Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 2 No. , April 2017, pp. 29 – 36

Estimation of Students' Graduation Using Multiple Linear Regression Method

Bintang Dewi Fajar Kurniatullah *[1], **Yuventius Tyas Catur Pramudi**[2]
[1,2]*Sistem Informasi S1, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro*
Jalan Imam Bonjol No. 207, Semarang 50131, Jawa Tengah, Indonesia
E-mail: bintangkurniatullah@gmail.com[1], , tyascatur@gmail.com*[2]
**Corresponding author*

Abstract - Utilization of students' academic data to produce information used by management in monitoring students' study period on Information System Department. Multiple linear regression methods will produce multiple linear regression equations used for estimating students' graduation equipped with the prototype. According to analysis carried out by using nine variable SKS1, SKS2, SKS3, SKS4, IPS1, IPS2, IPS3, IPS4, and the number of repeated courses of 2008 to 2012 the multiple linear regression equation is Y = 13.49 + 0.099 X1 + (-0.068) X2 + 0.025 X3 + (-0.059) X4 + (-0.585) X5 + (-0.443) X6 + (-0.155) X7 + (-0.368) X8 + (-0.082) X9. From the equation, there is an error of MSE and RMSE that is equal to 0.1168 and 0.3418. The prototype uses a PHP-based program using sublime text and XAMPP. The prototype monitoring the students' study time in this research is very helpful if supported by management.

Keywords: Data mining, multiple linear regression, estimation, monitoring, study time

1. INTRODUCTION

Academic data is a set of data that contains the results of the learning activities during students studying. The student academic data can be in the form of student identity, achievement index of each semester, the cumulative achievement index, time of furlough, the number of courses taken each semester, graduation period, the length of a student's study and others. The use of the academic data in university is still less than maximum except as database of the university, the data is used only for administrative data to complete the accreditation of university. The number of academic data will be directly proportional to the number of students each year. From the result of observation and literature study, academic data can be used optimally to support the improvement of the university. A Large number of academic data can be processed into highly useful information. One of the utilization of academic data is the data can be used as a model system development planning at each college. Some research on prediction of students' graduation rates at the university are Predicting the time of students' graduation with 4 attributes the registration way, the hometown, the previous school, and index of achievement. The result of this research is the value of the accuracy of 82.32% included in the category of *good classification* of *Naive Bayes Classifier* method [1]. *Jaccard Coefficient* method was used to predict the time of graduation produce accuracy levels in this study is 80.65% [2]. Estimating the time of students' study by using the combination of Bayesian network algorithm with k-nearest neighbors method. Analyzing the prediction of students' graduation based on an index of achievement in the first 2 semesters, a score of the national exam, majors at school, graduate school, the registration way to university [3]. Based on several studies in the top of one of the utilization of academic data the

researchers will try to do research on decision-making system model to estimate the time period of a student's study uses data mining algorithms by taking advantage of academic data of Information System Department of Dian Nuswantoro University. The estimation time of graduation will show when the student will graduate using multiple linear regression. Student graduation rates may affect Information Systems Department Accreditation. It thus will be seen in the number of students who get in compared to the number of students who have graduated in one generation.

2. RESEARCH METHOD

The processing of the data can use statistical techniques, mathematics, machine learning and artificial intelligence to dig up information that can be used [4]. By using methods to find patterns and relationships of data which has a large size are called data mining. The use of Linear Regression Estimation methods can determine the pattern of formulas to be used to determine the chances of a new data estimation student study period. Contains a description of the methods to be used. Data that have been acquired is still a row Data to be analyzed further needs to be done sorting the data referred to preprocessing. Data that are not needed is removed so as not to affect the outcome of the analysis. The determination of testing Data used for testing models of linear regression equation produced from the training data. Distribution of training data by 70%, while data on testing 30% of the dataset [5]. Predictor variables (X) are a variable that affects other variables (independent variables). Criterion variable (Y) is a variable that is affected by other variables (the dependent variable). Determining the estimation formula of linear regression equation based on the results of the identification predictor and criterion obtained a linear regression equation. The equation will be used in the estimation of the level of graduation students' of Information System Department.

