International Journal of Computer, Network Security and Information System (IJCONSIST)
Vol. 6 No. 1, September 2024

Hybrid Machine Learning to Evaluate the Incidence of Toddler Stunting through Integration of Multi-source Satellite Imagery and Official Statistics in East Nusa Tenggara Province

Suhendra Widi Prayoga
Department of Statistical Computing
Politeknik Statistika STIS
Jakarta, Indonesia
222112382@stis.ac.id

Setia Pramana
Department of Statistical Computing
Politeknik Statistika STIS
Jakarta, Indonesia
setia.pramana@stis.ac.id

***Abstract*— Stunting is a serious health problem that impacts the quality of life of children under five. In 2023, East Nusa Tenggara recorded the second highest prevalence of stunting in Indonesia, influenced by health, socio-economic and environmental factors. In terms of the environment, remote sensing technology can be utilised to monitor environmental factors that contribute to stunting, such as vegetation conditions, access to clean water, and soil conditions. This study aims to evaluate the incidence of stunting among children under five using a hybrid machine learning approach, combining predictive modeling and cluster analysis. The results indicate that eXtreme Gradient Boosting Regressor (XGBR) is the best model for estimating stunting prevalence, with a Root Mean Squared Error (RMSE) of 3.2076 and an R^2 value of 0.7223. Meanwhile, for clustering results, K-Means Clustering is identified as the most effective method for grouping districts/cities based on socioeconomic and environmental factors. The clustering process produced two groups, such as vulnerable (Cluster 1) and highly vulnerable (Cluster 2), with connectivity, Dunn Index, and silhouette coefficient values of 29.290, 0.6931, and 0.4509, respectively. These findings are expected to serve as a basis for policymakers in formulating targeted interventions to reduce stunting rates, particularly in highly vulnerable areas.**

Keywords*— machine learning, East Nusa Tenggara, official statistics, remote sensing, stunting

I. INTRODUCTION

Malnutrition poses a threat to the life of an individual and the development of a country, requiring significant attention to address the problem [1]. Stunting, one of the effects of malnutrition, is a serious health problem for under-fives as it can impair their physical and cognitive development and significantly impact quality of life. According to WHO, stunting is a growth problem in children resulting from chronic malnutrition and recurrent infections, characterised by lower than average height [2]. This is one of the main focuses in achieving Sustainable Development Goals (SDGs) goal 2, target 2.2, where all forms of malnutrition must end by 2030, including achieving targets related to stunting and wasting in children under five.

Based on WHO and UNICEF data, Indonesia ranks 27th out of 154 countries in the world for stunting prevalence and 5th in Asia [3], [4]. This shows the urgent need to take effective action to address this issue, especially in areas with a very high prevalence of stunting. The Indonesia Health Survey 2023 reported that East Nusa Tenggara Province has the second highest prevalence of stunting in Indonesia, at 37.9 percent [5]. Stunting can be influenced by various factors, such as environmental conditions, access to health services, and poor access to education [6]. According to [7], children aged 24-35 months, mothers with low education levels, and rural environments are factors that contribute to stunting in this region. In addition, environmental conditions can be an indirect causal factor for stunting in children [8].

NTT province is known as an archipelagic dryland region that has various limitations, including in health aspects. The region is dominated by hills, high levels of drought, and low rainfall, which contribute to limited access to clean water and health services. In general, NTT province has a hot and dry climate, with average temperatures reaching 27-28°C and very limited rainfall, ranging from 600–4,800 mm^3 per year. These factors worsen public health conditions and increase the risk of malnutrition. In addition, several infectious diseases such as helminthiasis and malaria are still significant health problems, with the incidence of helminthiasis in school children in Sumba Islands reaching 90% in 2019. Despite health efforts, including malaria elimination programmes, nutritional problems such as stunting remain a major challenge [9]. In fact, a healthy environment plays an important role in supporting optimal child growth, while a poor environment can increase the risk of disease and exacerbate malnutrition [10]. Stunting is a multidimensional problem that is not only caused by

malnutrition in pregnant women and toddlers, so its handling efforts require the involvement of various sectors [11].

Utilising remote sensing technology as an alternative to monitoring environmental conditions indirectly through multi-source satellite imagery offers a number of advantages, such as the ability to collect data on a large scale (big data) and in high detail. Study [12] utilised Landsat-8 satellite imagery to analyse the incidence of stunting in Lampung using Artificial Neural Network. In addition, the study [13], [14] utilized a predictive model to estimate stunting prevalence in Bangladesh and India using remote sensing data. Mapping was done based on stunting predictions with the algorithm, showing the potential of machine learning in identifying areas with high stunting risk. Study that integrates remote sensing in stunting study in Indonesia is still very limited. In addition, condition monitoring also requires statistical data from official government agencies, such as the Central Bureau of Statistics and the Ministry of Health. Study [1] used official statistics data to map provinces based on a special index for handling stunting in Indonesia through hierarchical and nonhierarchical clustering methods. In addition, [15] mapped the prevalence of stunting in South Kalimantan Province using K-Means with several distance measurement scenarios, such as Canberra, Chebyshev, Euclidean, and Manhattan. Distance measurement using Canberra produced three clusters, while the other three methods produced five clusters. Study by [16] also used the K-Means method, but with a smaller area coverage, namely at the village level, by utilising Posyandu data in Tegalwangi Village. The integration of multi-source satellite image data and official statistics allows for a more accurate and comprehensive mapping of stunting incidence in East Nusa Tenggara.

