Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 21, No. 2, Maret 2022, pp. 249∼256
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v21i2.1348

Forecasting The Number Of Students In Multiple Linear Regressions

Fristi Riandari[1], Hengki Tamando Sihotang[2], Husain[3]
[1,2]STMIK Pelita Nusantara, Indonesia
[3]Universitas Bumigora, Indonesia

Article Info

Article history:

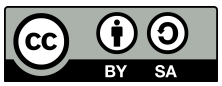
Received July 18, 2021
Revised December 15, 2021
Accepted Januari 15, 2022

Keywords:

Big data
Data Mining
Multiple linear regressions
Forecasting

ABSTRACT

The most important element of higher education was students, therefore every university must continue to improve services in the future, and one of them was by using decision support. This case could be done by utilizing the University of Big Data. Predicting the number of prospective students in higher education was done by utilizing data mining and multiple linear regression approaches. By using 2 independent variables, namely administration costs (X1), accreditation score (X2), and the number of students who was registered each year as dependent variable (Y). For the test data, it used database for the last 13 years. By using multiple linear regression, the intercept value was sought and the coefficient of determination until the regression coefficient was obtained with the equation $Y = 45.28 + -0.02.X1 + 121.58.X2$, noted that if X2 was constant, the increasing of one unit was in X1 would have the effect of increasing -0.02 units on Y. Secondly, if X1 was constant, the increasing of one unit was in X2, would have the effect of increasing 121.58 units in Y. Thirdly, if X1 and X2 were equal to zero, the magnitude of Y was 45.28 units. Therefore, the proposed approach could be provided the acceptable predictive results.



Corresponding Author:

Fristi Riandari,
Department of Computer Engineering,
STMIK Pelita Nusantara,
Email: fristy.rianda@ymail.com

1. INTRODUCTION

The popularity of big data in the last few decades has shown that big data can ne used to produce some information in the future. It can be seen from the number of topics relates to big data. In the era of digital transformation, big data has important role, as it can give valuable information in prediction [1]. Based on the technological point of view, big data often has 3 VS characteristics: volume (the large volume continuously of multiple devices and applications), velocity (the fast way in generating and needing the data for fast entry), variety (multiple data sources in multiple format) [2, 3].The other point of view add 3 Vs: veracity (the accuracy of meaningful data for analyzing the problem), variability (the changing data meanings constantly), and value (the possibility to extract useful information) [4]. It continues to o evolve until it is defined as having characteristics as 5V (Volume, Variety, Velocity, Value, Veracity) and 1C (Complexity) [5]. Different with traditional data, big data refers to a large growing data set which includes heterogeneous and complex formats [6]. The analysis of Big data describes the activities involve in specification, capture, storage, access, and analysis of large data sets in order to understand the content and to exploit the judgments in decision making [7]. It aims to break into new patterns and business insights in traditional research approaches [9]. One of the techniques which can be used in big data is big data mining technique known as data mining.

Data mining is a computer technology which classifies, categorizes and predicts the amounts of large data [9]. Data mining has been realized by various methods including statistical and empirical analysis, social network analysis, ML techniques, and NLP techniques [10]. One of the advantages of data mining is to produce information which can increase the competitiveness. Data mining can be related to machine learning is extracting the useful information from large data sizes [5]. The importance of data mining is based on several reasons. Firstly is the number of abundant academic resources. Secondly is the accessible reservoir data easily. Thirdly is the amount of large data on scientist collaborations, document sharing, and publications support the scientific impact of various entities including papers, authors and journal.This scientific impact is very important for the government and the business sector for the decision-making process, university ranking decisions, tenure, and recruitment decisions [10]. Quoting four reasons related to the importance of data mining in decision making, therefore the using of data mining is carried out in making predictions. Basically, a prediction is a prediction of the occurrence of an event in the future. Besides that, the purpose of this study is to produce information in formulating policies easily in the future.

