,}069(x_2) + 0{,}666(x_4)]\})$$

Model G is the logistic regression model using high school achievement data from class of 2008 until class 2014 is;

$$P(\mathbf{x}) = 1/(1+\exp\{-[1{,}575 - 0{,}160(x_1) - 1{,}121(x_2) - 0{,}367(x_3) + 0{,}712(x_4)]\})$$

and after a simplified model becomes:

$$P(\mathbf{x}) = 1/(1+\exp\{-[0{,}895 - 1{,}149(x_2) + 0{,}738(x_4)]\})$$

Summary result for the logistic regression coefficient before significance data can be summarized in Table 3.

Table 3: Summary Table of Logistic Regression Coefficients before Significance Test

Class	α	β_1	β_2	β_3	β_4	Equility Percentage
2008	-22.567	0.490	-0.015	20.124	1.523	79.4%
2008 sd 2009	-22.060	-0.587	-0.540	21.206	0.802	76.4%
2008 sd 2010	-0.565	0.050	-0.648	0.664	0.768	71.9%
2008 sd 2011	1.895	-0.181	-1.438	-0.125	0.573	75.1%
2008 sd 2012	1.509	-0.236	-1.160	0.320	0.465	71.6%
2008 sd 2013	1.625	-0.251	-1.082	-0.322	0.646	71.5%
2008 sd 2014	1.575	-0.160	-1.121	-0.367	0.712	73.7%

Table 3 shows that the average of its Equility Percentage is 74.1%

Table 4: Summary Table of Logistic Regression Coefficients after Significance Test

Class	α	β_1	β_2	β_3	β_4	Equility Percentage
2008	-1.460	0	0	0	1.420	76.2%
2008 sd 2009	-0.381	0	0	0	0.698	75%
2008 sd 2010	-0.793	0	0	0	0.833	73.4%
2008 sd 2011	1.409	0	-1.404	0	0.582	75.1%
2008 sd 2012	1.350	0	-1.057	0	0.456	72.4%
2008 sd 2013	0.809	0	-1.069	0	0.666	71.9%
2008 sd 2014	0.895	0	-1.149	0	0.738	73.9%

Table 4 shows that the average of its Equility Percentage is 73.9 %

By using all the seven models in Table 4, the first semester GPA category from Information Technology students can be predicted through high school achievement admission. The rows and columns in the crosstab table have two categories such as the category which is labeled as '0' indicating the first semester GPA category below the sample average and another category labeled '1' showing the first semester GPA category above average sample data. And the results is summarized by Crosstab tables.

Table 5: Number of students as a result predictive using Model A

Model A		Prediction Data		Total Row
		0	1	
Observation Data	0	59	47	106
	1	7	80	87
Total Column		**66**	**127**	**193**

Based on Table 5, it can be seen that there are 59 prediction data in the category 1 = 0 and it is also true in the category 1 = 0 corresponds to the actual observation data. There are also 80 prediction data in the category 2 = 1 and it is also true in the category 2 = 1 corresponds to the actual observation data. So the accuracy of prediction using model A has Equality Percentage = 72%.

Table 6: Number of students as a result predictive using Model B

Model B		Prediction Data		Total Row
		0	1	
Observation Data	0	0	0	0
	1	66	127	193
Total Column		**66**	**127**	**193**

Based on Table 6, it can be seen that there are 0 prediction data in the category 1 = 0 and it is also true in the category 1 = 0 corresponds to the actual observation data. There are also 127 prediction data in the category 2 = 1 and it is also true in the category 2 = 1 corresponds to the actual observation data. So the accuracy of prediction using model B has Equality Percentage = 65.8%.

Table 7: Number of students as a result predictive using Model C

Model C		Prediction Data		Total Row
		0	1	
Observation Data	0	0	0	0
	1	66	127	193
Total Column		**66**	**127**	**193**

Based on Table 7, it can be seen that there are 59 prediction data in the category 1 = 0 and it is also true in the category 1 = 0 corresponds to the actual observation data. There are also 127 prediction data in the category 2 =

1 and it is also true in the category 2 = 1 corresponds to the actual observation data. So the accuracy of prediction using model C has Equality Percentage = 65.8%.

Table 8: Number of students as a result predictive using Model D

Model D		Prediction Data		Total Row
		0	1	
Observation Data	0	65	85	150
	1	1	42	43
Total Column		**66**	**127**	**193**

Based on Table 8, it can be seen that there are 65 prediction data in the category 1 = 0 and it is also true in the category 1 = 0 corresponds to the actual observation data. There are also 42 prediction data in the category 2 = 1 and it is also true in the category 2 = 1 corresponds to the actual observation data. So the accuracy of prediction using model D has Equality Percentage = 55.4%.

Table 9: Number of students as a result predictive using Model E

Model E		Prediction Data		Total Row
		0	1	
Observation Data	0	59	47	106
	1	7	80	87
Total Column		**66**	**127**	**193**

Based on Table 9, it can be seen that there are 59 prediction data in the category 1 = 0 and it is also true in the category 1 = 0 corresponds to the actual observation data. There are also 80 prediction data in the category 2 = 1 and it is also true in the category 2 = 1 corresponds to the actual observation data. So the accuracy of prediction using model E has Equality Percentage = 72%.