Therefore, this study conducted a hybrid machine learning analysis in the form of predictive modelling and mapping of districts/cities based on indicators of stunting incidence in East Nusa Tenggara Province through the integration of multi-source satellite imagery and official statistics. Supervised learning methods used include Random Forest Regression (RFR), Support Vector Regression (SVR), and eXtreme Gradient Boosting Regression (XGBR) to predict stunting prevalence. Meanwhile, the unsupervised learning methods applied include K-Means Clustering, K-Medoids Clustering, and Fuzzy C-Means Clustering to group regions based on their level of vulnerability to stunting. The results of the study are expected to serve as a basis for the government in formulating more effective policies and strategies to address the problem of stunting, especially in East Nusa Tenggara, and support the achievement of Sustainable Development Goals (SDGs) goal 2, target 2.2, which is to end all forms of malnutrition by 2030.

II. METHODOLOGY

A. Study Area

This study takes the study location in East Nusa Tenggara Province, Indonesia in 2023 as shown in Fig. 1. East Nusa Tenggara Province is a province located in the eastern part of Indonesia, adjacent to Maluku and Papua Islands. This province was chosen as the study area for this study because it has a high prevalence of stunting each year.

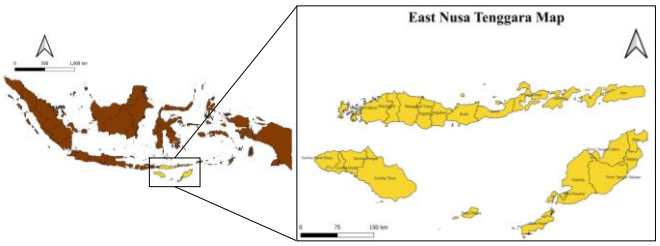


Fig. 1. Study area map.

B. Data Sources

This study uses official statistics data obtained from the Central Bureau of Statistics and the Ministry of Health as well as satellite images in the form of composite indices such as Normalised Difference Vegetation Index (NDVI), Soil Adjusted Vegetation Index (SAVI), Normalised Difference Built-up Index (NDBI), Normalised Difference Water Index (NDWI) from Landsat-8 and Land Surface Temperature (LST) from MODIS obtained through Google Earth Engine processing. The following are the details of the variables in this study. Table **1** is a breakdown of the variables used in this study.

TABLE I. DATA AND SOURCES

Variable	Description	Unit	Source	Reference
wasting	Children under five experiencing wasting	Percent	Ministry of Health RI	[17]
imunisasi	Population aged 0-59 months with complete immunization	Percent	Central Bureau of Statistics	[18]
sanitasi	Households with access to proper sanitation	Percent		[17]
asi_eks	Population aged 0-23 months who have ever been breastfed	Percent		[17]
air_minum	Households with access to safe drinking water	Percent		[19]
miskin	Population in poverty	Percent		[18]
rls_p	Average years of schooling for females	Years		[20]
sma_p	Females aged 15 years and above with a high school diploma or equivalent	Percent		[18]
huruf_p	Illiterate females	Years		[21]
konsumsi	Average daily per capita calorie consumption	kcal		[19]
NDVI	*Normalized Difference Vegetation Index*	Index	Landsat-8	[12]
NDWI	*Normalized Difference Water Index*	Index		
SAVI	*Soil Adjusted Vegetation Index*	Index		
NDBI	*Normalized Difference Built-up Index*	Index		
LST	Land Surface Temperature	Celcius	MODIS	[12]

C. Machine Learning

Machine learning approaches are classified into supervised and unsupervised learning. Supervised learning trains models using labeled data, where input features are paired with known outputs [22]. In this study, Random Forest Regression (RFR), Support Vector Regression (SVR), and eXtreme Gradient Boosting Regression (XGBR) were used in regression modeling to estimate stunting prevalence by establishing relationships between predictors and target values for accurate predictions. Unsupervised learning is a machine learning algorithm where the algorithm is trained using unlabelled data [22], [23]. This study uses unsupervised algorithms in the form of K-Means, K-Medoids, and Fuzzy C-Means clustering. The three algorithms include non-hierarchical clustering designed to group objects into a number of *k* clusters determined in advance [1].

1) Predictive Modelling

In this analysis, predictive modelling was conducted to estimate the prevalence of stunting using a machine learning approach based on remote sensing and official statistics variables. The three main models applied in this study are Random Forest Regression (RFR), Support Vector Regression (SVR), and eXtreme Gradient Boosting Regression (XGBR).

