DEVOTION
Journal of Research and Community Service
Volume 3, Number 13, November 2022
e-ISSN: 2797-6068 and p-ISSN: 2777-0915

PREDICTION OF CATFISH YIELD TO FULFILL COMMUNITY NEEDS USING MULTIPLE LINEAR REGRESSION ALGORITHM METHOD

Syifa Nurazizah, Sri Winarno
Universitas Dian Nuswantoro, Indonesia
Email: 112201906271@mhs.dinus.ac.id, sri.winarno@dsn.dinus.ac.id

KEYWORDS
Prediction, Data Mining, Multiple Linear Regression Algorithm.

ABSTRACT
Fishery is one sector that is important for human life, because most of human needs come from fisheries, one example is the need for catfish. Currently, there is a gap between crop yields and community needs, which is indicated by very high demand while yields are low. Therefore, this study was conducted to predict catfish yields using a linear regression algorithm so that community needs are met. The method used for the training process in this study uses Multiple Linear Regression. Multiple Linear Regression is an analysis conducted on the dependent variable / dependent variable and two or more independent or independent variables. In contrast to simple regression which only has one independent variable and one dependent variable. Predicting the size of the dependent variable using the independent variable data which is already known. The results of this study analyzed the value of Root Mean Squared Error, and Mean Absolute Error. The result of the line equation from the training data for catfish harvesting is yˆ = -310.6119 + 0.0759x1 + 0.0245x2 + 0.6104x3

INTRODUCTION

Indonesia has source extensive water power and facilities for fisheries and aquaculture. Fishery could Becomes eye livelihood main and source other income for society. Choose algorithm a for predict results is one Thing the most important thing you need to do with right. one algorithm that can used for predict results future catch is multiple linear regression (Puteri & Silvanie, 2020). For determine score from variable Y, must be consider other variables that affect it. So there variable dependent variable Y and variable independent variable X1, X2, Xn. For predict Y when all score variable independent is known, some linear regression models can used. one advantages from algorithm multiple linear regression is could To do Duty prediction with fast, accurate, and provide good result.

Algorithm regression linear many used in study forecasting , including application technique multiple linear regression for predict sale motorcycles (Rahayu, Parlina, & Siregar, 2022) . Some linear regression models can also estimate or predict consumption electricity in the sector commercial the Nigerian economy uses electricity commercial, temperature, bulk rain, electricity, total energy and water variables (Usman, Abdullah, & Mohammed, 2019) . Multiple linear regression models can be used for predict consumption power period short and heat input changes, with burden yesterday and burden week then show results prediction with accuracy up to 95% (Dhaval & Deshpande, 2020) .

Another study that predicts supermarket sales data. On research this used algorithm linear regression and moving average I use two group. One of results sale health and

second that is from results sale electronic For analyze second algorithm the that is with MSE and RMSE. From result analysis the produce MSE value of sales data 50,489, 7,106 and sales data electronic is 57.603, 7.59 (Nafi'iyah & Rakhmawati, 2021) . A study predict production coconut palm use multiple linear regression. In study this variable used is v variable predictive used in forecasting is month, rainfall rain, wide land, amount tree, number bunches, average weight, with production coconut palm as variable dependent. Source of data used is from PT. Nusantara I Plantation and external data in the form of bulk rain from the Meteorology, Climatology and Geophysics Agency. Training data is 180 and the test data is 20 % of the training data. Research results show that obtained equality multiple linear regression Y= -415337.95+1073.82208X1 + 373668741X2 + -15306.629X3 + -621.89932X4 + 11.7449262X5 +7.47948459X6 + 33421 ,5621 X 7 results MAPE as big as 14.28% (Prasetyo & Information Technology and Computers at Lhokseumawe State Polytechnic, 2021) .

