Comparison of Machine Learning Algorithms for Predicting Stunting Prevalence in Indonesia

Moh. Asry Eka Pratama[1] , Syaiful Hendra*[2] , Hajra Rasmita Ngemba[3] Rosmala Nur[4] Ryfial Azhar[5] Rahmah Laila[6]
Information Systems Study Program, Faculty of Engineering [1],[3]
Informatics Engineering Study Program, Faculty of Engineering[2], [5], [6]
Department of Public Health, Faculty of Public Health[4]
Tadulako University
Palu, Central Sulawesi, Indonesia
E-mail: asrhiekap@gmail.com[1], syaiful.hendra.garuda@gmail.com[2], hajra.rasmita@gmail.com [3] [4], malanur_id@yahoo.comphyal_dp08@yahoo.com [5], lailarahmah.ella@gmail.com [6]

***Stunting* is a serious public health problem, especially among under-fives, which can cause serious short- and long-term impacts. Efforts to tackle *stunting* in Indonesia involve national strategies and development priorities. Therefore, this study aims to compare the performance of machine learning regression algorithms in predicting stunting prevalence in Indonesia. The data collected is secondary data. The data collection was done carefully, taking explicit details regarding the source, scope, extent, and analysis of the dataset, and using a careful sampling methodology. The model evaluation results show that the *Random Forest Regression* algorithm has the best performance, with a success rate of 90.537%. The application of this model to the new dataset shows that East Nusa Tenggara province has the highest percentage of stunting at 31.85%, while Bali has the lowest percentage at 12.07%. Visualization of the dashboard using *Tableau* provides a clear picture of the distribution of stunting in Indonesia. In conclusion, this research contributes to the development of science, especially in the field of machine learning and public health, and provides policy recommendations for tackling stunting in Indonesia.**

Keywords-*Stunting, Machine Learning, Stunting Prevalence, Regression Algorithm, Tableau*

I. Introduction

A child is said to be stunted if his or her development is hampered by chronic malnutrition until his or her height is below the average age [1],[2]. Stunting is a key indicator when assessing the state of nutrition and public health, especially in children under the age of five [3]. By 2022, the Southeast Asian region will have a stunting prevalence rate of 26.4%, according to statistics compiled by the World Health Organization (WHO) and published in 2023 References [4],[5]. In addition, in 2021, the incidence of stunting in children under five years of age in Indonesia reached 24.4%, as reported by the Status Survey of Indonesia.

Nutrition Indonesia (SSGI) Despite progress, the incidence of stunting in children under the age of five in Indonesia will reach 21.6% by 2022, placing the country in the moderate category according to the World Health Organization's categorization [6], [7].

A child's health and academic performance may be negatively impacted by stunting, both now and in the future. When a child is stunted, their immune system is weakened, they are more likely to develop degenerative diseases, they have difficulty understanding, they learn more slowly, and they do not grow as tall as they should. Stunting has far-reaching consequences, including cognitive impairment, impaired learning capacity, increased susceptibility to chronic diseases, including diabetes, cancer, hypertension, and obesity, and large financial losses [5], [8].

Community initiatives to improve nutrition during the first thousand days of life and the national plan to accelerate stunting prevention are examples of such initiatives. It is one of the national development priorities listed in the National Medium-Term Development Plan's (RPJMN) main objective to reduce stunting in Indonesia [9]. Parental actions to prevent stunting can be done by providing exclusive breastfeeding until the baby is six months old and meeting the nutritional needs of children since pregnancy. This includes reducing the prevalence of stunting in children under five. Presidential Regulation (Perpres) Number 72 of 2021 concerning the Acceleration of Stunting Reduction was issued by the Indonesian government as an effort to reduce stunting rates. The government's goal is to reduce the incidence of stunting by fourteen percent by 2024 [6], [7],[10].

Critical to the success of this effort is that using relevant data and methodologies, we can estimate the incidence of stunting in Indonesia. The government and other stakeholders can help plan and allocate resources for nutrition and health interventions by anticipating stunting rates [11]. Stunting rate prediction can also be used as a monitoring and evaluation tool to see how well the program is performing. By using machine learning, stunting rates can be predicted. The foundation of machine learning-based prediction is historical data, which is transformed into data patterns that can predict future events [12].

Machine learning is a branch of artificial intelligence. Our research focuses on developing AI systems with the ability to learn from past mistakes and successes. One of the many

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI: 10.32736/sisfokom.v13i2.2097, Copyright ©2024
Submitted: February 23, 2024, Revised: April 15, 2024, Accepted: April 21, 2024, Published: June, 15, 2024

applications of machine learning is prediction. Making predictions involves using pre-existing data or knowledge to create unknown values (regression) or categories (classification). Evaluation, planning, and decision-making can all benefit from prediction [12],[13].

Machine learning regression is a guided learning technique that uses input factors related to output variables to estimate numerical or continuous values. The difference between linear and non-linear regression lies in the fact that linear regression makes the assumption of a linear relationship between the predictor variable and the target variable, whereas linear regression is able to capture more complicated non-linear patterns [14].

Assessing how well various machine learning regression models predict the frequency of stunting in Indonesia is the main objective of this study. The machine learning regression methods used for this study are *Support Vector Regression* (SVR), *Decision Trees Regression*, *Random Forest Regression*, and *Generalized Additive Models* (GAM). The selection of these algorithms is based on the non-linear nature of the data patterns of the datasets used. Therefore, these algorithms can effectively handle the complexity and variety of non-linear patterns contained in the dataset, enabling more accurate and reliable predictions. The selection of these algorithms is based on the unique features and methods of modeling and predicting data. This research uses measures that include MSE, which stands for mean squared error; RMSE, which stands for root mean squared error; and total mean absolute error (MAE), R-squared (R2), and mean absolute perception error (MAPE) to evaluate algorithm performance.