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8 + b_9X_9 + b_{10}X_{10} \quad (2.1)$$

Where:

Y	= Dependent variable (Dependent)
X	= Independent variable (Independent)
b_o	= Constanta
$b_1 ... b_n$	= Regression direction coefficient

By using the equation above, estimation of the graduation rate of Information System Department student can be performed with multiple linear regressions method [6]. Validation Mean Square Error (MSE) and Root Mean Squared Error (RMSE). The good prediction result is if the value of MSE produced shows little value, or smaller when compared with the results of other prediction methods of calculation. The smaller the number showed by MSE, the higher the accuracy of estimation result [7]. The small value of RMSE will show that variation value generated by the estimating equations approaches the value of observation. Errors in estimating a model is a middle value in the square root RMSE. The mathematical formula is written as the following equation [8].

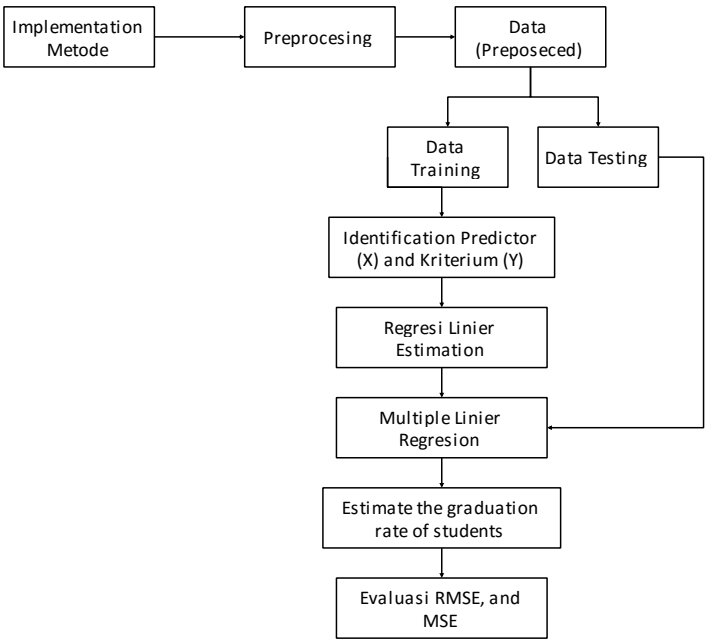


Figure 1. Flow implementation methods

Unified Modeling Language (UML) is one of the visual programming language used to create model and express system using diagrams and supporting texts [9]. UML is used to make a temporal design for the prototype. The prototype that is part of information systems can already be used, but only as an example of early models which will be improved to be easily used by the user [10]. The programming language used Hypertext Preprocessor (PHP) and MySQL.

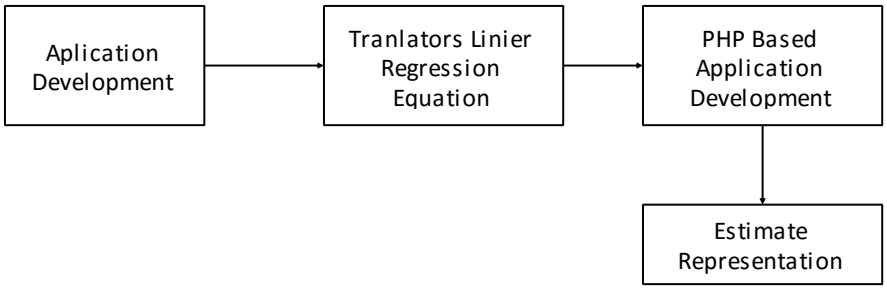


Figure 2. Chronology of the development of implementation methods

3. RESULTS AND DISCUSSION

3.1 Calculation Of Multiple Linear Regression

Multiple linear regression calculations begin by preprocessing the data that is by Eliminating Attributes, eliminating records, and data cleansing. Preliminary data of student of Information Department Dian Nuswantoro University starting in the class of 2008 to 2012 amounted to 1753 records obtained from the Center for Information Data. The data set is generated from the preprocessing stage acquired a number of 1000 record. Then the data is divided into two those are

training data total 700 records and data testing total 300 records. The following attributes are used in this study:

Table 1. Use of Attributes

Attribute	Notation	Attribute	Notation
Estimates passed in the semester to	Y	IPS 2	X_6
SKS 1	X_1	IPS 3	X_7
SKS 2	X_2	IPS 4	X_8
SKS 3	X_3	The number of subjects to repeat	X_9
SKS 4	X_4	constants	A
IPS 1	X_5	The regression coefficient	$b_1....b_n$

The general equation on multiple linear regressions so that the equation to calculate the estimation of students' graduation as follows:
$Y = b_0 + b_1$ (SKS 1) + b_2 (SKS 2) + b_3 (SKS 3) + b_4 (SKS 4) + b_5 (IPS 1) + b_6 (IPS 2) + b_7 (IPS 3) + b_8 (IPS 4) + b_9 (repeated courses)

Creating a multiple linear regression equation to calculate the estimated graduation which will be produced in accordance with the stages of the multiple linear regression equation the least square 10 variable as follows:

1 $700b_0 + 13973b_1 + 14621b_2 + 14064b_3 + 14125b_4 + 1829.81b_5 + 1838.44b_6 + 1907.77b_7 + 1917.1b_8 + 282b_9$ **= 6469**
2 $13973b_0 + 279983b_1 + 292243b_2 + 281216b_3 + 282212b_4 + 36653.91b_5 + 36826.58b_6 + 38247.71b_7 + 38378.86b_8 + 5563b_9$ **= 129005**
3 $14621b_0 + 292243b_1 + 309549b_2 + 295904b_3 + 296203b_4 + 39103.8b_5 + 39060.29b_6 + 40468.12b_7 + 40598.22b_8 + 5280b_9$ **= 133732**
4 $14064b_0 + 281216b_1 + 295904b_2 + 287242b_3 + 286075b_4 + 37265.36b_5 + 37654.65b_6 + 38915.25b_7 + 39048.3b_8 + 4994b_9$ **= 128900**
5 $14125b_0 + 282212b_1 + 296203b_2 + 286075b_3 + 290253b_4 + 37211.7b_5 + 37445.71b_6 + 39222.72b_7 + 39209.83b_8 + 5329b_9$ **= 12953**
6 $1829.81b_0 + 36653.91b_1 + 39103.8b_2 + 37265.36b_3 + 37211.7b_4 + 5108.0119b_5 + 5021.3866b_6 + 5178.5277b_7 + 5181.6186b_8 + 576.87b_9$ **= 16344,77**
7 $1838.44b_0 + 36826.58b_1 + 39060.29b_2 + 37654.65b_3 + 37445.71b_4 + 5021.3866b_5 + 5145.0118b_6 + 5233.654b_7 + 5232.601b_8 + 518.47b_9$ **= 16275,83**
8 $1907.77b_0 + 38247.71b_1 + 40468.12b_2 + 38915.25b_3 + 39222.72b_4 + 5178.5277b_5 + 5233.654b_6 + 5539.7919b_7 + 5435.1557b_8 + 601.2b_9$ **= 16867,31**
9 $1917.1b_0 + 38378.86b_1 + 40598.22b_2 + 39048.3b_3 + 39209.83b_4 + 5181.6186b_5 + 5232.601_6 + 5435.1557b_7 + 5435.1557b_8 + 581.07b_9$ **= 17166,23**
10 $282b_0 + 5563b_1 + 5280b_2 + 4994b_3 + 5329b_4 + 576.87b_5 + 518.47b_6 + 601.2b_7 + 581.07\ b_8 + 780b_9$ **= 2948**

The next step is to change the equation into the form of a matrix by multiplying the result with the least squares vector column b0; b1; ..... B9.

