

Prediction of linear model on stunting prevalence with machine learning approach

Mambang¹, Finki Dona Marleny², Muhammad Zulfadhilah¹

¹Department of Information Technology, Faculty of Science and Technology, Sari Mulia University, Banjarmasin, Indonesia

²Department of Informatics, Faculty of Engineering, University of Muhammadiyah Banjarmasin, Banjarmasin, Indonesia

Article Info

Article history:

Received May 2, 2022

Revised Sep 5, 2022

Accepted Oct 7, 2022

Keywords:

Linear model

Machine learning

Prevalence

Scikit learn

Stunting

ABSTRACT

An increase in the number of residents should be anticipated including in the health sector, especially the problem of stunting. Stunting in children disrupts height and lack of absorption of nutrients. Information and data drive change in many areas such as health, entertainment, economics, business, and other strategic areas. The stages carried out in this study are initiating, developing linear models, and making prediction results on linear machine learning models. The results of testing with the scikit-learn linear model with a minimum variable of 19 get the best test results, namely the polynomial regression with pipeline model with mean absolute percentage error (MAPE) 0.02, root mean square error (RMSE) 3.32, and coefficient of determination (R²) 1.00. Testing with the scikit-learn linear model with a maximum variable of 48 gets the best test results, namely the polynomial regression with pipeline model with MAPE 0.00, RMSE 3.79 and R² 1.00. Testing with the scikit-learn linear model with an average variable of 32 gets the best test results, namely the polynomial regression model with MAPE 0.01, RMSE 3.32, and R² 1.00. The results of testing with the scikit-learn linear model with the minimum, maximum, and average variables get the best test results, namely the polynomial regression with pipeline model.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Mambang

Department of Information Technology, Faculty of Science and Technology, Sari Mulia University

Jl. Pramuka No. 2, Pemurus Luar, Banjarmasin, South Kalimantan, Indonesia

Email: mambang@unism.ac.id

1. INTRODUCTION

Current and future population growth continues to experience a very significant increase. Economic growth activity will also increase in the coming decades [1]. Climate change and ecosystems are happening all over the world [2]. An increase in the number of residents should be anticipated in many areas, including the health sector, especially the problem of stunting. Stunting that often occurs in children causes impaired height and lack of absorption of nutrients [3]. Stunting is a common problem in many countries globally and occurs in 161 million children between 0 and 5 [4]. Stunting affects up to a third of children in low-to middle-income countries (LMICs) [5].

Stunting problems continue to occur in many developing countries, making it a serious problem experienced by children less than five years old, and this condition occurs in people's lives. Stunting problems in the short and long term in the process of child development, such as the occurrence of decline in cognitive function and other dangerous diseases, stunting problems need to be a priority, especially in the field of health [6]. Malnutrition is a serious problem in the global community resulting in chronic disease and death [7]. Malnutrition affects decreased muscle function, immune disorders, and brain dysfunction and can cause disorders of nerve development [8].

The World Health Organization (WHO) estimates the prevalence of stunted toddlers worldwide at 22% or as many as 149.2 million by 2020. Based on the results of the Indonesian nutrition status survey (SSGI) in 2021, the national stunting rate decreased by 1.6% per year from 27.7% in 2019 to 24.4% in 2021. Most of the 34 provinces showed a decrease compared to 2019 and only 5 provinces showed an increase. The decrease in the prevalence of stunting in toddlers is the main agenda of the government of Indonesia. The secretariat of the vice president coordinates efforts to accelerate stunting prevention to converge, both on planning, and implementation, including monitoring, and evaluation at various levels of government, including villages. The secretariat of the vice president encourages the involvement of all parties in the acceleration of stunting prevention so that its prevalence drops to 14% by 2024. Currently, the prevalence of stunting in Indonesia is better than in Myanmar (35%) but still higher than in Vietnam (23%), Malaysia (17%), Thailand (16%), and Singapore (4%). The decrease in the prevalence of stunting in 2021 can be used to improve the quality of life of the younger generation towards the Indonesian golden generation 2045.

The stunting data used in this study came from secondary data and open access obtained on [9]. The presence of data in all sectors provides a very big change in the order of life [10]. The advent of the big data era has had a huge impact on traditional management methods and companies have also begun to make changes. The huge impact of data growth makes data play a very important role in changing lives, from traditional to more modern and dynamic [11]. Artificial intelligence technology is very important to use to automatically detect and diagnose various types of diseases [12]. Building the model used in this study uses open data trained using machine learning to get the results of predictions and new information [13]. Information and data have driven change in many areas such as health, entertainment, economics and business, and other strategic areas [14].

