Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2022. Vol. 1. No. 3

Analysis Of Multiple Regression Data Mining Methods On The Prediction Of Ibtidaiyah School Registration

Fica Oktavia Lusiana[1]*, Muhammad Zarlis[2], Irfan Sudahri Damanik[3], Solikhun[4], Abdi Rahim Damanik[5]

[1,3,5]STIKOM Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia
[4]AMIK Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia
** ficalucyana2@gmail.com*

Abstract

Data mining originates from data explosion problems experienced by agencies / companies that have collected data from various kinds of transactions. Data mining is the process of looking for patterns or interesting information in selected data using certain techniques or methods. Techniques, methods, or algorithms in data mining vary widely. The choice of the right method or algorithm is very much dependent on the objectives and the overall Knowledge Discovery in Database (KDD) process. The algorithm used in this research is Multiple Linear Regression. School is a suitable place for the application of this method, therefore this research was conducted at the Madrasah Ibtidaiyah Sinaksak Foundation School. The purpose of this study, among others, was to determine the number of registrants at the Madrasah Ibtidaiyah Foundation School (YMI) Sinaksak. In this study, researchers used multiple linear regression association data mining methods. Sources of research data used are observation and interview methods. It is hoped that from the research the school can make a decision or strategy in the estimation of registrants in the following year.

Keywords: *Multiple Linear Regression, Data Mining, Registrant, Data Explosion*

1. Introduction

Various kinds of methods contained in data mining, for example is the regression method. The regression method works when the data set as the data sample is numeric and has labels[1]. This method is also used to guess an unknown value, for example the case taken is the population growth rate. So that this solution of this problem can be solved using data mining methods with methods including the regression method using the Multiple Linear Regression technique[2],[3],[4]. Data mining comes from data explosion problems experienced by agencies/companies that have collected data from various types of transactions[5],[6]. Examples are purchase data, customer data, transaction data, sales data, and much more. The data will mount up if left alone, like an item of property that is allowed to accumulate more and more, whether the item is simply thrown away or can "mine" it to look for important items or information that are still very useful. In this case, the accumulated information or data can still be parsed so that new data can produce new information that can still be used. If the resulting data can be a new information data then the processing is successful.

With the development of the world of education and the increasing breadth of information among the school environment, everyone is required to be able to compete in determining the education to be chosen. With the increasingly strict world of education, many job training institutions offer their services to prepare themselves for entry into the chosen institution. A school is an institution for students teaching students/students under the supervision of a teacher. Most countries have formal education systems which are generally mandatory. Utilization of technology in Indonesia has been applied in various fields including health, government, agriculture and education. Results in the field of education, the government requires educational institutions to utilize technology, including in the implementation of the selection to enroll in a school. Utilization of information technology is increasingly optimal along with the rapid development of technology-based information systems[7].

The problems that often arise that cause students who are in school to become uncomfortable are the lack of school facilities and infrastructure. The main factor is that the school cannot design the right strategy in dealing with the students' needs in the next year. Planning such as funding for school infrastructure development is not well planned, the addition of inventory for the needs of students is not carried out resulting in planning to be undirected. Schools are also a place for unused data storage because schools are agencies that have quite a lot of transactions and administration make a lot of data useless which can be reprocessed.

Based on the problems, it is hoped that this research can predict the number of registrants for the following years with 2 variables, namely the number of registrants, the number of male and female registrants. The output that will be produced is the prediction of the number of registrants for the following year with the Multiple Linear Regression method as a solution for solving predictions in the future [8],[9],[10] which is expected to be input for the Madrasah Ibtidaiyah Sinaksak Foundation School in making optimal annual planning.