$$\begin{pmatrix} 700 & 13973 & 14621 & 14064 & 14125 & 1829.8 & 1838.4 & 1907.8 & 1917.1 & 282 \\ 13973 & 279983 & 292243 & 281216 & 282212 & 36654 & 36827 & 38248 & 38379 & 5563 \\ 14621 & 292243 & 309549 & 295904 & 296203 & 39104 & 39060 & 40468 & 40598 & 5280 \\ 14064 & 281216 & 295904 & 287242 & 286075 & 37265 & 37655 & 38915 & 39048 & 4994 \\ 14125 & 282212 & 296203 & 286075 & 290253 & 37212 & 37446 & 39223 & 39210 & 5329 \\ 1830 & 36654 & 39104 & 37265 & 37212 & 5108 & 5021.4 & 5178.5 & 5181.6 & 577 \\ 1838 & 36827 & 39060 & 37655 & 37446 & 5021.4 & 5145 & 5233.7 & 5232.6 & 518 \\ 1908 & 38248 & 40468 & 38915 & 39223 & 5178.5 & 5233.7 & 5539.8 & 5435.2 & 601 \\ 1917 & 38379 & 40598 & 39048 & 39210 & 5181.6 & 5232.6 & 5435.2 & 5578.8 & 581 \\ 282 & 5563 & 5280 & 4994 & 5329 & 576.87 & 518.47 & 601.2 & 581.07 & 780 \end{pmatrix} \times \begin{pmatrix} b0 \\ b1 \\ b2 \\ b3 \\ b4 \\ b5 \\ b6 \\ b7 \\ b8 \\ b9 \end{pmatrix} = \begin{pmatrix} 6469 \\ 129005 \\ 133732 \\ 128900 \\ 129531 \\ 16473 \\ 16572 \\ 17225 \\ 17322 \\ 2948 \end{pmatrix}$$

Figure 3. Regression coefficient matrix multiplication

AB = H

$B = A^{-1} H$

Where :

A = Matrix (is known)

H = Vector columns (is known)

B = Vector columns (not known)

A^{-1} = Reverse (inverse) from matrix A

To obtain inverse matrix A is then performed operation element line between matrix A and the identity matrix, thus the matrix to the left became identity matrix then identity matrix on the right will become inverse matrix.

$$\left(\begin{array}{cccccccccc|cccccccccc} 700 & 13973 & 14621 & 14064 & 14125 & 1829.8 & 1838.4 & 1907.8 & 1917.1 & 282 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 13973 & 279983 & 292243 & 281216 & 282212 & 36654 & 36827 & 38248 & 38379 & 5563 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 14621 & 292243 & 309549 & 295904 & 296203 & 39104 & 39060 & 40468 & 40598 & 5280 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 14064 & 281216 & 295904 & 287242 & 286075 & 37265 & 37655 & 38915 & 39048 & 4994 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 14125 & 282212 & 296203 & 286075 & 290253 & 37212 & 37446 & 39223 & 39210 & 5329 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1830 & 36654 & 39104 & 37265 & 37212 & 5108 & 5021.4 & 5178.5 & 5181.6 & 577 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1838 & 36827 & 39060 & 37655 & 37446 & 5021.4 & 5145 & 5233.7 & 5232.6 & 518 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1908 & 38248 & 40468 & 38915 & 39223 & 5178.5 & 5233.7 & 5539.8 & 5435.2 & 601 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1917 & 38379 & 40598 & 39048 & 39210 & 5181.6 & 5232.6 & 5435.2 & 5578.8 & 581 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 282 & 5563 & 5280 & 4994 & 5329 & 576.87 & 518.47 & 601.2 & 581.07 & 780 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

Figure 4. A matrix multiplication with the identity matrix

The result of inverse matrix multiple with ΣY, ΣX1Y, ΣX2Y, ΣX3Y, ΣX4Y, ΣX5Y, ΣX6Y, ΣX7Y, ΣX8Y, ΣX9Y and resulting value of multiple linear regression equations as below:

0.5673	-0.01896	-0.00758	-0.00064	-0.0042	0.0154602	-0.0053	0.0174156	-0.001023	-0.0063
-0.019	0.001045	2.1E-05	-8.2E-05	5E-05	-0.000186	0.0001035	-0.000495	-2.65E-05	-8E-05
-0.0076	2.08E-05	0.00064	-0.00013	2.3E-05	-0.001501	0.0002066	-0.000194	-7.69E-05	0.00011
-0.0006	-8.2E-05	-0.00013	0.000429	-0.00015	0.0002833	-0.000852	0.0002814	2.86E-05	8.8E-05
-0.0042	4.97E-05	2.3E-05	-0.00015	0.00034	6.504E-05	0.0005442	-0.000805	-0.000209	-4E-06
0.01546	-0.00019	-0.0015	0.000283	6.5E-05	0.009571	-0.003253	-0.000806	-0.000579	-0.0002
-0.0053	0.000104	0.00021	-0.00085	0.00054	-0.003253	0.0105423	-0.003764	-0.001556	0.00098
0.01742	-0.0005	-0.00019	0.000281	-0.00081	-0.000806	-0.003764	0.0085963	-0.001562	-0.0001
-0.001	-2.7E-05	-7.7E-05	2.86E-05	-0.00021	-0.000579	-0.001556	-0.001562	0.00601	0.00052
-0.0063	-7.6E-05	0.00011	8.78E-05	-4.1E-06	-0.000223	0.0009758	-0.000118	0.000517	0.00207