Data computing using machine learning can provide optimal performance in data processing to be meaningful and have added value for its users [15]. Today and in the future, humans need strategies to manage information and data with various media and networks [16]. Various data management approaches increase innovative value and profitability in business and other strategic areas [17]. Data with various categories can be integrated to explore knowledge and support the latest research on a small scale and a large scale [18]. The prediction process is then carried out with a linear model in the scikit learn machine learning library from the available data. The use of machine learning and creating artificial intelligence models and methods is very important in classifying, clustering and performing other prediction models [19]. The use of machine learning in many cases and datasets in the field of health has contributed a lot of knowledge [20]. Today and in the future, algorithms in processing datasets are becoming increasingly necessary knowledge in many fields [21]. In machine learning, we can perform multistage classifiers such as logistic regression (LR), k-nearest neighbors (kNN), Gaussian Naive Bayes (GNB), support vector machine (SVM), and linear discriminant analysis (LDA) [22]. Group classification problems can be solved using classification algorithms such as SVM [23]. Machine learning models are very important in solving large and complex problems and can reach many disciplines [24]. Machine learning algorithms in classifying disease datasets can be done with various methods [25]. The method used to predict this dataset is the linear model, the linear regression algorithm on machine learning. Previous research has also used machine learning to solve dataset problems in health. The dataset used in this paper is very precise if it is processed with a linear model [26]. The linear output of the model can later provide conclusions to the data process carried out so that the results can be a measure of the strategy that will be carried out in managing stunting problems in the community.

From this condition, we compiled this paper with a linear methodology model on machine learning [27]. Some previous studies on stunting and machine learning approaches in conducting research related to stunting are the basis of theory and scientific studies. Research by Quamme and Iversen [3], they have been reviewing stunting among children for the past 20 years in Sub-Saharan Africa (SSA), where stunting is highly affecting and has high consequences for children's current and future physical and mental development. The method is used by identifying as many as 13 articles from 2000-to 2020 with established criteria. This study stated that children in SSA were severely affected by stunting, with an average prevalence of 41%. According to Harrison *et al.* [5] use machine learning models in determining stunting at birth and systemic inflammatory biomarkers as predictors of the next baby's growth. Machine learning with a random forest model is used in performing descriptive analysis. This study showed that of 107 children who were followed up to 48 months of age, 51% experienced stunting height-for-age Z-score ($HAZ < -2$) at birth which increased to 54% at 48 months of age. Research by Bitew *et al.* [7] use algorithms on machine learning to predict toddlers' malnutrition in Ethiopia. Descriptive results show that the `xgbTree` algorithm has better predictive capabilities than common linear mixed algorithms. Research by Shi *et al.* [8] use machine learning models to predict the occurrence of postoperative malnutrition in children. Their study used the machine learning model used five machine learning algorithms to build predictive models and performance models from machine learning measured with the area under the curve (AUC). Using five predictive algorithms to predict the occurrence of postoperative malnutrition in children, the XGBoost model achieved the largest

AUC in all outcomes. Research with machine learning methods is very important to predict the prevalence of stunting in the future by taking dataset samples from the minimum, average and maximum number of each dataset table so that the results of this study can be a reference in predicting stunting prevalence.

2. METHOD

This research has several stages in predicting the dataset using machine learning. The steps taken are the initiation process, the development of a linear model, and making prediction results that have been processed on a linear machine learning model which can be seen in Figure 1. The initiation stage is carried out by identifying the prevalence population of stunting from the dataset used. The second stage is to develop models and machine learning algorithm selection using dataset samples from the minimum, average and maximum number of all dataset variables, in the end, making predictions from the dataset to get information and value from the data owned.

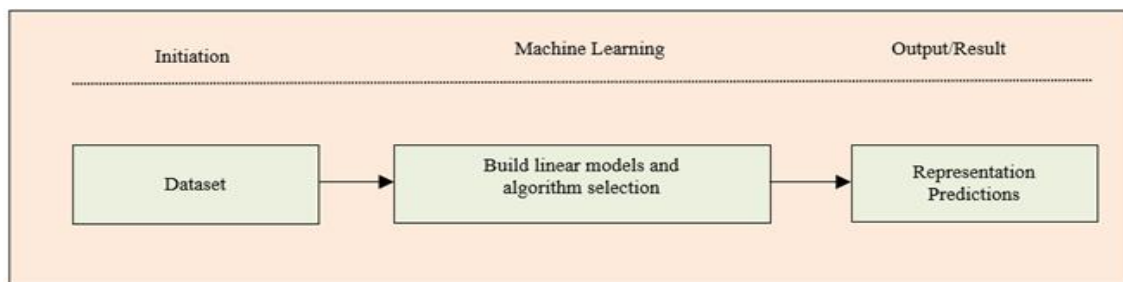


Figure 1. Research method

2.1. Dataset

The data used in this study used datasets from the website database "one data banua", which can be accessed openly to all circles with a URL address in [9]. From this dataset will be carried out the process of sampling data will be carried out the process to the prediction stage with linear models in machine learning. The minimum, maximum and average values in each variable will be summed up and made as an average of each variable's sums. This average value will be used in the prediction process with a linear model. The dataset used in the study has 6 columns and 13 rows.

From Table 1 of the stunting prevalence dataset used as a population, Table 2 is a sample of the dataset that will be processed on a linear machine learning model. The use of samples by creating an average of the values of all tables, such as a minimum every table value every column year, is further made the average of the entire table. The next step is also done simultaneously at maximum and average values. These three variables will be X inputs in the model's linear process.