There are several previous studies on predicting stunting rates, such as research made by [7]. This study aims to address stunting in East Java Province by analyzing the factors that influence its prevalence based on secondary health survey data. This study applies supervised machine learning methods, including linear regression, support vector regression, and random forest regression. The results showed that the factors highly correlated with stunting prevalence were not only limited to Low Birth Weight (LBW), but also included Human Development Index, sanitation, and the Poor Population Index. In addition, this study confirms that support vector regression has a lower error rate compared to other methods. While research conducted by [12]. This study compares five classification algorithms, namely Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), in identifying stunting in children under five. The dataset used consists of four main attributes: age, gender, weight, and height of the toddlers, as well as binary labels that distinguish between stunted and non-stunted toddlers. The results show that KNN with K = 3 provides the highest accuracy, reaching 94.85%, making it the best model for classifying stunting in children under five. In addition to accuracy, other metrics such as precision, recall, and F1-score were also used to evaluate the algorithm's performance. KNN stood out with the highest F1-score of 89.47% and in research [13]. This study aims to compare five algorithms to predict child stunting: Random Forest, Logistic Regression, Naïve Bayes, SVM, and Neural Networks. Using the 2023 child stunting data from Lubuk Linggau City as many as 400 samples, the research methodology involves the stages of initiation, model development, comparison of test results, and prediction analysis with KNIME. Naïve Bayes showed the highest performance with 98.57% accuracy, F1-Score 0.99, and high recall and precision, making it the best model, although Random Forest also gave good results with 98.29% accuracy. Based on previous research, it can be seen that the best algorithm obtained is based on different variables in the dataset in the study. In this study, the variables used in the dataset are also different from previous studies.

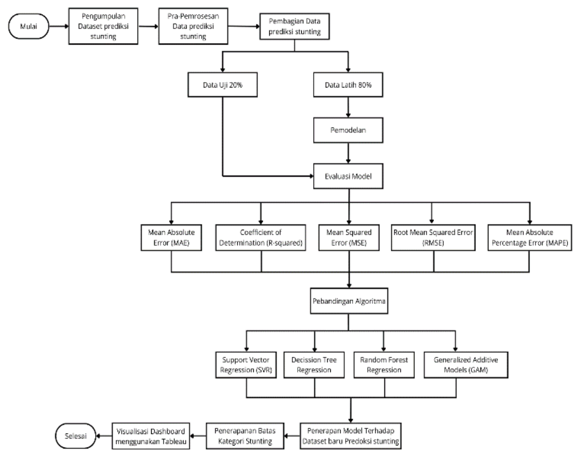
The researchers hope that their work will advance scientific knowledge, particularly in the fields of public health and machine learning. The government and other interested parties in Indonesia can use the findings of this study to inform the development of more effective policies and initiatives aimed at reducing stunting.

II. Methodology

The methodology in this study is characterized by experimental and quantitative approaches. The quantitative research approach utilizes numerical and statistical data collected by various means, such as experiments, surveys, tests, and other similar instruments. The purpose of this approach is to measure variables, test hypotheses, or find relationships between variables [15], [16]. Researchers using experimental research techniques collect data by providing treatment or intervention to research participants. The purpose of this approach is to find out how to influence the measured variables [15],[16]. The main objective of this study is to evaluate the ability of several regression algorithms to predict the frequency of stunting in Indonesia. Here, we can see the use of numerical data generated by regression algorithms and the implementation of treatment through the selection of several methods for the dataset.

A. Research Flow

In this research, there are nine steps: data collection, preprocessing, partitioning, modeling, scoring, algorithm comparison, applying the model to a new dataset, applying category constraints, and dashboard visualization using Tableau.



p-ISSN 2301-7988, e-ISSN 2581-0588
DOI: 10.32736/sisfokom.v13i2.2097, Copyright ©2024
Submitted: February 23, 2024, Revised: April 15, 2024, Accepted: April 21, 2024, Published: June, 15, 2024

Figure 1. Research Flow

B. Data Collection

This study uses secondary data that is legally available on the internet, coming from the Central Bureau of Statistics and the Ministry of Health that can be accessed by the public. This study uses survey data collected from various sources. For the purpose of predicting stunting prevalence, the 2021 and 2022 data are used as the initial dataset, while the 2023 data is used as the new dataset. Here is the data:

1. *Results of the Indonesian Nutrition Status Study (SSGI)[6], [7]*
2. *Health Statistics Profile[17].*
3. *Settlement and Housing[18]*
4. *Consumption and Income[19]*
5. *Early Childhood Profile[20]*
6. *Education Statistics[21], [22],[23].*
7. *Maternal and Child Health Profile[24].*
8. *Indonesia's Poverty Profile[25], [26],[27].*

Some of the data used as attributes in modeling in this study are as follows:

1. *Immunization (Percentage of children who have received all mandatory vaccinations in the last 12-23 months).*
2. *The distribution of health services by health care providers in health care facilities (HFs) is the proportion of the population aged 15 to 49 years who are being treated by health care providers in health facilities.*
3. *Contemporary family planning is the percentage of sexually active women of childbearing age (15-49 years) or their partners who want to have more children or do not want to have more children and use a contemporary contraceptive method.*
4. *Percentage of Stunted Toddlers in Indonesia. (Target Variable/Variable)*
5. *Percentage of exclusively breastfed infants and toddlers (especially those less than twelve months old).*
6. *The percentage of children aged 6-23 months who received breastmilk supplements is referred to as complementary foods.*
7. *Clean water (percentage of households that have access to clean water services).*
8. *Implementation (Percentage of households that have access to interests and sustainability).*
9. *Prevalence of Undernutrition (PoU) in the population that is not chronically malnourished due to excessive food consumption.*
10. *The 10th program is called "Early Childhood Education" (ECED) and lasts for three to six years.*
11. *Jamkesda (also known as Jaminan Kesehatan Nasional (JKN)) ownership is the proportion of the population who received JKN./JAMKESDA and used it for health promotion in the past year).*
12. *Percent Poverty by province, also known as Percent Poor.*

The attributes of interest are included in the EXCEL dataset, which can be processed using Python teaching technology and the Google Cloud Platform.

1. *Health Factors:*
   - Immunization (Percentage of children who have received all mandatory vaccinations in the last 12-23 months).
   - Distribution of health services by health care providers in health care facilities (CHF) is the proportion of the population aged 15 to 49 years who are being treated by health care providers in health facilities.
   - Contemporary family planning (KB) is the percentage of sexually active women of childbearing age (15-49 years) or their partners who want to have more children or do not want to have more children and use contemporary contraceptive methods.
   - Percentage of Stunted Toddlers in Indonesia. (Taeger variable /Dependent variable)
2. *Nutritional factors:*
   - Percentage of infants and toddlers who are exclusively breastfed (especially those less than twelve months old).
   - Percentage of children aged 6-23 months who receive breastmilk supplements referred to as complementary foods (MP).
3. *Settlement and housing factors:*
   - Clean water (Percentage of households with access to clean water services).
   - Livable sanitation (Percentage of households that have access to interest and sustainability).
4. *Food factors:*
   - Prevalence of Undernutrition (PoU) in the population that is not chronically malnourished due to excessive food consumption.
5. *Child education factor:*
   - The 10th program is called "Early Childhood Education" (ECED) and lasts for three to six years.
6. *Social protection factors:*
   - Jamkesda (also known as Jaminan Kesehatan Nasional (JKN)) ownership is the proportion of the population that received JKN./jamkesda and used it for health promotion in the past year).
7. *Economic factors:*
   - Percent Poverty by province, also known as Percent of Poor Population.

