International Journal Information System and Computer Science (IJISCS)

Full Paper
pISSN : 2598-0793
eISSN : 2598-246X

THE IMPLEMENTATION OF A SIMPLE LINIER REGRESSIVE ALGORITHM ON DATA FACTORY CASSAVA SINAR LAUT AT THE NORTH OF LAMPUNG

Dwi Marisa Efendi
STMIK Dian Cipta Cendekia Kotabumi
The Nort of Lampung, Indonesian

*Corresponding author
dwimarisa@gmail.com
dwi.marisa@dcc.id

Abstract
Cassava is one type of plant that can be planted in tropical climates. Cassava commodity is one of the leading sub-sectors in the plantation area. Cassava plant is the main ingredient of sago flour which is now experiencing price decline. The condition of the abundant supply of sago or tapioca flour production is due to the increase of cassava planting in each farmer. With the increasing number of cassava planting in farmer's plantation cause the price of cassava received by farmer is not suitable. So for the need of making sago or tapioca flour often excess in buying raw material of cassava This resulted in a lot of rotten cassava and the factory bought cassava for a low price. Based on the problem, this research is done using data mining modeled with multiple linear regression algorithm which aim to estimate the amount of Sago or Tapioca flour that can be produced, so that the future can improve the balance between the amount of cassava supply and tapioca production. The variables used in linear regression analysis are *dependent* variable and *independent* variable . From the data obtained, the *dependent* variable is the number of Tapioca (kg) symbolized by *Y* while the *independent* variable is milled cassava symbolized by *X*. From the results obtained with an accuracy of 95% confidence level, then obtained coefficient of determination (R2) is 1.00. While the estimation results almost closer to the actual data value, with an average error of 0.00.

Keywords: *sugar production, data mining, determination, simple linear, independent, dependent*

1.0 INTRODUCTION

Cassava is one of the important plantation commodities planted because as the main raw material for making sago or tapioca flour. Currently, many farmers of northern Lampung, especially those who have cultivated cassava so that the supply of raw materials is quite large and causing excessive production pr oses. With many raw materials making the price of buying cassava cheaper on the farm side this is very harmful, not only that, with its many raw materials in the factory menghibatkn many raw materials are rotten because the purchase is not dimbangi with melesatnya sales of production. Thus the need to improve effectiveness in processing production to improve the production process better [2].

From the description above can be done ana lisa to cassava production data in North Lampung regency by using *data mining* method . *Data mining* is the process of extracting added value in the form of unknown information manually [3].

One of the *data mining* process that will be used is the estimation method with Linear Regression algorithm. Estimates are estimates of a number of samples. Estimates are additional functions that exist in *data mining*.Linear regression algorithm is a

technique *data mining* to determine that there is a relationship between the variables to be predicted with other variables [4].

Several studies that have been done before, Edy Susanto Tataming in his research stated that regression analysis is a statistical tool that gives explanation about pattern of relationship (model) between two or more variables [5]. Another study by Sarita Permata Dewi stated that multiple linear regression analysis is used to predict the effect of two independent variables or more on one dependent variable [6]. M. Fathrurahman & Haeruddin stated that multiple linear regression analysis is one of the data analysis techniques which is often used to study the relationship between several variables and predict a variable [7]. Based on the research, the research will be made using the same method of multiple linear regression analysis because the method is more commonly used in the case study being studied.

2.0 THEORETICAL

2.1 Estimates

Estimation is to estimate a thing from a number of samples, more inclined to classification but the target estimate variables are more numerical than in the category. System development is done using a complete record that provides the value of the target variable as a prediction. So the value of the target variable will be made according to the predicted variable value.