Table 1. Dataset prevalence stunting

Location/year	2013	2015	2016	2017	2018	2020
1	39.8	38.49	26.97	40.7	29.3	9.4
2	74.8	32.82	36.39	46.7	27.2	15.9
3	48.2	30.09	26.89	26.1	29.1	20.2
4	47.6	39.03	35.91	36.3	28.9	15.2
5	45.3	37.78	35.84	45.7	32.7	13.6
6	49.9	48.08	28.76	39.9	20.7	7.1
7	51.7	37.65	35.94	39.1	29	10
8	56.1	53.54	48	39.4	38.8	21.3
9	44.6	44.64	32.71	36.2	35.3	11.5
10	44.9	31.5	25.41	17.9	26.1	6.4
11	42.7	40.32	15.77	35.3	34.6	26.2
12	35.8	32.13	33.64	31.5	24.9	5.8
13	34.1	21.48	20.99	29	23.5	17.3
Total	615.5	487.55	403.22	463.8	380.1	179.9

Table 2. Sample dataset, minimum, maximum, and average

Variable	Average						
Min	34.1	21.48	15.77	17.9	20.7	5.8	19.3
Max	74.8	53.54	48	46.7	38.8	26.2	48.0
Average	47.3	37.5	31.0	35.7	29.2	13.8	32.4

2.2. Linear model with scikit learn

Scikit-learn or commonly known as sklearn is a python programming language module that is built based on numpy, scipy, and matplotlib. Sklearn modules assist in processing data and performing data training for machine learning and data science purposes. Scikit-learn provides many features, such as classification models, clustering, regression-based machine learning models, and processes used in the feature engineering stage, such as dimensionality reduction using principal component analysis (PCA). Various data scientists use Scikit-learn a lot because of the many machine learning models that can be used with this module. Some of the algorithms that are often used in the liner model with scikit learn are minmaxscaler, linear regression, ridge regression, lasso, elasticnet, orthogonal matching pursuit, generalized linear models (GLM), and stochastic gradient descent.

2.3. Design linear model with Scikit learn

Figure 2 shows the stages of the design method made in this study. Method design is used in this phase to describe conceptually and comprehensively and identify processes and stages in the research process [28]. Research design is a collection of techniques and strategies used to analyze and collect data in order to determine the research variables. Research design is also a strategy carried out by researchers to systematically connect each element of research to analyze and determine the focus of research that becomes more effective and efficient. In creating an input variable X in Table 3 using parameters random.randint (low, high=None, size=None, dtype=int). Programming scripts created as shown in:

```
X = np.random.randint(19, size=(1000,6))
X = pd.DataFrame(X)
X.columns = ['x1', 'x2', 'x3', 'x4', 'x5', 'x6,']
```

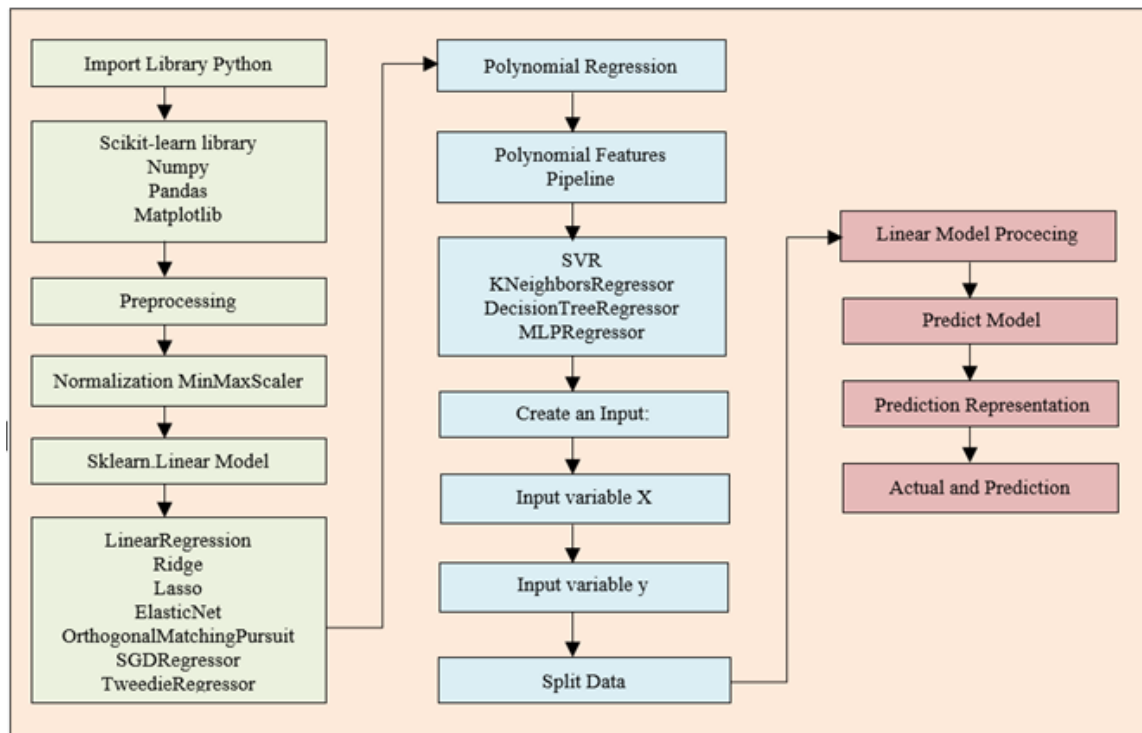


Figure 2. Proposed model linear method design

Table 3. Variable input X

Input variable X	Script input variable X
Minimal 19	X=np.random.randint (19.3, size=(1000,6))
Maximal 48	X=np.random.randint (48.3, size=(1000,6))
Average 32	X=np.random.randint (32.3, size=(1000,6))