In study predict power Sumalata PLTS output Gorontalo. Regency use multiple linear regression. D attache obtained from measurement in two factory for maximizing the resulting solar potential. Equation formed _ from prediction power the output for factory1 is Y = -22216632810.1123 - 771640073.1888X1 + 2349039057.8254X2-25796134709.3552X3 and for factory2 is Y = -2784.107 + 300.0146X1 – 173.7016X2 + 21773.3845X3. The correlation the coefficient on dataset factory1 is 0.52 and dataset factory2 is 0.92, so weather and conditions external other have the effect of 52% on the power output generated in factories1 and 92% in factories2 (Bramasto, Khairiani, Raya, Serpong, & Selatan, 2022) .

Still a lot other research with different problem however use same method that is method multiple linear regression where study this develop method multiple linear regression used as solution for predict sale Isuzu Astra cars. Study this using 2 variables free that is industry (X1) and type (X2) and 1 variable bound that is income (Y). Based on results calculation algorithm multiple linear regression with use tool help SPSS24 obtained F Count value of 48.657 with score 0.000 significance means variables X1 and X2 have an effect to variable Y with R2 = 74.7. span> % MAD value = 0.0607 Research results this give estimation sales of PT. Astra International Tbk - Isuzu 2020 from 12,223 (Al-Fadhilah Nur Wahyudin, Primajaya, & Susilo Yuda Irawan, 2020) .

Study about forecasting bulk rain in the city push using data from BMKG DEO office push with method regression multiple show performance more forecasting good if compared with method other (Yusuf, Setyanto, & Aryasa, 2022) . Mostly study above using linear regression model double, no there is show results catfish production year front use regression multiple linear.

METHOD RESEARCH

Stages study is the steps to be conducted researcher in To do research. Flow scheme stages used in research this shown in Figure 1. Research this use language python programming with the Anaconda Jupyter Notebook tool for To do data processing (Herwanto, Widiyaningtyas, & Indriana, 2019) . As for the stage study writer To do stages study from beginning as following

Defining Scope Problem

At stage define room scope for define formulas and limitations problem that will researched, space scope study this will focus on use multiple linear algorithm on catfish data.

Analyze Problem

At the stage of analyzing the problem, an understanding of the problem that has been determined in its scope or limits is carried out, it is obtained that the analysis of the problem that will be discussed in this study is to predict catfish data. This data can be used to obtain information related to the results the harvest that will come by using multiple linear algorithm.

Find Destination

Study this aim for predict results catfish production in Indonesia using multiple linear regression models, in order to help manager fisheries in the area arrest for increase fish production. Type variable used in study this have two variable, that is variable independent (X) and variable bound (Y). Variable independent cover large land (X1), total cultivator (2), total house ladder fisheries (3) and variable dependent is results production (Y). About problem in study this prediction results harvest estimated based on current data this so that level accuracy still not yet enough.

Learn Literature

Studying Literature At the stage of studying literature, learning is carried out from existing literature to be used as a basis or reference in this research. This study uses several literature studies obtained in the form of materials that are published regularly, for example journals, books and other literature to support the basis and as a reference in this study.

Collecting Data

On research this, the dataset used for model training obtained through the website of the Ministry of Marine Affairs and Fisheries. Following is sample dataset used in the experiment, as shown in Table 1.

Table 1
Sample Dataset

Land Area (ha)	amount cultivator (person)	Amount house ladder fishery (unit)	Yield (tons)
55247	62721	20907	3360.47
55565	6654	2218	3069.2
55647	66468	22156	5920.9
52109	50907	21619	5135.8
60295	54346	2279	6795.72
60295	54346	2279	13140.86
69138	44082	22041	10997.02
103691	20377	15046	12779.5
144713	23485	17726	11634.89
2263	30432	10144	1664.93

Modeling

After the dataset is collected, then stages next is carry out the training process for form a predictive model. Regression is one technique for predict future data come use independent variable and variable explanatory/free). Different with predictive classification score variable that is discrete, regression To do function mapping learning a data element to a variable prediction worth real, used for predict score variable that is continuous.