×

6469
129005
133732
128900
129531
16473
16572
17225
17322
2948

=

13.68744
0.101764
-0.06041
0.016933
-0.06987
-0.60519
-0.44221
-0.14531
-0.36919
0.011617

Figure 5. Inverse matrix multiplication with H (vector column)

Equation (Y) multiple linier regressions to determine estimation of graduation student of Information System Department used equation as follow:

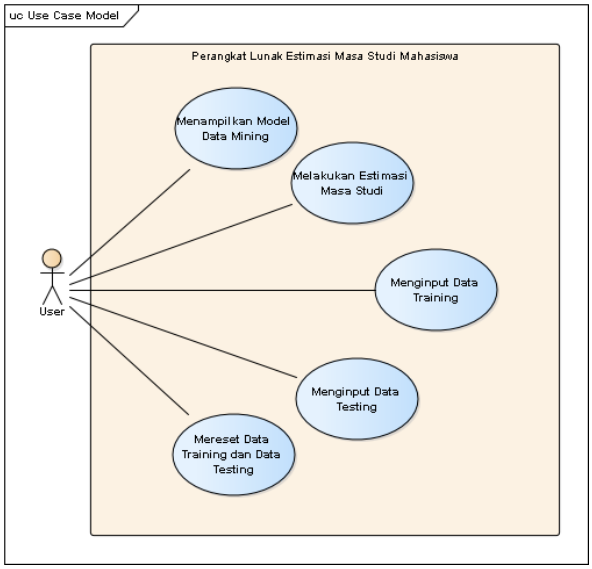
Y = 13.49 + 0.099 X1 + (-0.068) X2 + 0.025 X3 + (-0.059) X4 + (-0.585) X5 + (-0.443) X6 + (-0.155) X7 + (-0.368) X8 + (-0.082) X9