Input X=np.random.randint (19) is the minimum value of the dataset sample used, where the random range to be generated from 0 to the largest value is 19. Input value X will have 3 inputs, namely (19,48, and

32). Size (1,000) indicates the number of rows of data to be created. Value (6) indicates the feature we use in variable X. As for the input y that serves as the output later made with the equation $y=5*X[x1]+0.1*(X[x2]**2)+30*(X[x3]**0.5)+X[x4]*X[x5]+X[x6]+10$. The input process of variables X and y generate new data, where this data will be the weight value or (w), and subsequently, the prediction process will be carried out with linear models on sklearn. Table 4 of the results of X and y inputs. Once the X data input and y output are created, the next step is to divide the data used for the training and testing process to create the model and validate the data used. The Table 5 is the training and testing data used.

Table 4. Input results variables X and y

	Variable X						Variable y	
	x1	x2	x3	x4	x5	x6		
0	6	18	5	4	6	1	0	164.482039
1	0	2	3	6	5	12	1	104.361524
2	17	1	9	2	15	5	2	220.100000
3	9	4	14	1	7	7	3	182.849722
4	14	6	4	3	10	0	4	173.600000
...
995	9	3	8	15	1	14	995	169.752814
996	7	9	17	7	15	3	996	284.793169
997	3	2	0	6	12	17	997	114.400000
998	0	9	2	11	17	4	998	251.526407
999	0	10	17	7	9	11	999	217.693169
1,000 rows x 6 columns							Length: 1,000, dtype: float64	

Table 5. X_train, X_valid, y_train, and y_valid

X_Train						X_valid						y_train		y_valid			
	x1	x2	x3	x4	x5	x6	x1	x2	x3	x4	x5	x6					
0	6	18	5	4	6	1	0	7	10	10	2	5	16	0	164.482039	0	175.868330
1	0	2	3	6	5	12	1	1	15	11	8	7	2	1	104.361524	1	194.998744
2	17	1	9	2	15	5	2	12	9	17	4	11	15	2	220.100000	2	260.793169
3	9	4	14	1	7	7	3	0	11	17	16	6	2	3	182.849722	3	243.793169
4	14	6	4	3	10	0	4	10	6	15	9	11	5	4	173.600000	4	283.789500
...
795	18	2	6	4	3	5	795	9	3	8	15	1	14	795	190.884692	195	169.752814
796	3	3	5	2	5	13	796	7	9	17	7	15	3	796	115.982039	196	284.793169
797	5	10	17	2	16	13	797	3	2	0	6	12	17	797	213.693169	197	114.400000
798	6	14	15	2	14	8	798	0	9	2	11	17	4	798	211.789500	198	251.526407
799	1	1	16	2	6	17	799	0	10	17	7	9	11	799	164.100000	199	217.693169
[800 rows x 6 columns]						[200 rows x 6 columns]						Length: 800, float64		Length: 200, float64			

3. RESULTS AND DISCUSSION

After all the previous processes are done, this section describes the results with a linear model on scikit-learn. The linear regression model is based on inputs from "from sklearn.linear_model import linear regression". Form a model or run the training process using the "model.fit (X_train, y_train)". From the command model.fit (X_train, y_train) Will form an input of X_train and output y_train. After the model for input and output is created, then the continuation of the model is used for prediction with commands y_valid_pred=model. Predict (X_valid). Predicted results y_valid_pred It will then be compared to y_valid=y[800:1,000]. To compare the results of these predictions by using mean absolute percentage error (MAPE), root mean square error (RMSE), and coefficient of determination (R2).

3.1. Mean absolute percentage error

The MAPE measures a forecast system's accuracy. It measures this accuracy as a percentage and can be calculated as the average absolute percent error for each time minus actual values divided by actual values. The (1) for mean absolute percentage error is:

$$M = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right| \tag{1}$$

where:

n = is the number of fitted points,

At = is the actual value,

Ft = is the forecast value,

Σ = is summation notation (the absolute value is summed for every forecasted point in time).

The average MAPE is the most common measure used to estimate errors and works best if there are no extremes on the data (and no 0).

3.2. Root mean square error

The RMSE is the residuals' standard deviation (prediction errors). Residuals measure how far data points are off the regression line; RMSE measures how these residuals are dispersed. In other words, it indicates the degree of data concentration around the line of best fit. In climatology, forecasting, and regression analysis, RMSE is frequently used to verify experimental results. The (2) for RMSE is:

$$MRSE = \sqrt{(f - o)^2} \tag{2}$$

where:

f = forecasts (expected values or unknown results),

o = observed values (known results).

3.3. Coefficient of determination

The R2, often known as the R-squared method, is the proportion of the variation in the dependent variable that can be predicted based on the independent variable. It represents the level of dispersion in the provided data set. The (3) for R2 is:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n \Sigma x^2 - (\Sigma x)^2][n \Sigma y^2 - (\Sigma y)^2]}} \tag{3}$$

where:

n = total number of observations

Σx = total of the first variable value

Σy = total of the second variable value

Σxy = sum of the product of first & second value

Σx^2 = sum of the squares of the first value

Σy^2 = sum of the squares of the second value

The Table 6 shows the results of these predictions by using MAPE, RMSE, and R2. The result of Table 6 is the input variable $X = np.random.randint(19, size=(1000,6))$, while to get the output of variable y by using equations $y = 5 * X[x1] + 0.1 * (X[x2]**2) + 30 * (X[x3]**0.5) + X[x4] * X[x5] + X[x6] + 10$. From this test we did, polynomial regression with pipeline is the best method to predict stunting prevalence using minimum variables.