The approach used in this study is multiple linear regression approach by using 2 independent variables, namely the Administration Cost (X1) and the Study Program Rating Score (X2) by utilizing the number of students who have registered since the last 13 years as the dependent variable (Y). The results will allow in understanding the accuracy of multiple linear regression approach.

2. RESEARCH METHOD

2.1. Research Flow

The following are the steps carried out in the study or what is referred to as the research flow. The picture below is the research rules that must be passed in this research.
Stage 1: Problem Identification
Observing the problems that often occur in universities, one of which is the number of students who register every year and collect information which is the main problem.
Stage 2: Formulating the Problem
After the problem description is obtained, then an analysis of the main problem will be carried out by applying the forecasting model as a tool in decision making.
Stage 3: Data Collection

1. Literature Review
   Learn the theories and concepts of Data Mining, Forecasting, Multiple Linear Regression and various studies related to this research topic.
2. Expert Interview
   Interviewing parties who have authority in the field of student affairs at universities that are research locations with the aim of obtaining data that will be used as test data.

Stage 4: Data Analysis: Application of Multiple Linear Regression
At this stage the data that has been obtained and has been determined as a fixed variable and independent variable will then be processed using multiple linear regression in accordance with the steps in the approach.

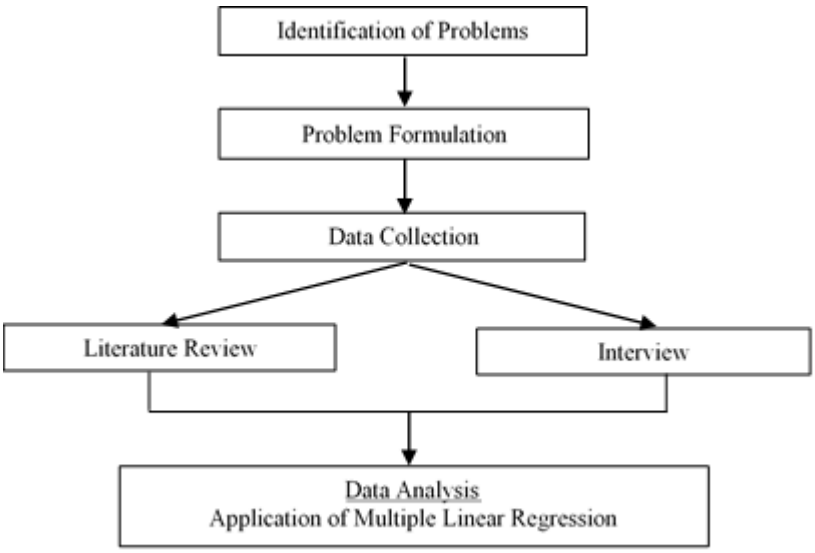


Figure 1. Research Flow

2.2. Multiple Linear Regression Analisis

Statistics has an important pole in big data. It is caused by many statistical methods are used to analyze big data. Statistical software provides much functionality for data and modeling analysis, but just for a small amount of data can be relied on. Regression can be seen in many widely used fields, for instance business, social and behavioral sciences, biological sciences, climate prediction, and so on. Regression is a study of the relationship between the variable and the independent variable, and the relationship between the independent variable and the dependent variable is expressed through regression equations. Multivariate linear regression model represents a variable and a number of independent variables [14]. The analysis of regression is used in statistical analysis of big data as the regression model is popular in data analysis. Linear algorithm has several advantages, one of which is simple structure [15]. Linear regression analysis has two types based on the number of input variables; the first type is called simple linear regression which takes a single input variable, while the other type has more than one input variable and is called multiple linear regressions [16].