2. Research Methodology

Data collection is the process of procuring primary data, for the needs of a research. The data collection techniques in this study are library Research, namely using the library as a suggestion in collecting data, by studying books as reference material. This is done by reading writings in the form of books and journals related to the case the author has raised. Field Research , namely research conducted

directly in the field using several techniques as Observation is a method of collecting data by conducting direct observations of various existing activities. This the author did by making direct observations at the research site in order to find the data needed in this study which was useful in determining the variables to be tested. This observation involves the school itself and the author's view that what data can be extracted for this research. Research place author has kindergarten, elementary, junior high, and high school so that the data needed is getting bigger and more complex, so the author can dig up big data. Interview is the process of collecting data or information through face to face between the questioner (interviewer) and the party being asked or the answerer (interviewee). The author has made the data that the author wants to test before the interview session. During the interview process, the Foundation Officer and the author talked about the history of the Foundation and the resource person explained what was remembered and a copy of the land deed containing an explanation of the history of the Foundation. And also the author asks for the data needed for research research. The data obtained are in Table 1 as follows as unprocessed data:

Table 1. Number of Registrant Data for 2015-2020

No.	Grade	2015	2016	2017	2018	2019	2020
1.	Playgroup	68	74	89	80	108	112
2.	primary school	381	398	421	439	450	482
3.	Jr. high school	394	407	390	421	436	441
4.	Sr, High School	182	192	203	230	221	248
	Total :	1.025	1.071	1.103	1.170	1.215	1.283

Table 2. Data Accumulation by Gender

Years	Registrant		Number of Registrants (*Y*)
	Male (*X*1)	Female (*X*2)	
2015	568	457	1.025
2016	585	486	1.071
2017	603	500	1.103
2018	611	559	1.170
2019	627	588	1.215
2020	649	634	1.283
Total	3.643	3.224	6.867

Data analysis is the process of inspecting, cleaning, and modeling data with the aim of finding useful information. Easier is the data analysis technique is to process the data into an information that can be used to draw conclusions of a study. This study uses the MAPE (Mean Absolute Percentage Error) method. The MAPE method is a statistical measurement of the accuracy of the forecast (prediction) in the forecasting method. Because this study is a prediction of the number of registrants, MAPE can provide information on how much the forecasting error is compared to the actual value of the series. The smaller the percentage error in MAPE, the more accurate the forecasting result will be.

Table 3. MAPE Range

MAPE Range	Score
< 10 %	The ability of the forecasting model is very good
10 - 20%	Good forecasting model ability
20 - 50 %	The ability of the forecasting model is feasible
> 50 %	Poor forecasting model ability

How to calculate the Mean Absolute Percentage Error (MAPE) is to do a total sum by first subtracting the actual data value with the forecasting data then dividing it by the actual data (absolute value is required) and multiplied by 100 then divided by the number of existing data What is meant by absolute is the value if negative remains positive[11].

The research instrument that author did was observation and interviews. Which is where the author's observations are aimed at knowing and determining the desired variables and data to be tested. Furthermore, after getting the desired data from the observation process, the authors conducted interviews to obtain the data. This process is a way to obtain primary data for research needs.

Figure 1. Research Instruments

In research, it is required to make activity diagrams that aim to make it easier for readers to understand the flow of this research. The author has also made a flowchart of research activities shown in Figure 2 below:

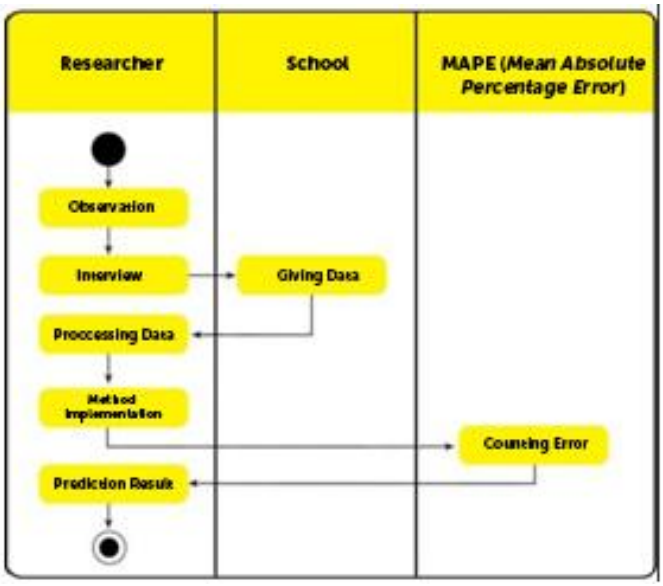

Figure 2. Research Work Activity Diagram

Described in the research work process, what the researchers did were Observation, Interview, Processing Data, Implementing Methods and Generating a prediction from the calculation results. Then the school contributes in providing historical data and MAPE as a means of calculating errors from a data result.