Table 6. Predicted results MAPE, RMSE, R2 with minimum variables

Linear model scikit-learn	MAPE	RMSE	R2
Linear regression	0.13	31.33	0.88
Ridge	0.15	40.19	0.80
Lasso	0.20	50.42	0.69
Orthogona matching pursuit	0.14	32.65	0.87
Tweedie regressor	0.11	26.32	0.92
polynomial regression with pipeline	0.02	3.32	1.00
No pipeline (manual)	0.02	4.61	1.00
Support vector regression (SVR)	0.09	21.96	0.94
KNeighborsRegressor	0.10	23.93	0.93
MLPRegressor	0.03	8.25	0.99

Figure 3 shows the representation of y validation with actual label and y validation prediction with prediction label. From the visualization in Figure 3, the predicted value is close to the actual value. The results of testing with the scikit-learn linear model with a minimum variable of 19 get the best test results, namely the polynomial regression with pipeline model with MAPE 0.02, RMSE 3.32 and R2 1.00. The result of Table 7 is variable input $X = np.random.randint(48, size=(1000,6))$, while to get the output of variable y by using equations $y = 5 * X[x1] + 0.1 * (X[x2]**2) + 30 * (X[x3]**0.5) + X[x4] * X[x5] + X[x6] + 10$. From this test we did, polynomial regression with pipeline is the best method to predict stunting prevalence using maximum variables.

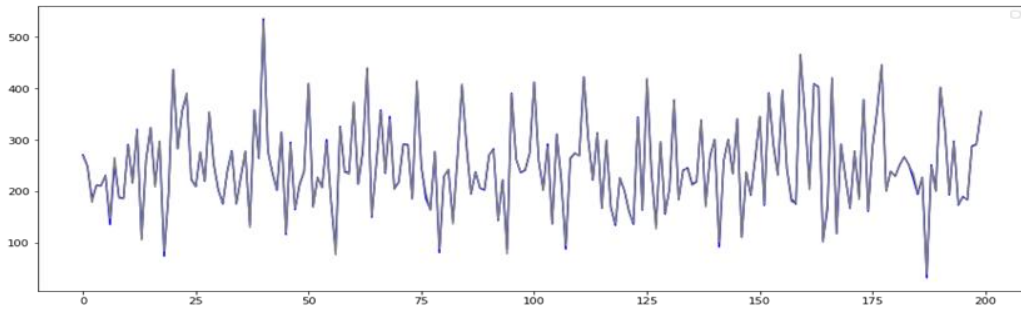


Figure 3. Variable X minimum with polynomial regression

Table 7. Predicted results MAPE, RMSE, and R2 with maximum variables

Linear model scikit-learn	MAPE	RMSE	R2
LinearRegression	0.22	194.92	0.86
Ridge	0.21	199.98	0.86
Lasso	0.22	197.40	0.86
OrthogonalMatchingPursuit	0.23	201.87	0.85
TweedieRegressor	0.15	127.04	0.94
Polynomial Regression with Pipeline	0.00	3.79	1.00
No Pipeline (Manual)	0.01	5.63	1.00
SVR	0.09	86.77	0.97
KNeighborsRegressor	0.11	133.39	0.94
MLPRegressor	0.04	61.61	0.99

Figure 4 shows the representation of y validation with actual label and y validation prediction with prediction label. From the visualization contained in Figure 4, get the result that the predicted value is close to the actual value or actual value. The results of testing with the scikit-learn linear model with a maximum variable of 48 get the best test results, namely the polynomial regression with pipeline model with MAPE 0.00, RMSE 3.79, and R2 1.00. The result of Table 8 is variable input $X=np.random.randint(32, size=(1000,6))$, while to get the output of variable y by using equations $y=5*X[x1]+0.1*(X[x2]**2)+30*(X[x3]**0.5)+X[x4]*X[x5]+X[x6]+10$. From the results of this test that we do, polynomial regression with Pipeline is the best method to predict stunting prevalence using the average variable.