Linear regression method is structured on the basis of relevant data relation patterns in the past. In general, predicted variables such as inventory, expressed as the variable sought by this variable is influenced by the magnitude of the independent variables. The relationship that occurs between independent variables with the variables sought is a function
Simple Linear Regression Equation:

Y = a + bX + e
Y = Forecasted value
a = Constant
b = Regression coefficient
X = Independent variable
e = Residual Value

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

$$a = \frac{\sum Y - b(\sum X)}{n}$$

Table.1.1 cassava and tapioca flour data

Nomor	singkong/kg(X	tapioka/kg(Y)	y pred	(y-ypred)2	y-yrata)2
1	314.335,0	62.500,0	77477,00191	224310586	1,85E+09
2	349.375,0	65.545,0	85547,40707	400096289	1,59E+09
3	406.450,0	80.000,0	98692,90862	349424833	6,48E+08
4	540.020,0	108.000,0	129456,7219	460390914	6471936
5	500.520,0	100.104,0	120359,0905	410268691	28643904
6	512.460,0	102.492,0	123109,1087	425065171	8785296
7	439.350,0	87.600,0	106270,4294	348584936	3,19E+08
8	518.100,0	103.001,0	124408,1123	458264456	6027025
9	433.140,0	86.344,0	104840,1436	342107328	3,65E+08
10,0	482.490,0	96.211,0	116206,4248	399817014	85470025
11	474.715,0	94.000,0	114415,6885	416800338	1,31E+08
12	504.566,0	100.655,0	121290,9643	425843024	23049601
42	358.145,0	71625	87567,31156	254157298	1,14E+09
43	399.835,0	79678	97169,34326	305947089	6,65E+08
jumlah	18739993	4534606,731	4534606,731	6,285E+11	6,43E+11

a. Coefficient of Determination

The coefficient of determination is used to measure the extent of the model's ability to explain the variation of the *dependent* variable [18].

The value of the determination coefficient (R2) has an interval between 0 and 1 (0 ≤ R2 ≤ 1). If the value of R2 is close to 1, the better for the regression model and if the value of R2 approaches 0 then the *independent*variable can not explain the *dependent* variable as a whole [19].

$$R^2 = 1 - \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2}$$

$$R^2 = 1 - \frac{(628.492.610.572,7)}{(643.343.151.182,7)} = 0,02308339$$

Coefficient of Determination Adjusted (adjusted

$$R_{adj} = R^2 - \frac{P(1 - R^2)}{N - P - 1}$$

$$R_{adj} = 0,02308339 - \frac{1(1 - 0,02308339)}{43 - 1 - 1} = -0,000743845$$

c. Standard Estimate Error

Used to measure the error rate of the regression model formed

$$Se = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n - k}}$$

$$Se = \sqrt{\frac{(628.492.610.572,7)}{43 - 2}} = 123810,6945$$

d. Standard Error Regression Coefficient

Used to measure the magnitude of the error rate of regression coefficients:

$$Sb = \frac{Se}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}}$$

$$Sb_1 = \frac{123810,6945}{\sqrt{(8.447.096.942.733,0) - \frac{(18.739.993,0)^2}{43}}}$$

$$= 4,42261E-07$$

e. Test F (Linearity Test)

F test is used to find out the relation between *independent* variable t with *dependent* variable have linear relation (significant) or not (not significant) [6]. Decision making is done by comparing between Fcount and Ftable with significance level that is alpha (a) equal to 5% (0,05), with confidence level generally 95% [14]. The F test is used to test the accuracy of the model, whether the predictive value is able to describe the actual condition:

Ho: Accepted if $F_{count} \leq F_{table}$

Ha: Accepted if $F_{arithmetic} > F_{table}$

$$F = \frac{R^2/(k-1)}{1-R^2/(n-k)}$$

$$F = \frac{0{,}02308339/(2-1)}{1-0{,}02308339/(43-2)} = 0{,}968781757$$

Because F count (***0.968781757***)> of F table (0.97) then the regression equation is stated **good** *(good of fit)* .

f. Test t

Used to know the influence of independent variables to dependent variables.

H_o : Accepted if $t_{count} \leq t_{table}$

H_a : Accepted if $t_{count} > t_{table}$

$$T_{hitung} = \frac{bj}{Sbj}$$

$$t_{hitung} = \frac{0{,}230319782}{4{,}42261\text{E-}07} = 520778{,}281$$

Because t_{count} (***520778,281***)> from t table (1,943) then H_a received there influence a cassava to sago flour