Table 10: Number of students as a result predictive using Model F

Model F		Prediction Data		Total Row
		0	1	
Observation Data	0	59	47	106
	1	7	80	87
Total Column		**66**	**127**	**193**

Based on Table 10, it can be seen that there are 59 prediction data in the category 1 = 0 and it is also true in the category 1 = 0 corresponds to the actual observation data. There are also 80 prediction data in the category 2 = 1 and it is also true in the category 2 = 1 corresponds to the actual observation data. So the accuracy of prediction using model F has Equality Percentage = 72%.

Table 11: Number of students as a result predictive using Model G

Model G		Prediction Data		Total Row
		0	1	
Observation Data	0	59	47	106
	1	7	80	87
Total Column		**66**	**127**	**193**

Based on Table 11, it can be seen that there are 59 prediction data in the category 1 = 0 and it is also true in the category 1 = 0 corresponds to the actual observation data. There are also 80 prediction data in the category 2 = 1 and it is also true in the category 2 = 1 corresponds to the actual observation data. So the accuracy of prediction using model G has Equality Percentage = 72%.

CONCLUSION:

We arrive at the following conclusion based on the literature.

1. In terms of compatibility between logistic regression models with training data, the model A is the best model, because it has a percentage kesamaaan 79.4%.
2. In terms of compatibility between the logistic regression models with training data after significance test, the model A is the best model, because it has a percentage kesamaaan 79.4%.
3. The best logistic regression model was used to predict first semester GPA category student of 2015 class year are the model A, E, F and G with each match percentage is 72%.
4. There is change in the logistic regression, model A, B, C influence only by level of English. While on the D, E, F and G model, influence by the level of English and high school location. It must be a concern for decision makers in the educational process.
5. The percentage of equality of this research is still not satisfied, so in future we will try to use other methods that can increase the equality percentage or add other variables that have not been taken into our consideration.

ACKNOWLEDGEMENTS:

The authors would like to thank Informatics Department of Faculty of Information Technology, DWCU on all funding and infrastructure so that the completion of this research.

REFERENCES:

Ahmad, F., Ismail, N., & Aziz, A. (2015). The Prediction of Students' Academic Performance Using Classification Data Mining Techniques. *Applied Mathematical Sciences, 9*(129), 6415 - 6426. Retrieved from http://dx.doi.org/10.12988/ams.2015.53289

Aziz, A., Ismail, N., & Ahmad, F. (2013). MINING STUDENTS ACADEMIC PERFORMANCE. *Journal Of Theoretical And Applied Information Technology, 53*(2). Retrieved from http://www.jatit.org/volumes/Vol53No3/21Vol53No3.pdf

Bhattacharyya, & Johnson R.A. (1977). *Statistical Concepts and Methods.* John Wiley & Sons, Inc.

Hajarisman, N. (1998). Kajian Perbandingan Model Regresi Beta Binomial dengan Model Regresi Logistik dan Penerapannya untuk Menduga Pola Kelulusan Mahasiswa TPB-IPB (Magister). *Institut Pertanian Bogor*. Retrieved from https://core.ac.uk/download/pdf/32354709.pdf

Han, J., & Kamber, M. (2011). *Data Mining : Concepts and Techniques.* Morgan Kaufmann Publishers.

Hosmer, D., & Lemeshow, S. (2003). *Applied Logistic Regression.* John Weley and Sons, Inc.

Jadrić, M., Garača, Ž., & Ćukušić, M. (2010). STUDENT DROPOUT ANALYSIS WITH APPLICATION OF DATA MINING METHODS. *Management, 15*(01), 31-46.

Kantardzic, M. (2003). *Data Mining Concepts, Models, Methods and Algorithms.* IEEE Press and Wiley-Interscience.

Kleinbaum, D., & Klein, M. (2002). *Logistic Regression : A Self-Learning Text.* New York: Springer Verlag New York , Inc.

RISET, M., TEKNOLOGI, & DAN PENDIDIKAN. (2015). *PERATURAN MENTERI RISET, TEKNOLOGI, DAN PENDIDIKAN TINGGI REPUBLIK INDONESIA NOMOR 44 TAHUN 2015 TENTANG STANDAR NASIONAL PENDIDIKAN TINGGI* (1st ed.). Jakarta: Departemen Riset Teknologi, dan Pendidikan Republik Indonesia. Retrieved March 01, 2017, from http://kopertis3.or.id/v2/wp-content/uploads/PERMENRISTEKDIKTI-NOMOR-44-TAHUN-2015-TENTANG-SNPT-SALINAN.pdf

Rud, O. (2010). *Data Mining Cook Book Modeling Data for Marketing.* Risk and Customer Relationship management. John Wiley & Sons Inc.

Santosa, R., & Chrismanto, A. (2016). *Regresi Logistik untuk Prediksi Kategori IP Mahasiswa Fakultas Teknologi Informasi UKDW.* Not Published Research Report.

Santosa, R., & Setiadi, H. (2015). *Analisis Faktorial untuk Uji Pengaruh Beberapa Faktor terhadap Indeks Prestasi Mahasiswa Fakultas Teknologi Informasi UKDW.* Not Published Research Report.

Thakar, P., Mehta, A., & Manisha. (2015). Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue. *International Journal Of Computer Applications, 110*(15).

Vadivu, P., & Bharathi, D. (2014). Survey on Students' Academic Failure and Dropout using Data Mining Techniques. *International Journal Of Advances In Computer Science And Technology, 3*(5). Retrieved March 01, 2017, from http://www.warse.org/pdfs/2014/ijacst03352014.pdf

---