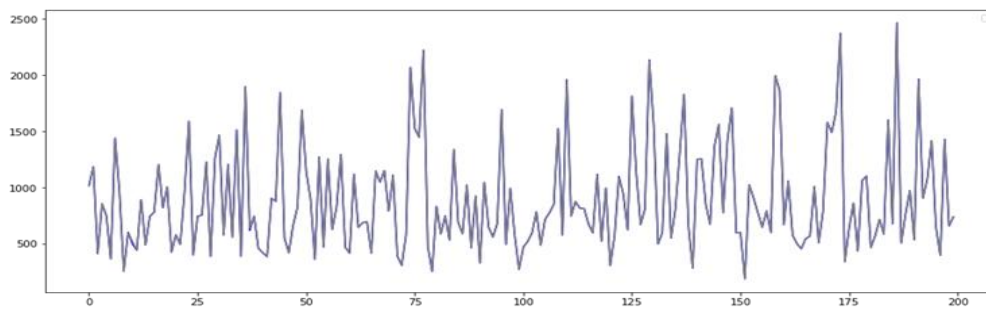


Figure 4. Variable X maximum with polynomial regression

Table 8. Predicted results MAPE, RMSE, and R2 with average variable

Linear Model Scikit-Learn	MAPE	RMSE	R2
LinearRegression	0.17	83.62	0.86
Ridge	0.16	87.75	0.85
Lasso	0.17	87.31	0.85
OrthogonalMatchingPursuit	0.17	85.88	0.85
TweedieRegressor	0.12	59.95	0.93
Polynomial Regression with Pipeline	0.01	3.32	1.00
No Pipeline (Manual)	0.01	5.24	1.00
Support vector regression (SVR)	0.07	36.77	0.97
KNeighborsRegressor	0.10	58.98	0.93
MLPRegressor	0.03	16.61	0.99

Figure 5 shows the representation of y validation with actual label and y validation prediction with prediction label. From the visualization in Figure 5, the predicted value is close to the actual value. The results of testing with the scikit-learn linear model with an average variable of 32 get the best test results, namely the polynomial regression with pipeline model with MAPE 0.01, RMSE 3.32, and R2 1.00. Testing the results of this study is also compared with previous studies such as Elizabeth *et al.* [5]. It uses machine learning models in determining stunting at birth and systemic inflammatory biomarkers as predictors of the next baby's growth. The random forest model identified HAZ at birth as the most important feature in predicting HAZ at 18 months. Of the biomarkers, alpha-1-acid glycoprotein (AGP), c-reactive protein (CRP), and interleukin-1 (IL1) were identified as strong predictors of subsequent growth in both classification and regression models. Research by Bitew *et al.* [7] use algorithms on machine learning to predict the malnutrition of toddlers in Ethiopia. Descriptive results show that the xgbTree algorithm has better predictive capabilities than common linear mixed algorithms. Research by Shi *et al.* [8] use machine learning models to predict the occurrence of postoperative malnutrition in children. Using five predictive algorithms to predict the occurrence of postoperative malnutrition in children, the XGBoost model achieved the largest AUC in all outcomes. Research by Fenta *et al.* [29] testing machine learning to identify determinants of childhood malnutrition. The results obtained by machine learning testing are determined from several algorithms used and show rf algorithms selected as the best ML model.

Polynomial regression is a special case of linear regression. We adjust the polynomial equation on the data to the curved relationship between the target and independent variables. Machine learning pipelines are a way to codify and automate the workflows needed to generate machine learning models. The machine learning pipeline consists of several sequential steps, from data extraction and preprocessing to training and implementing models. Polynomial regression is a statistical analysis method used to look at the influence of two or more variables. Variable relationships are functions that are manifested in a mathematical model. A pipeline is a Python scikit-learn utility for organizing machine learning operations. Pipelines function by allowing a series of linear data transformations to be linked together, resulting in a scalable modelling process. The goal is to ensure that all phases in the pipeline, such as training data sets or any parts involved in cross-validation techniques, are limited to the data available for assessment. The MAPE method provides information on how many forecasting errors compare to the actual value of the series. The smaller the value of the error presentation (percentage error) in MAPE, the more accurate the forecasting results. MAPE values close to zero indicate that the forecasting results follow actual data and can be used for forecasting calculations in future periods. A low RMSE value indicates that the variation in value generated by a forecast model is close to the variation in its observant value. RMSE calculates how different a set of values is. The smaller the RMSE value, the closer the predicted and observed value and the smaller the RMSE value, the better coefficient of determination or commonly symbolized by "R2", which means as a contribution of influence given by free variables or independent variables (X) to bound variables or dependent variables (Y), or in other words, the value of the R2 or R square is useful for predicting and seeing how much influence contribution that variable X gives simultaneously (together) to variable y. The smaller the value of the R2 (R square), then the means that the influence of the free variable (X) on the bound variable (y) is weaker. Conversely, if the value of R square is getting closer to the number 1, then the influence is getting stronger.

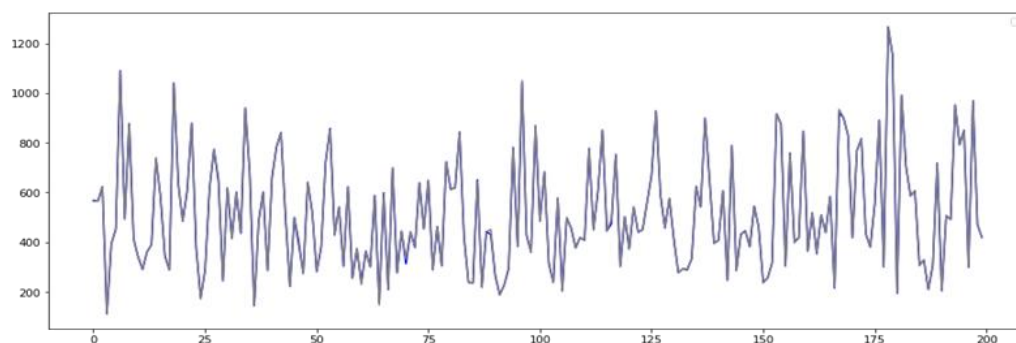


Figure 5. Variable X average with polynomial regression

4. CONCLUSION

An increase in the number of residents should be anticipated in many fields, including the health sector, especially the problem of stunting. Stunting that often occurs in children disrupts height and lack of