Multiple linear regressions can be applied to understand dependent relationship variables and several variables; by using analysis of relationship between the independent variable (x) and the connected variable (y) in developing correlation model which provides the researcher for the dependent variable (y) [17]. Multiple linear regressions is a statistical model used to describe a linear relationship between a variable called ”explain” and a set of independent variables or predictors called ”explanatory” variables [13]. This model was developed to simplify the planning and the development process based on predictive results, and in predicting the using of the multiple linear regressions approach giva more expected results [18] and showed the congruence in predictions [19]. Multiple linear regressions aim to model the relationship between two or more independent and dependent variables with linear equations to the observed data. The theoretical assumption of multiple linear regressions is that each independent variable causes simultaneous changes in the dependent variable [20]. The general model of multiple linear regressions by using 2 independent variables can be seen as follows: Multiple regressions model by using 1 dependent variable (Y) and independent variable (X) is:

$$Y = a + b_1.X_1 + b_2.X_2 + \ldots + b_n.X_n \tag{1}$$

Example for $n = 2$, the regressions model is:

$$Y = a + b_1.X_1 + b_2.X_2$$

Noted:

Y	=	The Y value prediction
X_1	=	The Independent Variable 1
X_2	=	TheIndependent Variable 2
b_1	=	The coefficient of independent variable regression 1, is the change on Y for every change in X1of 1 unit by assumingX2 is constant
b_2	=	The coefficient of independent variable regression 2, is the change on Y for every change in X2 of 1 unit by assuming X1is constant

The analysis of multiple linear regressions can be expressed below:

$$Y = a + b_1.X_1 + b_2.X_2 + \ldots + b_n.X_n$$

For the case of 2 independent variable, the linear equation is expressed as:

$$Y = a + b_1.X_1 + b_2.X_2$$

For getting the value of a, b_1 and b_2 can be used formulas below:

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 \quad (2)$$

$$b_1 = \frac{(\sum X_2^2)(\sum X_1Y) - (\sum X_1X_2)(\sum X_2Y)}{(\sum X_1^2)(\sum X_2^2) - (\sum X_1X_2)^2} \quad (3)$$

$$b_2 = \frac{(\sum X_1^2)(\sum X_2Y) - (\sum X_1X_2)(\sum X_1Y)}{(\sum X_1^2)(\sum X_2^2) - (\sum X_1X_2)^2} \quad (4)$$

In which:

$$\begin{aligned}
\sum X_1^2 &= \sum X_1^2 - \frac{(\sum X_1Y)^2}{n} \\
\sum X_2^2 &= \sum X_2^2 - \frac{(\sum X_2Y)^2}{n} \\
\sum X_1Y &= \sum X_1Y - \frac{(\sum X_1)(\sum Y)}{n} \\
\sum X_2Y &= \sum X_2Y - \frac{(\sum X_2)(\sum Y)}{n} \\
\sum X_1X_2 &= \sum X_1X_2 - \frac{(\sum X_1)(\sum Y)}{n} \\
\sum Y^2 &= \sum Y^2 - \frac{(\sum Y)^2}{n} \\
\bar{Y} &= \sum \frac{Y}{n} \\
\bar{X}_1 &= \frac{\sum X_1}{n} \\
\bar{X}_2 &= \frac{\sum X_2}{n}
\end{aligned}$$

3. RESULT AND ANALYSIS

In this study, the source data was from database of university for the last 13 years, the dataset used was related to the independent variable (Y) and the dependent variable (X) which had been suitable to the needs of the prediction.

Table 1. Multiple Linear Regression Formula Based on 2 (X1, X2) Independent Variables and 1 Dependent Variable (Y)

Year	Number of Registrants Y	BAP X_1	Rating Score X_2	X_1Y	X_2Y	X_1X_2	X_1^2	X_2^2
2008	151	3000	1	453000	151	3000	9000000	1
2009	47	3000	1	141000	47	3000	9000000	1
2010	126	3500	1	441000	126	3500	12250000	1
2011	81	3500	1	283500	81	3500	12250000	1
2012	100	3500	1	350000	100	3500	12250000	1
2013	253	3500	2	885500	506	7000	12250000	4
2014	212	4000	2	848000	424	8000	16000000	4
2015	257	4500	2	1156500	514	9000	20250000	4
2016	215	5250	2	1128750	430	10500	27562500	4
2017	235	5350	2	1257250	470	10700	28622500	4
2018	284	6000	3	1704000	852	18000	36000000	9
2019	337	6500	3	2190500	1011	19500	42250000	9
2020	238	6500	3	1547000	714	19500	42250000	9
	2536	58100	24	12386000	5426	118700	279935000	52