C. Data preprocessing

1. *Data understanding.*

The first step at this stage is to use a scatterplot to examine the relationship between each attribute in the dataset, the percentage of stunted children under five, as the y-axis. One way to see the relationship between two variables is through a scatterplot, which is a type of graph. A scatter plot consists of two parts: the mean (x) as the horizontal total, the median (y) as the vertical sum, and the x and y values indicated by tick marks. In this step, we will form a graph that shows the relationship between the two variables. The interaction between variables can show whether there is a positive, negative, or no relationship at all.[28].

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI: 10.32736/sisfokom.v13i2.2097, Copyright ©2024
Submitted: February 23, 2024, Revised: April 15, 2024, Accepted: April 21, 2024, Published: June, 15, 2024

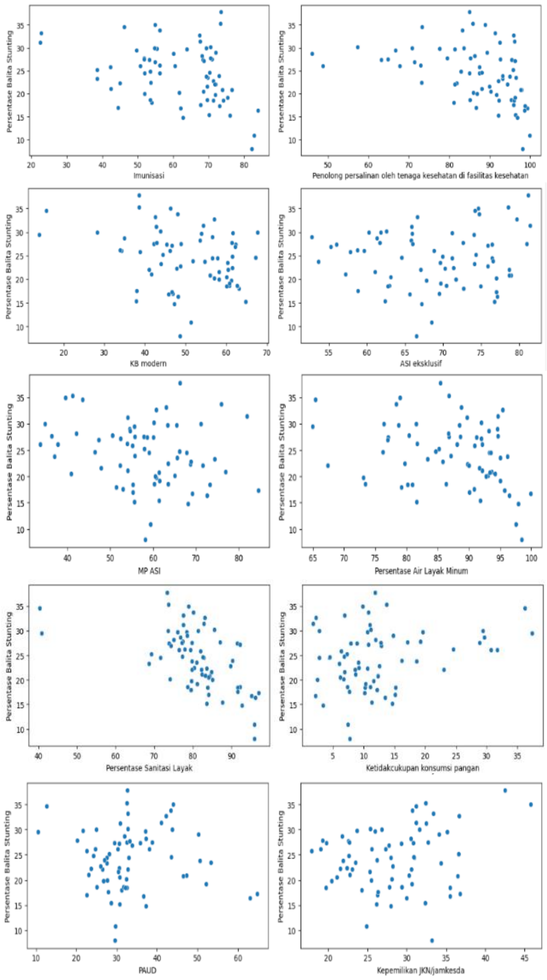


Figure 2. Visualization using *scatterplot*

It can be seen from Figure 2 that the data bars on each attribute have no relationship (non-linear). The algorithm used in the modeling step is a non-linear regression algorithm.

The next step is to test the relationship between variables using Pearson correlation analysis. One of the analytical techniques that can be used to determine the value and direction of the linear relationship between two variables is the Pearson correlation coefficient. The coefficient of determination ranges from -1 to 1, with values close to 1 indicating a greater positive relationship, -1 indicating a stronger negative relationship, and 0 indicating no linear relationship between the two variables. The Pearson correlation analysis technique was used to determine the absence of a relationship between two interval or ratio-sized variables. Normal data distribution and parametric data sets are the basis of Pearson correlation analysis. The results of Pearson correlation analysis can be used to determine the significance of the relationship between two variables and display the value and direction of the trend.

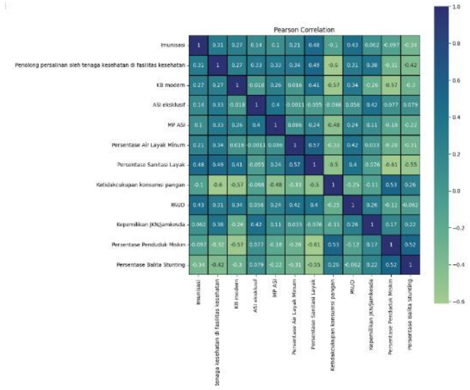


Figure 3. *Heatmap Pearson correlation analysis*

The results of the Pearson correlation analysis, which has the weakest relationship with the percentage of stunting, namely the poor penetration rate of 0.52, can be seen in Figure 3. This means that there is a positive correlation indicating that the stunting rate increases as the proportion of the poor population increases.

2. *Data cleaning*

This data preprocessing step involves checking for missing values and missing data types on each attribute, as well as performing data cleaning or data cleansing operations.

```
print('Data Null ?')
print(df.isnull().sum())

Data Null ?
Provinsi                                                            0
Tahun                                                               0
Imunisasi                                                           0
Penolong persalinan oleh tenaga kesehatan di fasilitas kesehatan    0
KB modern                                                           0
ASI eksklusif                                                       0
MP ASI                                                              0
Persentase Air Layak Minum                                          0
Persentase Sanitasi Layak                                           0
Ketidakcukupan konsumsi pangan                                      0
PAUD                                                                0
Kepemilikan JKN/jamkesda                                            0
Persentase Penduduk Miskin                                          0
Persentase Balita Stunting                                          0
dtype: int64

print(df.dtypes)

Provinsi                                                             object
Tahun                                                                 int64
Imunisasi                                                           float64
Penolong persalinan oleh tenaga kesehatan di fasilitas kesehatan    float64
KB modern                                                           float64
ASI eksklusif                                                       float64
MP ASI                                                              float64
Persentase Air Layak Minum                                          float64
Persentase Sanitasi Layak                                           float64
Ketidakcukupan konsumsi pangan                                      float64
PAUD                                                                float64
Kepemilikan JKN/jamkesda                                            float64
Persentase Penduduk Miskin                                          float64
Persentase Balita Stunting                                          float64
dtype: object
```

Figure 4. Data *Cleaning*

It can be seen from Figure 2 that there are no missing values in the dataset, and all attributes used are numeric. Data type conversion is not required for this result, as one aspect of the regression method is that the attributes to be used in the modeling step are numeric.