There is a number of type regression among others:

a. Regression, is a approach for model connection Among variable bound to Y and one or more variable X which is variable free, where whole the variable are quantitative data. It is called linear because every estimation on score expected experience enhancement or drop follow a straight line. Method this used for knowing how dependent variable can predictable through independent variable or variable predictor. Impact from use regression could used for decide does it go up and down? dependent variable can conducted through raise and lower state independent variable , or increase state dependent variable can conducted with increase independent variable and or otherwise . Based on amount variable independent X, linear regression divided Becomes two type that is simple linear regression and multiple linear regression (multiple linear regression).
b. Non-Linear Regression, is connection Among variables Y and X are not linear. Non - linear means, speed Y change due to rate change X not constant for certain X values. Like regression quadratic, cubic. For example: production paddy will increase moment given fertilizer level low to medium. However if given fertilizer with level high, then level production rather the more decreased.
c. Dummy Regression, is connection Among variable y (quantitative data) and variable x (qualitative data). Example: see influence packaging to price sell food. With code '1' represents packaging interesting and '0' if packaging no interesting. Codes '1' and '0' are dummy variable.
d. Regression Logistics, is connection Among variable y (qualitative data) and variable x (quantitative data). For example: If want to is known is consumer will buy food at home eat based on evaluation consumer to location, service, income. In case this only there are 2 possibilities response consumers, namely consumer buy (1) and not buy (0).

A number of algorithm that can used in method regression among others:

a) Simple linear regression
b) Multiple linear regression
c) Polynomial regression
d) Support vector regression
e) Decision tree regression
f) Random forest regression

Method used for the training process in research this use Multiple Linear Regression. Multiple Linear Regression is an analysis conducted on the dependent variable/dependent variable and two or more independent or independent variables. In contrast to simple regression which only has one independent variable and one dependent variable n. Predicting the size of the dependent variable using data on the independent variable whose magnitude is known (Hahs-Vaughn, 2021) .

Multiple linear regression model can depicted with equality as following:

Y = + 1 X1 + 2 X2 + n Xn + e

Description:

Y = Variable bound (Dependent)
X = Variable free (independent)
= Constant (Intercept)
= Slope or Coefficient estimate.
e = Error

Steps in MLR method of study this is as following:

1. Define data as variable free and bound At stage this conducted required data collection. The data will grouped each variable. Required data for variable free that is large land (X1), total cultivator (2), total house ladder fisheries (3) and variable data bound is results harvest (Y).
2. Preprocessing in stages this beginning of the appropriate data processing with specified variable. in this data the raw data in the form of boolean data type will changed into numeric data or integer, so that can processed in stages next. Following stages preprocessing as following: a. Change data type change purpose data type for convert data types so that they can be read by machine learning systems. b. Normalization Normalization aim for make variable with range the same value. c. Train test split Train test split aims for train and test data that has been divided into 2 training data and test data with 80:20 ratio.
3. Modeling Modeling this aim for knowing score coefficient and intercept/error on the variable.
4. Prediction in stages this multiple linear regression model will tested with results on the train test split.
5. Evaluation Evaluation this aim for knowing level model accuracy and feasibility system.

Results Analysis

At stage analysis results this is done analysis to RMSE and MAE values for knowing is prediction this worthy used or no.

RESULTS AND DISCUSSION

Based on the stages of data collection that have been carried out, the number of datasets obtained is 300 data. The dataset used for the formation of the prediction model consists of 4 main data, namely: area land, amount cultivator, amount house ladder fishery and results harvest resulting from. Furthermore, the ready dataset will be trained using the multiple linear regression method, where the independent variable or predictor is the area of land, amount cultivator and quantity house ladder fishery. As for the dependent variable or the target is the result harvest. After the training process is complete, the system is ready to make predictions. This data then processed use language Python programming with Jupyter Notebook tool.