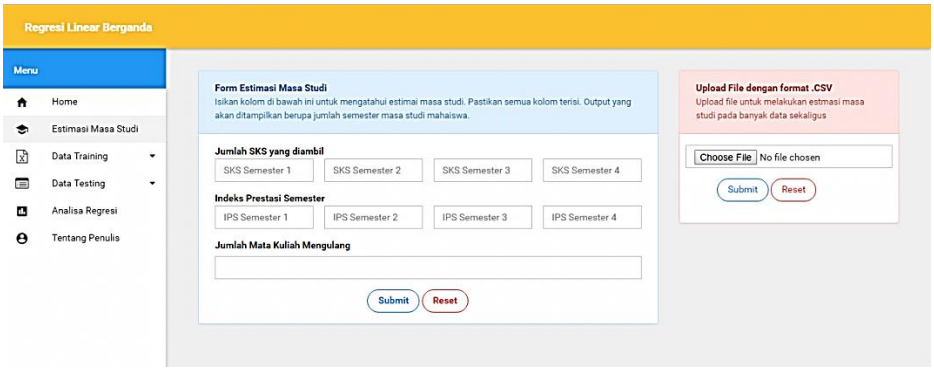
3.2 Evaluation and Validation

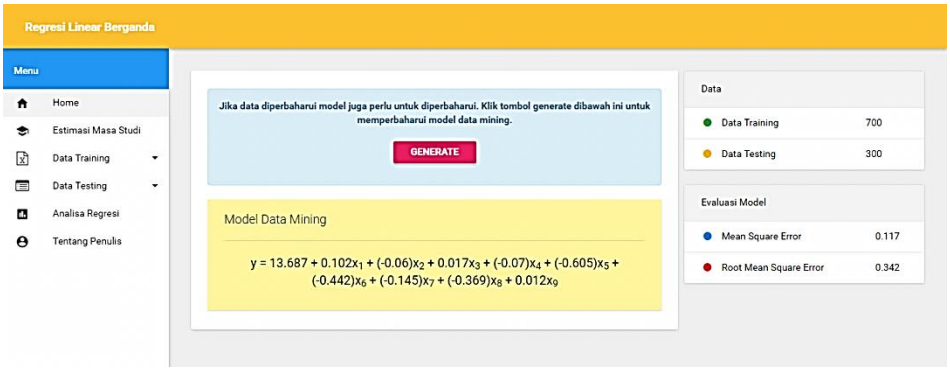
Mean Square Error (MSE) and *Root Mean Square Error* (RMSE) are the methods used in this model test. Calculating using Microsoft Excel by input the 300 testing data. Using the equation above the resulting in MSE and RMSE value respectively are 0.1168 and 0.3418. According to some references stated that the greater the result of RMSE value, the accuracy model is less accurate.

3.3 Prototype Design

The *Use Case* design is an actor's activity diagram in prototype to observe the actor's activity in carrying out the prototype.







4. CONCLUSION

The results of data processing by using data mining algorithms of multiple linear regressions produce a model equation (Y) multiple linear regression to determine the estimation graduation of Student of Information System with 9 independent variables X1 is SKS1; X2 is SKS2; X3 is SKS3; X4 is SKS4; X5 is IPS1; X6 is IPS2; X7 is IPS3; X8 is IPS4; X9 is repeated courses. 1000 data was used as sample data where the data was obtained from 2008 to 2012 generation on Information System Department produce an equation below:

Y = 13.49 + 0.099 X1 + (-0.068) X2 + 0.025 X3 + (-0.059) X4 + (-0.585) X5 + (-0.443) X6 + (-0.155) X7 + (-0.368) X8 + (-0.082) X9

REFERENCES

[1] Vira, “Pemodelan data mining untuk prediksi Kelulusan mahasiswa dengan metode Naive bayes classifier,” Tugas Akhir Sistem Informasi Universitas Dian Nuswantoro. Semarang. 2015

[2] K. Hafidh, “Memprediksi Masa Studi Mahasiswa Menggunakan Metode Jaccard Coefficient (Studi Kasus: Mahasiswa Program Studi Teknik Informatika Jurusan Teknik Elektro Fakultas Teknik Universitas Tanjungpura),”Kalimantan Tengah. 2015

[3] Windarti M, “Prediksi Masa Studi Mahasiswa Menggunakan Kombinasi Algoritma Bayesian Network Dan K- Nearest Neighbors,” Tesis Universitas Adma Jaya Yogyakarta, Program Studi Magister Teknik Informatika. Yogyakarta. 2016

[4] D.H Kamagi dan S. Hansun, “Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan mahasiswa,” Jurnal ULTIMATICS, Vol. VI, No. 1. 2014

[5] Pramitarini Y. Dkk. Analisa Rekam Medis Untuk Menentukan Status Gizi Anak Balita Menggunakan Naive Bayes Classifier. Prosiding Seminar Nasional Manajemen Teknologi XVII. Program Studi MMT-ITS, Surabaya 2 Februari 2013

[6] Ali Fikri, “Penerapan Data Mining Untuk Mengetahui Tingkat Kekuatan Beton Yang Dihasilkan Dengan Metode Estimasi Menggunakan Linear Regression,” *eprints.dinus.ac.id/12789/1/jurnal_12969.pdf. Semarang. 2013*

[7] A. A. Ghofur dan U. D. Widianti, “Sistem Peramalan Untuk Pengadaan Material Unit Injection di PT. XYZ,” Jurnal Ilmiah Komputer dan Informatika (KOMPUTA), vol. III, no. 2, pp. 13-18, 2013.

[8] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning (p. 68).* New York: Springer.

[9] Shalahuddin, M. dan Rosa A.S. Rekayasa Perangkat Lunak Terstruktur dan Berorientasi Objek. Bandung : Informatika. 2013

[10] Anisya, “Aplikasi Sistem Database Rumah Sakit Terpusat Pada Rumah SAkit Umum (RSU) 'Aisyiyah Padang Dengan Menerapkan Open Source (PHP - MySQL),” Jurnal Momentum, vol. 15, no. 2, pp. 1-10, 2013.