3. *Convert data*

This step involved data conversion to assign categories to the appearance attributes of stunted children. The aim was to assess the categories of stunting across all provinces in Indonesia.

```
# batas kategori
bins = [0, 20, 29, 39, float('inf')]
labels = ['Rendah', 'Menengah', 'Tinggi', 'Sangat Tinggi']

# Buat kolom baru dengan kategori
df['Kategori Stunting'] = pd.cut(df['Persentase Balita Stunting'], bins=bins, labels=labels, right=False)
```

Figure 5. *Stunting* category limit

The following categories are based on the stunting standards set by the World Health Organization (WHO). Which is less than 20%, medium 20% - 29%, highest 30% - 39%, and very

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI: 10.32736/sisfokom.v13i2.2097, Copyright ©2024
Submitted: February 23, 2024, Revised: April 15, 2024, Accepted: April 21, 2024, Published: June, 15, 2024

high >= 40% [11].

b. Data Sharing

At this stage, the data is transformed into training data and test data. The data used for modeling is called training data, while the data used to evaluate the learning results of machine learning is called test data. The data is summarized into 80% for training data and 20% for test data.

```
# split data kita menjadi training dan testing, SPLIT 80% dan 20% testing

x_train, x_test, y_train,  y_test=train_test_split(x,y, test_size=0.2, random_state=0)
# Tampilkan ukuran dari training set dan testing set
print("Ukuran Training Set:", len(x_train))
print("Ukuran Testing Set:", len(x_test))

Ukuran Training Set: 54
Ukuran Testing Set: 14
```

Figure 6. Training and test data

b. Modeling

At the stage of understanding the data using the scatterplot technique, it is known that the algorithm to be used in modeling is non-linear regression.

1. *Support Vector Regression* (SVR)

One machine learning regression algorithm that uses the same principles as SVM, or Support Vector Machine, is Support Vector Regression (SVR). Simple Vs. Random (SVM) is a classification algorithm that searches for a hyperplane that divides two data sets by a large margin. SVR also searches for hyperplanes but for continuous value prediction based on the mean variable. Using the concept of epsilon-insensitive loss, SVR is able to overcome prediction errors that are less than the target value of epsilon. In addition, SVR utilizes a kernel, a function that transforms the input data into a higher dimensional space to process non-linear data [29].

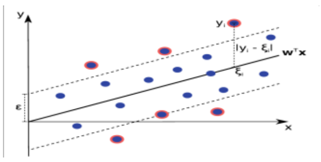


Figure 7. Ilustrasi SVR

```
# Support Vector Regression (SVR)
from sklearn.svm import SVR

svr_model = SVR(kernel='linear')
svr_model.fit(x_train, y_train)
y_pred_svr = svr_model.predict(x_test)
```

Figure. 8. Penerapan model SVR

2. *Decision Trees Regression*

One of the machine learning methods used to create data-driven prediction models is decision tree regression. The decision-making part of this prediction model consists of nodes and edges. Nodes represent input variables, while edges represent the relationship between input and output variables. You can use this prediction model to determine the outcome of a decision by analyzing the curves on the decision graph [30].

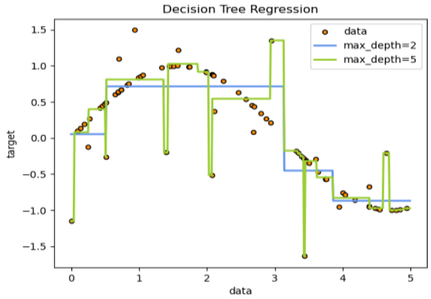


Figure 9 . Ilustrasi *Decision Tree Regression*

```
# Decision Trees Regression
from sklearn.tree import DecisionTreeRegressor

dt_model = DecisionTreeRegressor()
dt_model.fit(x_train, y_train)
y_pred_dt = dt_model.predict(x_test)
```

Figure 10. Penerapan model *Decision tree regression*

3. *Random Forest Regression*

As a machine learning technique, Random Forest Regression is one way to build a prediction model using the provided data. This prediction model consists of a combination of many independent decision trees. You can use this prediction model to estimate the continuous value of a lag variable [31].

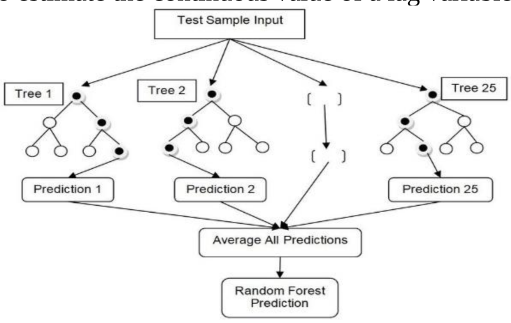


Figure 11. . Ilustrasi Random Forest Regression

```
# Random Forest Regression
from sklearn.ensemble import RandomForestRegressor

rf_model = RandomForestRegressor(n_estimators=1000)
rf_model.fit(x_train, y_train)
y_pred_rf = rf_model.predict(x_test)
```

Figure 12. penerapan model random forest regression

4. *Generalized Additive Models* (GAM)

One of the machine learning techniques used to create prediction models using provided data is Generalized Additive Models (GAM). These prediction models are additive functions of many basis functions that describe the relationship between the response variable and the predictor variables. The response variable in GAM may have an exponential distribution, such as normal, binomial, or Poisson. In addition, GAM can use a linking function that relates the mean of the response variable to the additive function of the predictor variables [32].

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI: 10.32736/sisfokom.v13i2.2097, Copyright ©2024
Submitted: February 23, 2024, Revised: April 15, 2024, Accepted: April 21, 2024, Published: June, 15, 2024

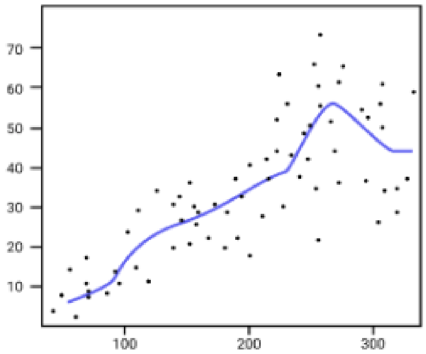
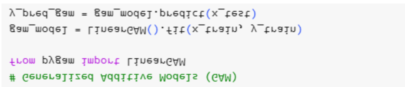


Figure 13. Ilustrasi GAM



```
# Generalized Additive Models (GAM)
from pygam import LinearGAM
gam_model = LinearGAM().fit(X_train, y_train)
y_pred_gam = gam_model.predict(X_test)
```

Figure *14*. penerapan model GAM

c. Mode Evaluation

In the model evaluation stage, there are several types of regression model evaluations used, namely:

1. *Mean Absolute Error* (MAE)

This is the only evaluation metric that takes into account the absolute difference between the value predicted by the model and the true value in the data set, which is the Mean Absolute Error (MAE). In the regression section, MAE shows how far the average prediction of the model is from the true value. To provide inconsistent results about the arousal model, MAE does not take into account the direction of the error (overestimation or underestimation) [33], [34]. There is a large absolute error in the predictions made by MAE. As the MAE value decreases, the quality of the model increases.