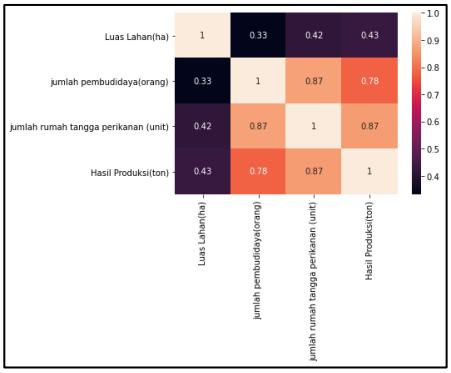

Figure 1
Correlation between variables with heatmap

Figure 2 is a schematic drawing of the relationship between variables. If the number in each box is close to 1, then the relationship between the variables is very strong (Fahlepi & Widjaja, 2019) . Seen from Figure 2, the following information can be obtained:

1. 'Land area' has a value of 0.43 on the 'Production Yield' variable. This indicates that land area has a relationship with production yields.
2. The 'number of cultivators' and 'number of fishery households' have a value of 0.78 and 0.87 for the variable 'Production Results', this indicates that the relationship between the two variables is related, each production output has a large number of cultivators and a large number of fishery households.

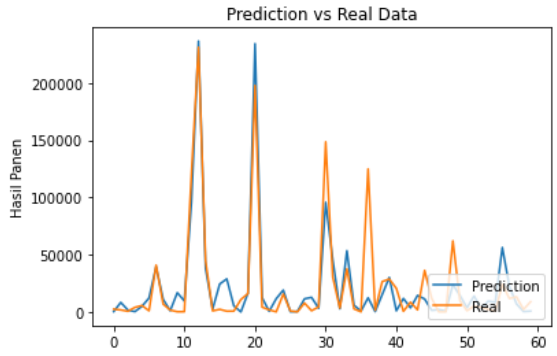

Figure 2
Real and predicted data comparison chart

On the plotting done with compare original data and result data prediction, can seen that results prediction still one track (still direction) and the result no so far with the original data. For that results prediction this can accepted.

Table 2
Comparison of real and predicted data

Harvest	Prediction Results
3360.47	18183.52
3069.2	5425.59
5920.9	19068.13
5135.8	18090.13
6795.72	6991.46
13140.86	6991.46
10997.02	19473.42
12779.5	17246.31
11634.89	22073.25
1664.93	6799.07

R-square (R2) or Coefficient Determination show how much good data match with the regression model. Next for R2 value, value best that is approach number 1. Rsquare value of > 0.67 indicates a very strong model, if score Rsquare > 0.33 means the model is moderate which means the model can accepted and if score Rsquare > 0.19 then the model is said to be weak and not worth. If more small from that so the model should not used. On research this produces an R2 score of 0.75 which indicates that the model is very strong (Wardhana, Aldi, & Siregar, 2022) .

Table 3
Comparison of Testing Data and Training Data

Data	RMSE	MAE
Training	25482.5	12954.7
Testing	20358.4	11192.1

On regression best RMSE and MAE values is the closest value 0. While that, influenced difference form of data owned, MSE, RMSE, and MAE values owned no touching number 0, will but no there is threshold fixed limit for RMSE, Possibility no there is value that can be received for any criteria (Fatah & Subekti, 2018) (Mulyana & Marjuki, 2022) .

This MAE and RMSE value still could upgraded again with add a dataset so that the prediction model that is formed more good again. That thing has prove that the more If many datasets are used, then the prediction model is formed the more good again.

For that our see the gap between value of training data and test data. If the value of RMSE/MAE data train < test data, then could confirmed that overfitting occurs. Then, if difference Among value of train data and test data too far so possibility underfitting occurs.

Studies case this linear regression To do training use tool learning machine or language Python programming. Training results in the form of line equation, training data line equation results harvest catfish is $\hat{y} = -310.6119 + 0.0759x1 + 0.0245x2 + 0.6104x3$.