After the value of the multiple linear regression equation is based on 2 independent variables (X1, X2) and 1 Dependent Variable (Y) then the intercept value will be sought to ensure the possibility of other coefficients appearing in the regression model.

Table 2. Intersep

Matrix A			Matrix A1			Matrix A2			Matrix A3		
13	58100	24	2536	58100	24	13	2536	24	13	58100	2536
58100	279935000	118700	12386000	279935000	118700	58100	12386000	118700	58100	279935000	12386000
24	118700	52	5426	118700	52	24	5426	52	24	118700	5426

Furthermore, the value of determination will be determined to determine the percentage contribution of the influence of the independent variables X1 and X2 simultaneously on the dependent variable (Y). The coefficient of determination can be seen in table 4 below:

Table 3. Coefficient of Determination

Det. A, A1, A2, A3	
Det [A]	326370000
Det [A1]	14778860000
Det [A2]	-5452600
Det [A3]	39681050000

The last step to get the equation is to find the regression coefficient with an alternative method, namely the matrix method (least squares method) a, b1 and b2. The regression coefficient can be seen in table 5 below:

Table 4. Coefficient of Regression

B Value	
b1	45.28
b2	-0.02
b3	121.58

Therefore, the multiple linear regression formula was:

$$Y = 45.28 + -0.02.X_1 + 121.58.X_2$$

The meaning of this formula is: First: if X2was constant, the addition of 1 unit on X1would have the effect of increasing -0.02 unit on Y. Secondly, if X1was constant, the addition of 1 unit on X2, would have the effect of increasing 121.58 unit on Y. Thirdly, if X1 and X2 were equal to zero, the Y value was 45.28 units.

By taking this formula if X1: 6750 and X2: 3, then the estimated number of prospective students which have registered in the following year was 297 person, which allowed the university was not increase the administrative costs and if the independent variable values X1: 6500 and X2: 3, it could be predicted that the number of prospective students who registered in the following year would be 301 person.

4. CONCLUSION

According to a series of studies which have been carried out by using a simple linear regression approach in 2 independent variables, the proposed approach has a simple structure and provides acceptable initial prediction results. In addition, it can be used to predict the number of prospective students who will register in the future. Therefore, the university can make some planning better. For further researcher is suggested to make comparisons of prediction methods, for instance Neural Network, Logistic Regression and so on in finding out which method is more appropriate to predict.

REFERENCES

[1] L. Ardito, R. Cerchione, P. Del Vecchio, and E. Raguseo, Big Data in Smart Tourism: Challenges, Issues and Opportunities, Curr. Issues Tour., vol. 22, no. 15, pp. 18051809, 2019.

[2] B. Furht and F. Villanustre, Big Data Technologies and Applications, Big Data Technol. Appl., pp. 1400, 2016.

[3] R. Dautov and S. Distefano, Quantifying Volume, Velocity, and Variety to Support (Big) Data-Intensive Application Development, Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018-January, pp. 28432852, 2017.

[4] I. A. T. Hashem et al., The Role of Big Data in Smart City, Int. J. Inf. Manage., vol. 36, no. 5, pp. 748758, 2016.

[5] T. M. Song and J. Song, Prediction of Risk Factors of Cyberbullying-Related Words in Korea: Application of Data Mining Using Social Big Data, Telemat. Informatics, vol. 58, p. 101524, 2021.

[6] T. Gajdok, Big Data Analytics in Smart Tourism Destinations. A New Tool for Destination Management Organizations?, pp. 1533, 2019.