MAE formula:

$$MAE = (1/n) * Σ|i=1|^n |y\_i - ŷ\_i| \quad (1)$$

Where:
n is the number of samples in the data
y_i is the actual value
ŷ_i is the predicted value

2. *Coefficient of Determination* (R-squared)

One evaluation metric that can be used to determine how well a model explains the target data is R-squared. R-squared values range from 0 to 1, with higher values indicating a better model explaining the data variables. The R-squared value compares the performance of the model with a baseline model that uses only the absolute values of the target data [34].

R formula² _Score :

$$R^2\_Score = 1 - (SS\_res / SS\_tot) \quad (2)$$

where
SS_res is the residual sum of squares
SS_tot is the total sum of squares.

3. *Mean Squared Error* (MSE)

Mean squared error (MSE) is a measure of the reliability of estimates between model predictions and actual values in the data set. The model will be more robust to outliers or large errors as MSE gives a stronger boost to large errors. Although MSE provides good information regarding the sensitivity of the model, it fails to understand its interpretation as a single parameter [33], [34]. Some large mean square errors can be found in the predictions. The higher the quality of the model, the smaller the MSE value.

MSE formula:

$$MSE = (1/n) * Σ|i=1|^n (y\_i - ŷ\_i)^2 \quad (3)$$

Where:
n is the number of samples in the data
y_i is the actual value
ŷ_i is the predicted value

4. *Root Mean Squared Error* (RMSE)

Root-mean-squared error, often known as RMSE, is a modified version of mean-squared error, also known as MSE, that provides a unit of measurement identical to the target variable. Having one measure in common with the target variable, RMSE is often easier to interpret than MSE. The root-mean-squared error (RMSE) provides more information on how far the prediction model is from the true value [33], [34]. The RMSE measures the comparison of the predicted value to the value that was actually determined before applying the squared difference. The quality of the model increases as the RMSE value decreases.

RMSE formula:

$$RMSE = sqrt((1/n) * Σ|i=1|^n (y\_i - ŷ\_i)^2) \quad (4)$$

Where:

n is the number of samples in the data
y_i is the actual value
ŷ_i is the predicted value

5. *Mean Absolute Percentage Error* (MAPE)

The relative evaluation metric known as MAPE is a kind of accounting for the absolute difference between the value predicted by the model and the true value. In the context of interpreting model results, MAPE provides information about how much the prediction model is accurate as a percentage of the true value [33]. As a percentage of the current value, MAPE estimates a large amount of error in the prediction. The higher the quality of the model, the smaller the MAPE value.

MAPE Formula:

$$MAPE = (1/n) * Σ|i=1|^n |(y\_i - ŷ\_i) / y\_i| * 100\% \quad (5)$$

Where:
n is the number of samples in the data
y_i is the actual value
ŷ_i is the predicted value.

III. Results and Discussion

In this section, we will look at many algorithms, including algorithms that evaluate algorithms using model analysis, algorithms that train new data sets, and algorithms that visualize dashboard results using tables.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI: 10.32736/sisfokom.v13i2.2097, Copyright ©2024
Submitted: February 23, 2024, Revised: April 15, 2024, Accepted: April 21, 2024, Published: June, 15, 2024

A. Person Correlation Analysis Results

From Figure 43 the following person correlation analysis results are obtained:

1. *Correlation between Immunization and Percentage of Stunted Toddlers (-0.33). There is a moderate negative correlation between immunization and the percentage of stunted children under five (-0.33). This indicates that the higher the level of immunization, the lower the percentage of stunted toddlers, and vice versa.*
2. *Correlation between Birth Attendance by Health personnel in health facilities and Percentage of Stunted Children (-0.42). There is a strong negative correlation between delivery assistance by health workers in health facilities and the percentage of stunted children under five (-0.45). This means that the more birth attendance by health workers in health facilities, the lower the percentage of stunted children under five.*
3. *Correlation between Modern Family Planning and Percentage of Stunted Children (-0.3). There is a moderate negative correlation between the use of modern family planning and the percentage of stunted children under five (-0.3). This suggests that the use of modern family planning can be associated with a decrease in the percentage of stunted children under five.*
4. *Correlation between Exclusive Breastfeeding and Percentage of Stunted Children (0.079). There is a very weak positive correlation between exclusive breastfeeding and the percentage of stunted children under five (0.079). Although positive, this relationship is not strong enough to infer a significant linear relationship.*
5. *Correlation between MP ASI and Percentage of Stunted Children (-0.22). There is a weak negative correlation between breastmilk MP and the percentage of stunted children under five (-0.22). This suggests that the higher the percentage of breastmilk MP, the lower the percentage of stunted children under five, although the relationship is not very strong.*
6. *Correlation between Percentage of Potable Water and Percentage of Stunted Children (-0.31). There is a moderate negative correlation between the percentage of potable water and the percentage of stunted children under five (-0.31). This indicates that the higher the percentage of potable water, the lower the percentage of stunted children under five.*
7. *Correlation between Percentage of Proper Sanitation and Percentage of Stunted Children (-0.55). There is a strong negative correlation between the percentage of proper sanitation and the percentage of stunted children under five (-0.55). This indicates that the higher the percentage of proper sanitation, the lower the percentage of stunted children under five.*
8. *There is a weak positive correlation between insufficient food consumption and the percentage of stunted children under five (0.26). This suggests that the presence of insufficient food consumption may be associated with an increase in the percentage of stunted children under five.*
9. *Correlation between ECD and Percentage of Stunted Children (-0.062) There is a very weak negative correlation between ECD and percentage of stunted children (-0.062). This correlation is not strong enough to infer a significant relationship between these two variables.*
10. *Correlation between JKN/Jamkesda Ownership and Percentage of Stunted Children (0.22) There is a weak positive correlation between JKN/Jamkesda ownership and percentage of stunted children (0.22). This suggests that JKN/Jamkesda ownership may be associated with an increase in the percentage of stunted children under five.*
11. *Correlation between Percentage of Poor Population and Percentage of Stunted Children (0.52) There is a moderately strong positive correlation between the percentage of poor population and the percentage of stunted children under five (0.52). This indicates that the higher the percentage of poor people, the higher the percentage of stunted children under five.*

B. Model Evaluation Results

After each algorithm is used in the modeling stage, the model evaluation results can be seen in the following table:

TABLE I. MODEL EVALUATION RESULTS

Algorithm	Model Evaluation					Ranking
	MAE	R^2-Square	MSE	RMSE	MAPE	
SVR	3.616	0.497	18.370	4.286	16.096	3
Decision Tree Regression	2.5	0.637	13.235	3.638	10.117	2
Random Forest Regression	2.367	0.703	10.837	3.292	9.463	1
GAM	3.452	0.265	26.842	5.181	14.206	4

The Random Forest Regression model has the best performance among all models tested, according to the comparison in Table 1. This is because it has the highest R-squared value (0.703), which means this model is able to explain 70.3% of the data variance. The model also has true MAE, MSE, RMSE, and MAPE values that show a low level of error and deviation from the raw data. The success rate of the current model is 90.537% based on the MAPE value of 9.463%. The Generalized Additive Models (GAM) performed poorly compared to all the models tested as it had a low R-squared value (0.265), which means it was only able to explain 26.5% of the observed variance. It also has high MAE, MSE, RMSE, and MAPE values, indicating that it has a high degree of error and deviation from the raw data. The MAPE of this model is 14.207%, which means that the success rate is 85.793%. Among all the models tested, the Decision Tree Regression Model performed the best, with the second-highest R-squared value of 0.638, indicating that it can explain 63.8% of the data variance. It also has two-sided MAE, MSE, RMSE, and MAPE

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI: 10.32736/sisfokom.v13i2.2097, Copyright ©2024
Submitted: February 23, 2024, Revised: April 15, 2024, Accepted: April 21, 2024, Published: June, 15, 2024

values, indicating that it has small standard deviations and errors from the raw data. With a MAPE of 10.118%, this model has a success rate of 89.882%. As it has a low R-squared value (0.497), indicating that it cannot adequately explain the data variables, the Support Vector Regression (SVR) model performed better than all the models tested. This model also has high MAE, MSE, RMSE, and MAPE values, indicating that it has a high degree of error and deviation from the raw data. The MAPE of this model is 16.097%, which means the success rate is 83.903%. The above analysis is a comparison of each model. And the model comparison mechanism is as described in the previous stage, stage F algorithm comparison.

From the results of the model evaluation above, it can be seen the novelty of previous research conducted by [7]. To predict the stunting rate, the best model obtained from the comparison results is the support vector regression model. While in this study the best model is random forest regression. Previous research also reached the stage of model evaluation to compare the best model.

C. Applying the Best Model to a New Dataset

After finding the best model, namely Random Forest Regression. Then the next step is to apply the best model to the new dataset is the Random Forest Regression algorithm to predict stunting rates.

	Provinsi	Tahun	Imunisasi	Penolong persalinan oleh tenaga kesehatan di fasilitas kesehatan	KB modern	ASI eksklusif	MP ASI	Persentase Air Layak Minum	Persentase Sanitasi Layak	Ketidakcukupan konsumsi pangan	PAUD	Kepemilikan JKN/jamkesda	Persentase Penduduk Miskin
0	ACEH	2023	24.79	94.43	43.44	67.05	52.25	89.74	77.38	9.44	32.36	32.32	14.45
1	SUMATERA UTARA	2023	41.04	89.37	41.87	61.98	52.19	92.19	83.36	8.60	23.78	14.90	8.15
2	SUMATERA BARAT	2023	39.12	95.94	45.01	75.84	54.65	85.59	70.66	7.63	28.75	27.26	5.95
3	RIAU	2023	46.03	87.51	46.97	71.14	53.71	90.47	86.26	12.33	23.37	18.10	6.68
4	JAMBI	2023	52.99	81.12	62.89	74.14	56.62	80.02	82.40	12.83	32.44	19.37	7.58
...	...	...	...	...	...	...	...	...	...	...	...	...	...
29	SULAWESI BARAT	2023	58.44	89.50	46.37	75.04	43.68	79.86	78.21	8.29	45.80	23.63	11.49
30	MALUKU	2023	60.12	52.70	33.38	61.52	39.29	92.98	76.93	30.27	31.13	16.59	16.42
31	MALUKU UTARA	2023	52.55	68.33	45.53	69.66	33.03	89.01	77.76	29.56	38.38	18.08	6.45
32	PAPUA BARAT	2023	53.24	70.90	28.18	62.92	45.82	81.57	73.31	24.00	25.66	23.70	20.49
33	PAPUA	2023	50.37	76.19	15.07	55.41	46.17	66.49	50.99	35.63	11.94	23.61	26.03

34 rows × 13 columns

Figure 14. New dataset invocation

```
[27] y_pred_rf = rf_model.predict(x1)
     y_pred_rf

array([30.2929, 22.4872, 24.9295, 20.9142, 21.6128, 21.9246, 21.2954,
       19.3624, 16.4932, 17.9326, 16.6726, 24.4145, 20.7563, 16.4379,
       18.7618, 21.5322, 12.3946, 23.5831, 31.8713, 23.0653, 27.0321,
       22.213 , 21.0654, 23.9077, 20.8977, 27.0009, 20.9451, 23.0767,
       24.0247, 25.9513, 26.7295, 26.7432, 27.2566, 26.8712])
```

```
# Mengubah hasil pemodelan yang  berbentuk array menjadi dataframe
df_y_predict_rf = pd.DataFrame({'Prediction_Persentase Balita Stunting': y_pred_rf}).round(2)
df_y_predict_rf
```

	Prediction_Persentase Balita Stunting
0	30.29
1	22.49
2	24.93
3	20.91
4	21.61
...	...
29	25.96
30	26.73
31	26.74
32	27.26
33	26.87