Residual Plot

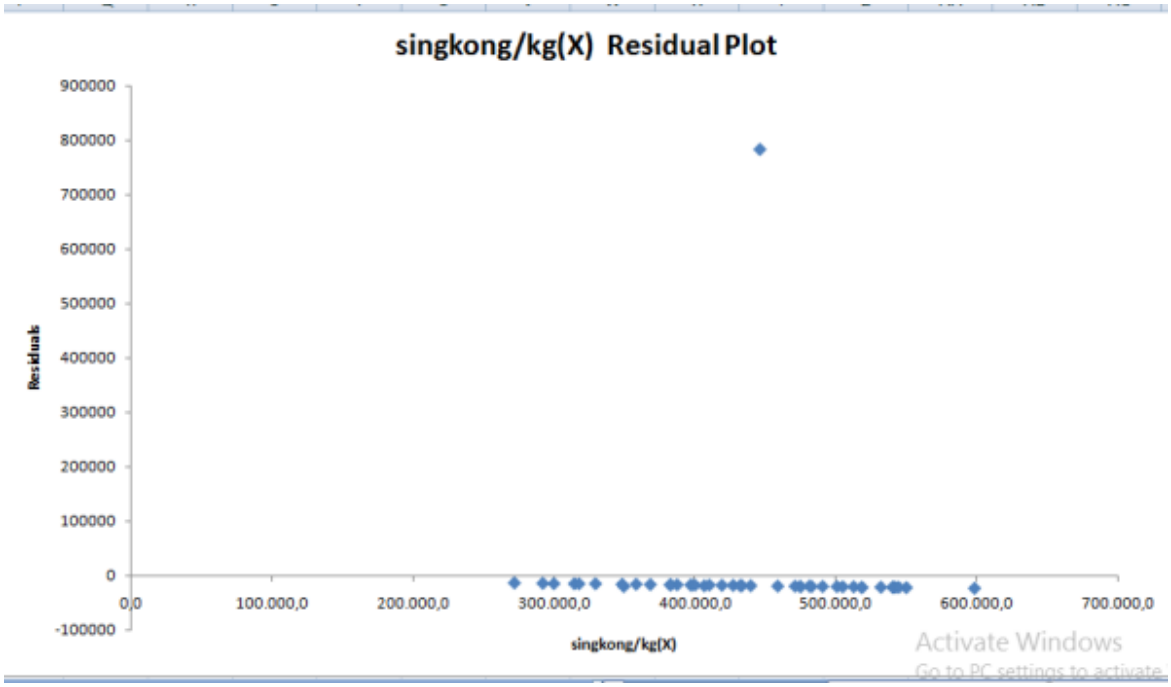

Figure 1.1 Residual Plot

Residual Output

table 1.2 residual output

RESIDUAL OUTPUT

Nom	*Predicted tapioka/kg(Y)*	*Residuals*	*Standard Residuals*
1	77477,00191	-14977,00191	-0,122433268
2	85547,40707	-20002,40707	-0,163514706
3	98692,90862	-18692,90862	-0,152809882
4	129456,7219	-21456,72188	-0,175403368
5	120359,0905	-20255,0905	-0,165580331
6	123109,1087	-20617,10869	-0,168539739
7	106270,4294	-18670,42944	-0,15262612
8	124408,1123	-21407,11226	-0,174997822
9	104840,1436	-18496,1436	-0,151201377
...	.....	.....	.....
....	.....	.....	.....
....	.....	.....	
41	80922,58585	-15066,58585	-0,123165595
42	87567,31156	-15942,31156	-0,130324435

2.3 Model Testing

The model testing stage is the evaluation stage where the model of linear regression equation is predicted how big the error. The method used in model testing is *Root Mean Square Error* (RMSE). Root Mean Square Error (RMSE) is a measure used as a distinction between predicted values and actual values. [4] Where the greater the RMSE value is, the accuracy of a model is less or less accurate, while the smaller the RMSE the better the accuracy of a linear regression model [20].

3.0 RESULANTS

The variables used in this research are cassava and tapioca flour/aci. Tapioca/aci starch is classified as *dependent* variable because influenced by *independent* variable is cassava. From the data that has been obtained known that the data used is 43 data. From the variables that have been defined before then the *dependent* variable is the amount of tapioca flour is assumed as Y, while the *independent* variable is cassava is assumed as X.

Results obtained from the calculation using multiple linear regression analysis is to produce linear model *y = 5079,43 + 0,23x* . And the large linearity relationship generated with a 95% confidence level, then obtained coefficient of determination (R2) is $0{,}02308339$.