3. Results And Discussion

The development of the Data Mining method is increasingly growing along with the increasing number of Data Mining enthusiasts. Processing of data back which is historical data that has accumulated so that this method is not complicated according to the researchers. The following is also the current research, namely the data used are the data of the registrants stored in the archives of the Sinaksak Ibtidaiyah Madrasah Foundation School which has been established for decades. The stored data can be used to predict the number of registrants. With the application of the Multiple Linear Regression algorithm to calculate predictions that are already popular for calculating predictions, the application of this algorithm is very suitable to be used. The following are several stages of problem solving with multiple linear regression methods:

a. Prepare training data, namely historical data that has happened before and has been grouped into certain classes.
b. Determining the independent variable and dependent variable
   Independent variable : Number of Boys (X1)
   Number of Women (X2)
   Dependent varibale : Number of Registrants (Y)
c. Finding the equation value Y = a + b1.x1 +b2.x2
d. Determine the constant value and regression coefficient

The data that will be taken at the Madrasah Ibtidaiyah Sinaksak Foundation School is data for 2015-2020. The data will be used in the calculation of the multiple linear regression method regarding the prediction of the number of registrants. This data will later form into new information for the coming year.

Table 4. Number of Registrant Data for 2015-2020

No.	Grade	2015	2016	2017	2018	2019	2020
1.	TK	68	74	89	80	108	112

2.	SD	381	398	421	439	450	482
3.	SMP	394	407	390	421	436	441
4.	SMA	182	192	203	230	221	248
	Total :	1.025	1.071	1.103	1.170	1.215	1.283

Table 5. Data Accumulation by Gender

Years	Registrant		Number of Registrant (*Y*)
	Male (*X*1)	Female (*X*2)	
2015	568	457	1.025
2016	585	486	1.071
2017	603	500	1.103
2018	611	559	1.170
2019	627	588	1.215
2020	649	634	1.283
Total	3.643	3.224	6.867

Multiple linear regression is a form of relationship in which both the independent variable X and the dependent variable Y are raised to the power of two[12]. The general equation is:

$$Y = a0 + a1X1 + a2X2 + + anXn$$

Information:

Y = dependent variable (predicted value)

a0 , a1 , a2 , an = regression coefficient

X1 ,X2, , Xn = independent variable

Prediction of the number of registrants as (Y) at the Madrasah Ibtidaiyah Sinaksak Foundation in terms of 2 variables, namely the number of male registrants (X1) and the number of female registrants (X2) which will be predicted using the multiple linear regression method. Then the next step is to find the value of the constant and the regression variable for each independent variable.

Table 6. Calculation Overview

Years	X1^2	X1*X2	X1*Y	X2^2	X2*Y
2015	322.624	259.576	582.200	208.849	468.425
2016	342.225	284.310	626.535	236.196	520.506
2017	363.609	301.500	665.109	250.000	551.500
2018	373.321	341.549	714.870	312.481	654.030
2019	393.129	368.676	761.805	345.744	714.420
2020	421.201	411.466	832.667	401.956	813.422
Total	2.216.109	1.967.077	4.183.186	1.755.226	3.722.303

To get the regression coefficients a, b1 and b2 can be done simultaneously from three equations, namely:

Detailed submission guidelines can be found on the journal web pages. All authors are responsible for understanding these guidelines before submitting their manuscript.

$an+b1\Sigma X1+b2\Sigma X2=\Sigma Y$ ..... (Equation 1)

$a\Sigma X1+b1\Sigma X1^2+b2\Sigma X1X2=\Sigma X1Y$ ..... (Equation 2)

$a\Sigma X2+b1\Sigma X1X2+b2\Sigma X^2=\Sigma X2Y$ ..... (Equation 3)

Then enter the calculation summary number (Table 6) and value ΣX1, ΣX2 and ΣY (Table 5)

a6 + b1(3.643) + b2(3.224)=6.867 (Equation 1)

a(3.643) + b1(2.216.109) + b2(1.967.077)=4.183.186. (Equation 2)

a(3.224) + b1(1.967.077) + b2(1.755.226)=3.722.303. (Equation 3)

The next steps for solving it are as follows:

1. Equations 1 and 2 are eliminated

6a + b1(3.643)+b2(3.224)	=6.867	x3.643
a(3.643) + b1(2.216.109) + b2(1.967.077)	=4.183.186	x6
21.858 a + 13.271.449 b1 +11.745.032 b2	=25.016.481	
21.858 a + 13.296.654 b1 +11.802.462 b2	=25.099.116	
-25.205 b1 +(-57.430)b2	=-82.635 (Equation 4)	

2. Equation 1 and 3 are elimination

6a + b1(3.643)+b2(3.224)	=6.867	x3.224
a(3.224) + b1(1.967.077) + b2(1.755.226)	=3.722.303	x6
19.344 a + 11.745.032 b1 +10.394.176 b2	=22.139.208	
19.344 a + 11.802.462 b1 +10.531.356 b2	=22.333.818	
-57.430 b1 +(-137.180)b2	=-194.610	(Equation 5)

3. Equation 4 and 5 are eliminated

-25.205 b1 +(-57.430)b2	=-82.635	x –57.430
-57.430 b1 +(-137.180)b2	=-194.610	x –25.205
1.447.580.580 b1 +3.298.204.900 b2	=4.745.728.050	
1.447.580.580 b1 +3.457.759.080 b2	=4.905.339.660	

-159.554.180 b2 =-159.611.610
b2 = - 159.611.610 : -159.554.180
b2 = 1,0004

4. Then to find the value of b1 enter the value of b2 in equation 4 by substituting

-25.205 b1 + (-57.430) b2	= -82.635
-25.205 b1 + (-57.430) (1,0004)	= -82.635
-25.205 b1 + (-57.451)	= -82.635

-25.205 b1 = -82.635 – (-57.451)
-25.205 b1 = -25.184
b1 = -25.205 : -25.205
b1 = 0,9991

5. Find the value of a by entering the values of b1 and b2 into equation 1 by substitution

a6 + b1(3.643) + b2(3.224) = 6.867
6a + (0,9991) (3.643) + (1,0004) (3.224) = 6.867
6a + 3.638,82 + 3.225,1604= 6.867
6a + 6.864,9808 = 6.867
6a = 6.867 - 6.864,9808
6a =2,0192
a = 0,3286

Then the values of a, b1 and b2 have been obtained, namely:
a =0,3286
b1 =0,9991
b2 =1,0004

So that it produces the regression equation:

Y = 0,3286 + 0,9991X1 + 1,0004X2

After the regression equation has been obtained, the next step to predict the number of registrants in 2021 is to enter the values of X1 and X2 in the 2020 period, namely with X1 as much as 649 and X2 634. So we reuse the multiple linear regression equation to predict the number of registrants:

Y = a + b1X1 + b2X2
Y = 0,3286 + 0,9991 (649) + 1,0004 (634)
Y = 0,3286 + 648,4335 + 643,2282
Y = 1281 number of registrants

In the next stage, the final results will be displayed using the RapidMiner 5.3 tools. as follows:

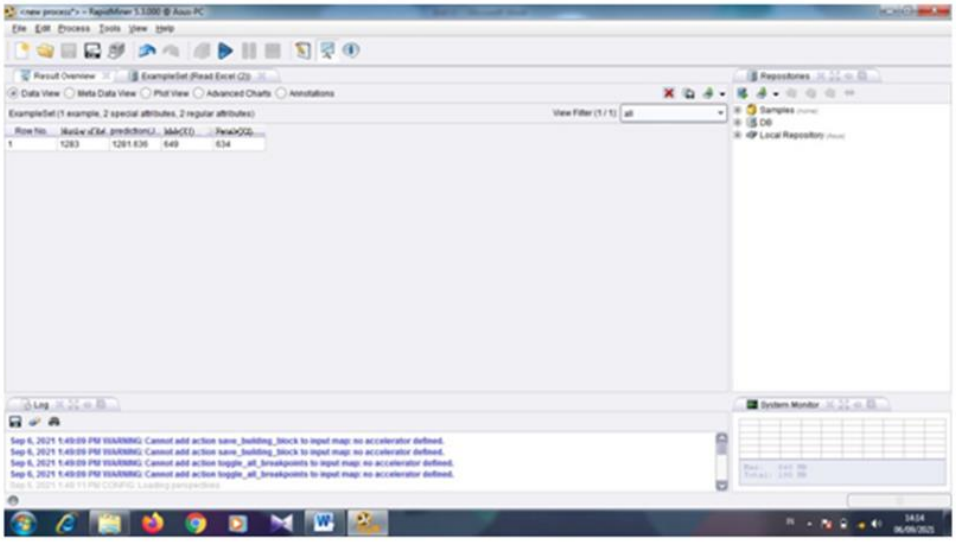

Figure 3. Calculation with RapidMiner's

RapidMiner's calculation results predict that the number of registrants in 2021 will be 1281 registrants.