CONCLUSION

Based on the RMSE results show of 25482.5. our could conclude that when this model used for predict results catfish harvest in Indonesia in range score as trained on the model, the average estimate will be miss about 25,482.5 and an R2 score of 0.75 which means this model is very strong. So could concluded that the resulting MAE and RMSE values still could reduced again with add dataset later day for increase results prediction. Besides it can conducted try later day with compare methods prediction other, for knowing which method is the most accurate in To do prediction.

REFERENCES

Al-Fadhilah Nur Wahyudin, Alif, Primajaya, Aji, & Susilo Yuda Irawan, Agung. (2020). *Application of Double Linear Regression Algorithm On Sales Estimation of Astra Isuzu Car* . *19* (4), 364–374.

Bramasto, Suryo, Khairiani, Dian, Raya, Jl, Serpong, Puspiptek, & South, Tangerang. (2022). *Prediction of Output Power of Solar Power Generation System (PLTS) Using Multiple Linear Regression* . *15* (3), 1979–276.

Dhaval, Bhatti, & Deshpande, Anuradha. (2020). Short-term load forecasting with using multiple linear regression. *International Journal of Electrical and Computer Engineering* , *10* (4), 3911–3917.

Fahlepi, Muhammad, & Widjaja, Andreas. (2019). Application of Multiple Linear Regression Method for Predicting Rental Prices for Boarding Rooms. *Journal of Strategy* , *1* (November), 615–629.

Fatah, Haerul, & Subekti, Agus. (2018). Cryptocurrency Price Prediction Using the K-Nearest Neighbors Method. *Journal of Pilar Nusa Mandiri* , *14* (2), 137.

Hahs-Vaughn, Debbie L. (2021). Multiple Linear Regression. *Applied Multivariate Statistical Concepts* , 71–130.

Herwanto, Heru Wahyu, Widiyaningtyas, Triyanna, & Indriana, Poppy. (2019). Application of Linear Regression Algorithm for Predicting Rice Crop Yield. *National Journal of Electrical Engineering and Information Technology (JNTETI)* , *8* (4), 364.

Mulyana, Dadang Iskandar, & Marjuki. (2022). Optimization of Vaname Shrimp Price Prediction With Rmse And Mae Methods In Linear Regression Algorithm. *Scientific Journal of Betrik* , *13* (1), 50–58.

Nafi'iyah, Nur, & Rakhmawati, Eka. (2021). *Linear Regression and Moving Average Analysis in Predicting Supermarket Sales Data* (Vol. 12).

Prasetyo, Adji, & Information and Computer Technology Lhokseumawe State Polytechnic, Department. (2021). Palm Oil Production Prediction Using Multiple Linear Regression Method. *Multimedia & Networking*, *6* (2).

Putri, Kandari, & Silvanie, Astried. (2020). Machine Learning For Basic Food Price Prediction Model With Multiple Linear Regression Method. *National Journal of Informatics* , *1* (2), 82–94.

Rahayu, Elvri, Parlina, Iin, & Siregar, Zulia Almaida. (2022). *Application of Multiple Linear Regression Algorithm for Motorcycle Sales Estimation Application of Multiple Linear Regression Algorithm for Motorcycle Sales Estimation* . *1* (1).

Usman, OY, Abdullah, MK, & Mohammed, AN (2019). Estimating electricity consumption in the commercial sector of nigeria's economy. *International Journal of Recent Technology and Engineering* , *8* (2 Special Issue), 41–47.

Wardhana, Sanggeni Gali, Aldi, M., & Siregar, Indra Rivaldi. (2022). Prediction of Shear Wave Velocity (Vs) Using Machine Learning in Well X. *Journal of Exploration Geophysics* , *8* (1), 67–77.

Yusuf, Muhammad, Setyanto, Arief, & Aryasa, Komang. (2022). Analysis of Monthly Rainfall Prediction in Sorong City Area Using Multiple Regression Method. *Journal of Computer Science And Informatics* , *6* , 405–417.