34 rows × 1 columns

Figure 15. Application of the model to new datasets

	Provinsi	Tahun	Imunisasi	Penolong persalinan oleh tenaga kesehatan di fasilitas kesehatan	KB modern	ASI eksklusif	MP ASI	Persentase Air Layak Minum	Persentase Sanitasi Layak	Ketidakcukupan konsumsi pangan	PAUD	Kepemilikan JKN/jamkesda	Persentase Penduduk Miskin	Prediction_Persentase Balita Stunting	Kategori Stunting
0	ACEH	2023	24.79	94.43	43.44	67.05	52.25	89.74	77.38	9.44	32.36	32.32	14.45	30.29	Tinggi
1	SUMATERA UTARA	2023	41.04	89.37	41.87	61.98	52.19	92.19	83.36	8.60	23.78	14.90	8.15	22.49	Menengah
2	SUMATERA BARAT	2023	39.12	95.94	45.01	75.84	54.65	85.59	70.66	7.63	28.75	27.26	5.95	24.93	Menengah
3	RIAU	2023	46.03	87.51	46.97	71.14	53.71	90.47	86.26	12.33	23.37	18.10	6.68	20.91	Menengah
4	JAMBI	2023	52.99	81.12	62.89	74.14	56.62	80.02	82.40	12.83	32.44	19.37	7.58	21.61	Menengah
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
29	SULAWESI BARAT	2023	58.44	89.50	46.37	75.04	43.68	79.86	78.21	8.29	45.80	23.63	11.49	25.96	Menengah
30	MALUKU	2023	60.12	52.70	33.38	61.52	39.29	92.98	76.93	30.27	31.13	16.59	16.42	26.73	Menengah
31	MALUKU UTARA	2023	52.55	68.33	45.53	69.66	33.03	89.01	77.76	29.56	38.38	18.08	6.45	26.74	Menengah
32	PAPUA BARAT	2023	53.24	70.90	28.18	62.92	45.82	81.57	73.31	24.00	25.66	23.70	20.49	27.26	Menengah
33	PAPUA	2023	50.37	76.19	15.07	55.41	46.17	66.49	50.99	35.63	11.94	23.61	26.03	26.87	Menengah

34 rows × 15 columns

Figure 16. Application of category boundaries in prediction results

Based on the stunting prediction results above, it can be seen that the province with the highest percentage of stunting is East Nusa Tenggara at 31.8 percent, which is in the highest category. This province requires special attention and priority in stunting prevention. Bali is the province with the highest proportion of stunted toddlers at 12.07%, which is in the middle-rich category. This province can be an example and source of inspiration for other provinces in implementing stunting programs. Of the 34 provinces surveyed, 28 provinces have stunted children in the male category, with percentages ranging from 20% to 30%. To reduce stunting, these provinces need a more intense and integrated campaign.

D. Dashboard Visualization

Here, data visualization in dashboard format is done using the stunting probability prediction dataset that has been previously exported to EXCEL format. In this research, Tableau is used as a tool for creating dashboards. When it comes to data visualization and business analysis, Tableau is the tool of choice. Data from various sources can be connected and underlined by visualizing graphs, tables, and interactive dashboards that use Tableau. Tableau can be used to generate reports and analyses to help make better business decisions [35].

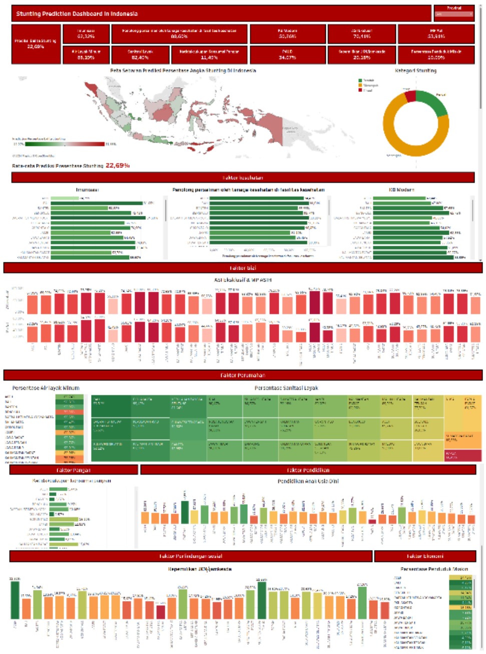


Figure 17. *Dashboard* visualization results

Based on the data above, it can be concluded that stunting will become a problem, amounting to 22.69% in Indonesia by 2023. Based on the category in each province, the high category is at 5.88%, the medium is at 73.53%, and the low is at 20.59%. These viewpoints may come from factors identified in the previous modeling stage and incorporated into the dashboard.

IV. Conclusion

In children under the age of five, stunting is a significant health problem as it is a condition of stunted growth caused by chronic geriatric (under-five) disease. Despite efforts to address this problem, stunting has a prevalence of 24.4% in Indonesia

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI: 10.32736/sisfokom.v13i2.2097, Copyright ©2024
Submitted: February 23, 2024, Revised: April 15, 2024, Accepted: April 21, 2024, Published: June, 15, 2024

by 2022. The limitation of this research is only to predict stunting rates in Indonesia through machine learning after model comparison. This research aims to compare the performance of regression algorithms in predicting stunting prevalence in Indonesia. This study compares the performance of regression algorithms in predicting the prevalence of stunting in Indonesia. Combining information from many surveys. The model with a 90.537% evaluation success rate was determined as the best model through a comparison of random forest regression models. The highest stunting prediction rate using the best algorithm was 31.8 percent in East Nusa Tenggara Province. The limitations of this study are still the lack of research yet in predicting stunting with the same model and evaluation of this model. And also the data obtained is quite difficult to obtain.

Arguments put forward by researchers Nutrition interventions are sensitive to the factors that cause stunting and specific nutrition interventions that affect pregnant women, infants, and toddlers in reducing stunting rates in Indonesia. For effective monitoring of soil moisture levels, it is necessary to drill down to the under-five nutrition level. Program effectiveness and avoiding overlap achieve coordination across sectors and levels. Finally, advocacy, campaigns, and socialization are needed to increase public awareness about the importance of balanced nutrition and good posture during stunting. It is hoped that further research can be developed and implemented directly such as given to policy makers so that the results of this study can be taken into consideration to reduce stunting rates.

Acknowledgments

The researcher would like to thank Tadulako University for the extraordinary opportunity to participate in the Merdeka Belajar Kampus Merdeka (MBKM) program, especially in the Independent Study Program.

References

[1] A. M. K. Evi Nuryuliyani, "Getting to Know More about Stunting," https://yankes.kemkes.go.id/, 2023. https://yankes.kemkes.go.id/view_artikel/2657/mengenal-lebih-jauh-tentang-stunting#:~:text=Although declining%2C the rate is still, in the first 1,000 days of life. (accessed Feb. 06, 2024).

[2] V. Ratriani, "Stunting According to WHO is a Chronic Malnutrition Problem, This is its Impact and Prevention," *health.kontan.co.id*, 2023. https://kesehatan.kontan.co.id/news/stunting-menurut-who-adalah-masalah-kurang-gizi-kronis-ini-dampak-dan-pencegahannya-1 (accessed Feb. 10, 2024).

[3] stunting. Go.id, "What is stunting?" *stunting.go.id*, 2020. https://stunting.go.id/faq/apakah-yang-dimaksud-dengan-stunting/ (accessed Feb. 07, 2023).

[4] Who, "Stunting prevalence among children under five years of age (%) (model-based estimates)," www.who.int, 2023.