4. Conclusion

In building the Data Mining method, the algorithm used is the Multiple Linear Regression algorithm in predicting the number of registrants at the Sinaksak Madrasah Ibtidaiyah Foundation School with test calculations using RapidMiner 5.3 and prediction error calculations using MAPE (Mean Absolute Percentage Error). The application of the Multiple Linear Regression algorithm can be applied in calculating the prediction of the number of registrants by testing the MAPE method of 0.43% which is categorized as very good. With this research, it is easier for schools to plan annual infrastructure development carried out at the beginning/end of the year to build schools to be more structured.

Acknowledgement

Acknowledgments to the supervisors and examiners who are lecturers at AMIK and STIKOM Tunas Bangsa so that this research can be arranged as one of the requirements for completing Bachelor's education (S1) at STIKOM Tunas Bangsa. I hope this research can be a reference for other research related to the methods and algorithms used. I hope for constructive suggestions for the readers for the perfection of this research in the future.

References

[1] M. Sarstedt and E. Mooi, "Regression Analysis," 2014, pp. 193–233.

[2] J. Majumdar, S. Naraseeyappa, and S. Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data," *J. Big Data*, vol. 4, no. 1, p. 20, 2017, doi: 10.1186/s40537-017-0077-4.

[3] D. Abdullah *et al.*, "Data Mining to Determine Correlation of Purchasing Cosmetics with A priori Method," in *Journal of Physics: Conference Series*, 2019, vol. 1361, no. 1, doi: 10.1088/1742-6596/1361/1/012056.

[4] F. Maksood and G. Achuthan, "Analysis of Data Mining Techniques and its Applications," *Int. J. Comput. Appl.*, vol. 140, pp. 6–14, Apr. 2016, doi: 10.5120/ijca2016909249.

[5] N. Khan *et al.*, "Big Data: Survey, Technologies, Opportunities, and Challenges," *Sci. World J.*, vol. 2014, p. 712826, 2014, doi: 10.1155/2014/712826.

[6] I. A. Ajah and H. F. Nweke, "Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications," *Big Data and Cognitive Computing* , vol. 3, no. 2. 2019, doi: 10.3390/bdcc3020032.

[7] W. He, Z. J. Zhang, and W. Li, "Information technology solutions, challenges, and suggestions for tackling the COVID-19 pandemic," *Int. J. Inf. Manage.*, vol. 57, p. 102287, Apr. 2021, doi: 10.1016/j.ijinfomgt.2020.102287.

[8] A. M. H. Pardede *et al.*, "Application of Data Mining Prediction of Electricity Deviation Flow Using Metode Backpropogation at PLN Binjai Area," in *Journal of Physics: Conference Series*, 2019, vol. 1363, no. 1, doi: 10.1088/1742-6596/1363/1/012067.

[9] O. Aissaoui, Y. Madani, L. Oughdir, A. Dakkak, and Y. EL ALLIOUI, "A Multiple Linear Regression-Based Approach to Predict Student Performance," 2020, pp. 9–23.

[10] Z. Ismail, A. Yahya, and A. Shabri, "Forecasting Gold Prices Using Multiple Linear Regression Method Department of Mathematics , Faculty of Science Department of Basic Education , Faculty of Education," *Am. J. Appl. Sci.*, vol. 6, no. 8, pp. 1509–1514, 2009.

[11] U. Khair, H. Fahmi, S. Al Hakim, and R. Rahim, "Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error," *J. Phys. Conf. Ser.*, vol. 930, no. 1, 2017, doi: 10.1088/1742-6596/930/1/012002.

[12] E. C. Alexopoulos, "Introduction to multivariate regression analysis," *Hippokratia*, vol. 14, no. Suppl 1, pp. 23–28, Dec. 2010.