absorption of nutrients. Stunting affects up to a third of children in LMICs. The presence of data in all sectors provides a very large change in the order of life. Data with various types of categories can be integrated to explore knowledge and support the latest research both on a small scale and on a large scale. Machine learning algorithms in classifying disease datasets can be done with a wide variety of methods. The method used in predicting this dataset is the linear model, which is used to perform the linear regression algorithm process in machine learning. In this study, there are several stages in predicting datasets using machine learning. The stages carried out are the initiation process, developing linear models, and making prediction results that have been processed on linear machine learning models. The results of testing with the Scikit-learn linear model with minimum, maximum and average variables get the best test results, namely the polynomial regression with pipeline model. From the test of 3 variables that have been done with the polynomial regression with pipeline model, the MAPE value in variable X maximum has a very accurate value of 0.00. The MAPE method provides information on how many forecasting errors compare to the actual value of the series. The smaller the value of the error presentation (percentage error) in MAPE, the more accurate the forecasting results. MAPE values close to zero indicate that the forecasting results follow actual data and can be used for forecasting calculations in future periods. This prediction method can help stakeholders make predictions following the data testing results conducted by machine learning methods using the scikit-learn linear model.

ACKNOWLEDGEMENTS

The author gratefully the support from the Department of Information Technology at Sari Mulia University and the Department of Informatics at the Muhammadiyah University of Banjarmasin.




REFERENCES

- [1] E. Rehman and S. Rehman, "Modeling the nexus between carbon emissions, urbanization, population growth, energy consumption, and economic development in Asia: Evidence from grey relational analysis," *Energy Reports*, vol. 8, pp. 5430–5442, Nov. 2022, doi: 10.1016/j.egy.2022.03.179.
- [2] L. J. M. Milec *et al.*, "Double-edged sword of desalination: Decreased growth and increased grazing endanger range-margin Fucus populations," *Journal of Experimental Marine Biology and Ecology*, vol. 547, pp. 1–11, Feb. 2022, doi: 10.1016/j.jembe.2021.151666.
- [3] S. H. Quamme and P. O. Iversen, "Prevalence of child stunting in Sub-Saharan Africa and its risk factors," *Clinical Nutrition Open Science*, vol. 42, pp. 49–61, Apr. 2022, doi: 10.1016/j.nutos.2022.01.009.
- [4] R. W. Fonseka *et al.*, "Measuring the impacts of maternal child marriage and maternal intimate partner violence and the moderating effects of proximity to conflict on stunting among children under 5 in post-conflict Sri Lanka," *SSM - Population Health*, vol. 18, pp. 1–9, Jun. 2022, doi: 10.1016/j.ssmph.2022.101074.
- [5] E. Harrison *et al.*, "Machine learning model demonstrates stunting at birth and systemic inflammatory biomarkers as predictors of subsequent infant growth – a four-year prospective study," *BMC Pediatrics*, vol. 20, no. 1, pp. 1–10, Dec. 2020, doi: 10.1186/s12887-020-02392-3.
- [6] A. A. Gani, V. Hadju, A. N. Syahrudin, A. S. Otuluwa, S. Palutturi, and A. R. Thaha, "The effect of convergent action on reducing stunting prevalence in under-five children in Banggai District, Central Sulawesi, Indonesia," *Gaceta Sanitaria*, vol. 35, pp. 421–424, 2021, doi: 10.1016/j.gaceta.2021.10.066.
- [7] F. H. Bitew, C. S. Sparks, and S. H. Nyarko, "Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia," *Public Health Nutrition*, vol. 25, no. 2, pp. 269–280, Oct. 2021, doi: 10.1017/S1368980021004262.
- [8] H. Shi *et al.*, "Explainable machine learning model for predicting the occurrence of postoperative malnutrition in children with congenital heart disease," *Clinical Nutrition*, vol. 41, no. 1, pp. 202–210, Jan. 2022, doi: 10.1016/j.clnu.2021.11.006.
- [9] South Kalimantan Provincial Health Office, "Prevalensi stunting," *Dinas Komunikasi dan Informatika Provinsi Kalimantan Selatan*, 2022. <https://data.kalselprov.go.id/dataset/data/1012> (accessed Feb. 28, 2022).
- [10] L. Lv, "RFID data analysis and evaluation based on big data and data clustering," *Computational Intelligence and Neuroscience*, pp. 1–10, Mar. 2022, doi: 10.1155/2022/3432688.
- [11] W. Luo, J. Xu, and Z. Zhou, "Design of data classification and classification management system for big data of hydropower enterprises based on data standards," *Mobile Information Systems*, pp. 1–7, Jan. 2022, doi: 10.1155/2022/8103897.
- [12] S. A. Abdulkareem, H. Y. Radhi, Y. A. Fadil, and H. Mahdi, "Soft computing techniques for early diabetes prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 2, pp. 1167–1176, Feb. 2022, doi: 10.11591/ijeecs.v25.i2.pp1167-1176.
- [13] Y. Li and H. Gan, "Tourism information data processing method based on multi- source data fusion," *Journal of Sensors*, pp. 1–12, Jul. 2021, doi: 10.1155/2021/7047119.
- [14] S. Hannah *et al.*, "Blockchain-based deep learning to process IoT data acquisition in cognitive data," *BioMed Research International*, pp. 1–7, Feb. 2022, doi: 10.1155/2022/5038851.
- [15] H. Wang, Y. Lin, and F. Xiao, "A lightweight data integrity verification with data dynamics for mobile edge computing," *Security and Communication Networks*, pp. 1–15, Mar. 2022, doi: 10.1155/2022/1870779.
- [16] Y. Park, "Research evidence for reshaping global energy strategy based on trend-based approach of big data analytics in the corona era," *Energy Strategy Reviews*, vol. 41, pp. 1–11, May 2022, doi: 10.1016/j.esr.2022.100835.
- [17] X. Cheng *et al.*, "Combating emerging financial risks in the big data era: A perspective review," *Fundamental Research*, vol. 1, no. 5, pp. 595–606, Sep. 2021, doi: 10.1016/j.fmre.2021.08.017.
- [18] L. Balcombe and D. De Leo, "Human-computer interaction in digital mental health," *Informatics*, vol. 9, no. 1, p. 14, Feb. 2022, doi: 10.3390/informatics9010014.
- [19] T. A. Assegie, T. Karpagam, R. L. Mothukuri, R. L. Tulasi, and M. F. Engidaye, "Extraction of human understandable insight from