[5] N. O. Nirmalasari, "Child Stunting: Causes and Risk Factors of Stunting in Indonesia," *Qawwam J. Gend. Mainstreaming*, vol. 14, no. 1, pp. 19-28, 2020, doi: 10.20414/Qawwam.v14i1.2372.

[6] Ministry of Health of the Republic of Indonesia, "Handbook on the Results of the Study of the Status of Nutrition of Indonesia (SSGI) at the National, Provincial, and District/City Levels in 2021," 2021.

[7] Ministry of Health, "Nutrition Status Survey of Indonesia (SSGI) 2022 Results," *Ministry of Health*, pp. 1-150, 2022.

[8] M. S. Haris, A. N. Khudori, and W. T. Kusuma, "Comparison of Supervised Machine Learning Methods for Prediction of Stunting Prevalence in East Java Province," *J. Technol. Inf. and Comput. Science*, vol. 9, no. 7, p. 1571, 2022, doi: 10.25126/jtiik.2022976744.

[9] National Development Planning Agency (Bappenas), "Rpjmn 2020-2024," *Natl. Mid-Term Dev. Plan 2020-2024*, p. 313, 2020, [Online]. Available: https://www.bappenas.go.id/id/data-dan...dan.../rpjmn-2015-2019/

[10] peraturan.bpk.go.id, "Acceleration of Stunting Reduction," *peraturan.bpk.go.id*, 2021. https://peraturan.bpk.go.id/Details/174964/perpres-no-72-tahun-2021#:~:text=This presidential decree regulates, among others, reporting%3B and 5) funding. (accessed Feb. 11, 2024).

[11] stunting. Go.id, "*Stunting* and its Causes," *stunting.go.id*, 2021. https://dashboard.stunting.go.id/latar-belakang-home/ (accessed Feb. 06, 2024).

[12] P. D. Kusuma, "Machine Learning Theories, Programs, and Case Studies," https://books.google.co.id/, 2020. https://books.google.co.id/books/about/Machine_Learning_Teori_Program_Dan_Studi.html?id=4k3sDwAAQBAJ&redir_esc=y (accessed Feb. 13, 2024).

[13] E. Retnoningsih and R. Pramudita, "Getting to Know Machine Learning with Supervised and Unsupervised Learning Techniques Using Python," vol. 7, no. 2, pp. 156-165, 2020.

[14] Farah Imaniar R, "Linear Regression in Machine Learning," *medium.com*, 2019. https://medium.com/group-3-machine-learning/regresi-linear-pada-machine-learning-97f34ce18633 (accessed Feb. 13, 2024).

[15] D. Store, "7 Types of Research Methodology and Full Explanation," *deepublishstore.com*, 2023. https://deepublishstore.com/blog/jenis-metodologi-penelitian/ (accessed Feb. 13, 2024).

[16] R. Pratama, "Research Methods: Definition, Types, and Examples (+PDF)," *bocahkampus.com*, 2024. https://bocahkampus.com/metode-penelitian (accessed Feb. 13, 2024).

[17] B. Central Bureau of Statistics, "Health Statistics Profile 2023," 2023.

[18] B. Statistics Indonesia, "Settlement and Housing," *bps.go.id*, 2023. https://www.bps.go.id/id/statistics-table?subject=525 (accessed Feb. 14, 2024).

[19] B. Statistics Indonesia, "Consumption and Income," *bps.go.id*, 2023. https://www.bps.go.id/id/statistics-table?subject=523 (accessed Feb. 14, 2024).

[20] B. Central Bureau of Statistics, *Front cover of CHILDHOOD PROFILE 2021 i*. 2021.

[21] R. Agustina, "Front cover of EDUCATION STATISTICS 2021 i," *Pus. Stat.*, 2021.

[22] BPS, "Indonesian Education Statistics 2022," *Badan Pus. Stat.*, no. February, pp. 1-353, 2022.

[23] BPS 2022, "Statisti Pendidikan 2023," *Badan Pus. Stat.*, vol. 1101001, p. 790, 2023.

[24] B. Statistics Indonesia, "Maternal and Child Health Profile 2022," pp. 7823-7830.

[25] National Bureau of Statistics, "Poverty Profile in Indonesia," Badan *Pus. Stat.*, vol. 01, no. 05, pp. 1-8, 2022.

[26] Central Bureau of Statistics of Indonesia, "Poverty Profile in Indonesia September 2022," *Ber. Official Stat.*, vol. 1, no. 5, p. 8, 2023.

[27] "Poverty Profiles in Indonesia March 2023," no. 47, 2023.

[28] A. Pragota, "Scatterplot: Visualization for Regression," *learn.nural.id*, 2021. https://learn.nural.id/course/statistics/regresi-linier/scatterplot (accessed Feb. 15, 2024).

[29] Trivusi, "Support Vector Regression (SVR) Algorithm:) A Type of SVM for Regression," *trivusi. Web. Id*, 2022. https://www.trivusi.web.id/2022/08/algoritma-svr.html (accessed Feb. 15, 2024).

[30] Herlambang, "Machine Learning: Decision Tree Regression," *megabagus. Id*, 2019. https://www.megabagus.id/machine-learning-decision-tree-regression/ (accessed Feb. 15, 2024).

[31] R. Mulywan, "Random Forests," *rifqimulyawan.com*, 2024. https://rifqimulyawan.com/kamus/random-forests/ (accessed Feb.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI: 10.32736/sisfokom.v13i2.2097, Copyright ©2024
Submitted: February 23, 2024, Revised: April 15, 2024, Accepted: April 21, 2024, Published: June, 15, 2024

15, 2024).

[32] Binus, "Generative Additve Model (GAM)," *binus.ac.id*, 2021. https://mti.binus.ac.id/2021/12/31/generative-additve-model-gam/ (accessed Feb. 15, 2024).

[33] Trivusi, "Difference between MAE, MSE, RMSE, and MAPE in Data Science," *trivusi. Web. Id*, 2023. https://www.trivusi.web.id/2023/03/perbedaan-mae-mse-rmse-dan-mape.html (accessed Feb. 15, 2024).

[34] Datatechnotes, "Regression Accuracy Check in Python (MAE, MSE, RMSE, R-Squared)," *datatechnotes.com*, 2019. datatechnotes (accessed Feb. 15, 2024).

[35] Tableau, "What is Tableau?" *tableau.com,* 2021. https://www.tableau.com/why-tableau/what-is-tableau (accessed Feb. 15, 2024)

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI: 10.32736/sisfokom.v13i2.2097, Copyright ©2024
Submitted: February 23, 2024, Revised: April 15, 2024, Accepted: April 21, 2024, Published: June, 15, 2024