4.0 . SUGGESTION

Based on the research that has been done in estimating the production of tapioca flour using *Liniar Regression* algorithm can be used as one of the reference to know sebau pa many tapioca that can be produced and useful to know how big the potency of tapioca commodity in north lampung.

The variables used are *dependent* variable and *independent* variable. The *dependent* variable in the data used is the assumed amount of tapioca with *Y* as well as its *independent* variable is milled cassava assumed as *X*. The result of estimate of sugar production using Gauss Elemination approach with 95% confidence level, then obtained coefficient of determination (R2) is $0{,}02308339$.

REFERENCES

[1] S. Purnama, "Website of Plantation Service of West Java Province," August 28, 2014. [Online] . Available: http://disbun.jabarprov.go.id/index.php/artikel/detailartikel/39. [Accessed December 23, 2014].

[2] ER Anandita, "Sugar Cane Classification Using Naive Bayes Classification Algorithm at the Forest Service and Pati Plantation".

[3] ER Anandita, "Sugar Cane Classification Using Naive Bayes Classification Algorithm at the Department of Forestry and Pati Plantation".

[4] A. Fikri, "Application of *Data Mining* to Know the Level of Concrete Strength Generated by Estimation Method Using Linear Regression".

[5] ES Tataming, "Large Analysis of the Contribution of Side Barrier to Speed by Using Multiple Linear Regression Model (Case Study: Road Segment in Segment of Jalan Serapung Road)," *Civil Statistics,* vol. 2, 2014.

[6] SP Dewi, "Influence of Internal Control and Leadership Style on Employee Performance SPBU Yogyakarta (Case Study at SPBU Anak Cabag Company RB.Group)," *Nominal,* vol. 1, 2012

[7] M. Fathurahman and H., "Linear Regression Modeling for Time Data Data," *Exponential,* vol. 2, 2011

[8] S. Sigilipu, "The Influence of Application of Management Accounting Information and Performance Measurement System to Managerial Performance," *EMBA,* vol. 1, 2013.

[9] H. Susanto and S., " *Data Mining* for Predicting Student Achievement Based on Socio-Economic, Motivation, Discipline and Past Achievements," *Vocational Education,* vol. 4, 2014.

[10] F. Rushdy, "repository.usu.ac.id," 2012. [Online]. Available: http://repository.usu.ac.id/bitstream/123456789/30937/4/Chapter%20II.pdf. [Accessed December 24, 2014].
[11] L. Ernawati and E. Suryani, "National Sugar Productivity Factor Analysis and Its Influence on Domestic Sugar Prices and Demand of Imported Sugar by Using Dynamic Systems," *POMITS Technique,* vol. 1, 2013.
[12] DT Larose, Discovering Knowledge in Data: An Introduction to *Data mining* , Jhon Willey & Sons.Inc, 2005.
[13] D. Kurniawan, "Linear Regression," 2008
[14] J. Sarwono, 12 SPSS Powerful Phase for Thesis Research, Jakarta: Elexmedia Komputindo, 2013
[15] JTSH Nur Setiaji Pamungkas, "Linear Regression Model Influence of Vehicle Composition to Accident Rate on Surabaya-Gempol Toll Road".
[16] YH Ngumar, "Application of Numerical and Matrix Methods in Coefficients of Multiple Linear Regression Coefficients for Forecasting," *National Conference on System and Informatics,* 2008.
[17] R. Istiarini and S., "The Influence of Teacher Certification and Teacher Motivation on Teacher Performance at SMA Negeri 1 Sentolo Kulon Progo Regency 2012," *Jurnal Pendidikan Akutansi Indonesia,* vol. X, pp. 98-113, 2012.
[18] I. Ghozali, Multifariate Analysis with SPSS Program, Semarang: Badan Penerbit UNDIP, 2005.
[19] W. Sulaiman, Regression Analysis using SPSS Case & Solution Example, Yogyakarta: Andi, 2004.
[20] A. Yusuf, H. Ginardi and I. Arieshanti, "Student Predictor Software Development Using Spectral Clustering and Linear Regression Bagging Methods," *ITS Journal of Engineering,* vol. 1, 2012.
[21] Sudjana, Statistics Method, Bandung: Tarsinto, 2005.