- machine learning model for diabetes prediction,” *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, pp. 1126–1133, Apr. 2022, doi: 10.11591/eei.v11i2.3391.
- [20] A. H. Ahmed, M. N. A. Al-Hamadani, and I. A. Satam, “Prediction of COVID-19 disease severity using machine learning techniques,” *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, pp. 1069–1074, Apr. 2022, doi: 10.11591/eei.v11i2.3272.
- [21] F. Ribeiro, F. Fidalgo, A. Silva, J. Metrólho, O. Santos, and R. Dionisio, “Literature review of machine-learning algorithms for pressure ulcer prevention: Challenges and opportunities,” *Informatics*, vol. 8, no. 4, p. 76, Nov. 2021, doi: 10.3390/informatics8040076.
- [22] R. Ali, M. M. Yusro, M. S. Hitam, and M. I. Abdullah, “Machine learning with multistage classifiers for identification of ectoparasite infected mud crab genus *Scylla*,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 2, pp. 406–413, Apr. 2021, doi: 10.12928/telkomnika.v19i2.16724.
- [23] J. E. Aurelia, Z. Rustam, I. Wirasati, S. Hartini, and G. S. Saragih, “Hepatitis classification using support vector machines and random forest,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 2, pp. 446–451, Jun. 2021, doi: 10.11591/ijai.v10.i2.pp446-451.
- [24] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, “Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis,” *Informatics*, vol. 8, no. 4, p. 79, Nov. 2021, doi: 10.3390/informatics8040079.
- [25] A. W. Olthof *et al.*, “Machine learning based natural language processing of radiology reports in orthopaedic trauma,” *Computer Methods and Programs in Biomedicine*, vol. 208, pp. 1–9, Sep. 2021, doi: 10.1016/j.cmpb.2021.106304.
- [26] P. Tatjewski and M. Ławryńczuk, “Algorithms with state estimation in linear and nonlinear model predictive control,” *Computers & Chemical Engineering*, vol. 143, pp. 1–19, Dec. 2020, doi: 10.1016/j.compchemeng.2020.107065.
- [27] J. Hernández-González, I. Inza, I. Granado, O. C. Basurko, J. A. Fernandes, and J. A. Lozano, “Aggregated outputs by linear models: An application on marine litter beaching prediction,” *Information Sciences*, vol. 481, pp. 381–393, May 2019, doi: 10.1016/j.ins.2018.12.083.
- [28] L.-H. Shih, Y.-T. Lee, and F. Huarng, “Creating customer value for product service systems by incorporating internet of things technology,” *Sustainability*, vol. 8, no. 12, pp. 1–16, Nov. 2016, doi: 10.3390/su8121217.
- [29] H. M. Fenta, T. Zewotir, and E. K. Muluneh, “A machine learning classifier approach for identifying the determinants of under-five child undernutrition in Ethiopian administrative zones,” *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–12, Dec. 2021, doi: 10.1186/s12911-021-01652-1.

BIOGRAPHIES OF AUTHORS






Mambang    is a lecturer at Sari Mulia University Banjarmasin Information Technology study program, research interests: big data, deep learning, machine learning, databases, networking, and any new techniques and subjects in computer science. He is also active in publishing opinions in online mass media at the regional and national levels and making article publications in national and international journals. He can be contacted at email: mambang@unism.ac.id.



Finki Dona Marleny    is a lecturer in the Department of Informatics of the Faculty of Engineering, University of Muhammadiyah Banjarmasin, Indonesia, where her research interest is mainly in the fields of big data, data science, and machine learning. Until now, she is still active as a content creator, blogger. She can be contacted at email: finkidona@umbjm.ac.id.



Muhammad Zulfadhilah    is a lecturer in the Department of Information Technology of the Faculty of Sains and Technology, Sari Mulia University, Banjarmasin, South Kalimantan, Indonesia. His research interests are in the fields of data science, expert systems, and data mining. He can be contacted at email: zulfadhilah@unism.ac.id.