[7] A. Gandomi and M. Haider, Beyond The Hype: Big Data Concepts, Methods, and Analytics, Int. J. Inf. Manage., vol. 35, no. 2, pp. 137144, 2015.

[8] D. Wang, X. Robert, and Y. Li, Chinas Smart Tourism Destination Initiative: A Taste of The Service-Dominant Logic, J. Destin. Mark. Manag., vol. 2, no. 2, pp. 5961, 2013.

[9] A. Yang, Y. Han, C.-S. Liu, J.-H. Wu, and D.-B. Hua, D-TSVR Recurrence Prediction Driven by Medical Big Data in Cancer, IEEE Trans. Ind. Informatics, vol. 3203, no. c, pp. 11, 2020.

[10] A. Dridi, M. M. Gaber, R. M. A. Azad, and J. Bhogal, Scholarly Data Mining: A Systematic Review of Its Applications, Wiley Interdiscip. Rev. Data Min. Knowl. Discov., no. October, pp. 123, 2020.

[11] Y. Ge and H. Wu, Prediction of Corn Price Fluctuation Based on Multiple Linear Regression Analysis Model Under Big Data, Neural Comput. Appl., vol. 32, no. 22, pp. 1684316855, 2020.

[12] J. Hong, Z. Wang, W. Chen, L. Y. Wang, and C. Qu, Online Joint-Prediction of Multi-Forward-Step Battery SOC Using LSTM Neural Networks and Multiple Linear Regression for Real-World Electric Vehicles, J. Energy Storage, vol. 30, no. February, p. 101459, 2020.

[13] K. L. L. Khine and T. T. S. Nyunt, Predictive Big Data Analytics Using Multiple Linear Regression Model, vol. 744. Springer Singapore, 2019.

[14] X. Xu, Z. Sun, L. Wang, J. Fu, and C. Wang, A Comparative Study of Customer Complaint Prediction Model of Time Series, Multiple Linear Regression and BP Neural Network, J. Phys. Conf. Ser., vol. 1187, no. 5, 2019.

[15] F. Wang, Z. Shi, A. Biswas, S. Yang, and J. Ding, Multi-Algorithm Comparison for Predicting Soil Salinity, Geoderma, vol. 365, no. February 2019, p. 114211, 2020.

[16] H. Rawashdeh et al., Intelligent System Based on Data Mining Techniques for Prediction of Preterm Birth for Women with Cervical Cerclage, Comput. Biol. Chem., vol. 85, no. February, p. 107233, 2020.

[17] Y. S. Lee, J. R. Wang, J. W. Zhan, and J. M. Zhang, Data Mining Analysis of Overall Team Information Based on Internet of Things, IEEE Access, vol. 8, pp. 4182241829, 2020.

[18] C. N. Burger, T. L. Grobler, and W. Kleynhans, Discrete Kalman Filter and Linear Regression Comparison for Vessel Coordinate Prediction, Proc. - IEEE Int. Conf. Mob. Data Manag., vol. 2020-June, no. Mdm, pp. 269274, 2020.

[19] Y. S. Kong, S. Abdullah, D. Schramm, M. Z. Omar, and S. M. Haris, Development of Multiple Linear Regression-Based Models for Fatigue Life Evaluation of Automotive Coil Springs, Mech. Syst. Signal Process., vol. 118, pp. 675695, 2019.

[20] Bochumer Institut fr Technologie GmbH, Data Science - Data Science, no. September 2016. 2018.

[21] Liu, C., Jin, R., Gong, E., Liu, Y., Yue, M., Prediction for The Performance of Gas Turbine Units Using Multiple Linear Regression,Proc.- Of the Chinese Society of Electrical Engineering., vol. 37, pp. 4731-4738, Aug 2017.

[22] X. Li, H. Dong, and S. Han, Multiple Linear Regression with Kalman Filter for Predicting End Prices of Online Auctions, 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Aug. 2020.