Random Forest Regression (RFR)

Random Forest Regression is an ensemble machine learning method that combines multiple decision trees to produce more accurate and stable predictions. This algorithm builds several decision trees randomly during training and then aggregates the predictions from each tree to obtain a final, robust value. The main advantages of RFR include its ability to handle data with many features, its use of bagging (bootstrap aggregation) to reduce the risk of overfitting, and its capability to capture non-linear relationships in data [13].

Support Vector Regression (SVR)

Support Vector Regression is a machine learning method based on Support Vector Machines (SVM) designed for regression tasks. This algorithm works by finding the best function that models the relationship between independent and dependent variables within a specified margin of error. SVR utilizes hyperplanes and kernel functions to capture non-linear patterns in data, making it applicable to various datasets, including those with high-dimensional features or non-linear distributions [24]. The main advantages of SVR lie in its strong generalization ability, even with a limited training dataset, and its robustness against outliers due to the margin-based optimization approach [25].

eXtreme Gradient Boosting Regression (XGBR)

eXtreme Gradient Boosting Regression is a tree-based boosting algorithm designed to enhance prediction accuracy by optimizing the learning process iteratively. This algorithm employs gradient boosting techniques, where errors from previous models are corrected by building new models that improve overall performance. XGBR is known for its efficiency in handling large datasets, its ability to capture complex variable interactions, and its built-in regularization features that help reduce the risk of overfitting. Additionally, XGBR has a high training speed and often delivers more accurate results than other regression models, making it one of the most popular algorithms for predictive modeling [26].

2) Clustering Analysis

Before clustering, multicollinearity checks need to be carried out on each variable in the data by looking at the correlation between variables. The correlation criteria can be said to have multicollinearity if ≥ 0.8 so that the data needs to be resolved. Principal Component Analysis (PCA) is one method to overcome multicollinearity in data by reducing the complexity of data dimensions to be simpler [27]. The variable reduction process forms new components that retain total variation so as not to eliminate the entire initial information. The resulting new components are independent of each other, and each component can represent or replace information from the variables in the data [1]. The number of principal components can be determined using a sloping scree plot, a cumulative proportion of total variance of at least 80%, and an eigenvalue > 1 [28].

K-Means Clustering

K-Means clustering method has been introduced by Macqueen since 1967. K-Means is a distance-based unsupervised learning algorithm whose ability is suitable for large datasets quickly and efficiently. Small distance differences will give significant similarities between objects. The K-Means clustering algorithm starts by determining the desired number of clusters k, usually with Elbow and Silhouette plots. After that, data points are randomly selected to be the initial centres of each cluster. Next, the distance between each data point and the cluster centre is calculated using euclidean distance. Based on this calculated distance, each data point is then placed into the cluster with the closest centre. After all data points are grouped into clusters, the cluster centres are updated by calculating the average position of all data points in the cluster. This process of placing data points into clusters and updating the cluster centre is repeated continuously until the position of the cluster centre no longer changes, indicating that the cluster has reached stability [15].

K-Medoids Clustering

One of the developments of K-Means clustering is K-Medoids which is designed to be more resistant to outliers and noise. K-Medoids uses median values that are not affected by outliers and noise, making it more robust than K-Means [29]. In the K-Medoids method, each cluster is represented by one object located near the cluster centre, with the aim of reducing the amount of dissimilarity between each object and its representative object. The process of cluster formation with K-Medoids begins by determining the desired number of initial clusters k. Next, each object is grouped into clusters. Next, each object is grouped into clusters based on its closest distance to the representative object. New objects are then randomly selected to serve as medoids, and this new set of medoids is used to recalculate the cost. If the resulting cost is less than zero,

the jth medoid is replaced with a new random medoid from the dataset. These steps are repeated until there is no further change in the medoids, which signifies that no objects have moved anymore [1].

Fuzzy C-Means Clustering

Fuzzy C-Means method is a fuzzy version of the hard clustering algorithm. Fuzzy C-Means is a clustering method that is often used because it is easy and efficient [30]. The algorithm involves fuzzy membership μ_{ij} and cluster centroid V_i, with exponent m used to minimise the objective function and set the fuzziness level. A value of $m = 1$ produces a clear result, while a value of m close to ∞ produces perfect fuzziness. Usually, the best value of m between 1 and 3 is used for more accurate results. Equation (1) is an objective function in Fuzzy C-Means clustering [1].

$$J_m = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} d\left(X_j, V_i\right)^2 \quad (1)$$

where μ_{ij}: the membership value of the jth object in the i-th cluster, $d(X_i, V_i)$: the distance between the j-th object (X_i) and the centroid of the i-th cluster (V_i), n: the number of observations, and m: the fuzziness parameter of the clustering result ($m > 1$). The Fuzzy C-Means algorithm starts by initialising the membership value μ_{ij} of the object (X_j) in the i-th cluster, provided that the membership number of each object is equal to 1. Next, the fuzzy centroid (V_i) is calculated and continued by updating the fuzzy membership value. The step is repeated until no J_m value decreases [1].

D. Model Evaluation

Evaluation of the cluster results was conducted to determine the best method based on the following internal validation criteria:

1) Predictive Modelling

This study employs two model evaluation metrics in predictive modeling, such as coefficient of determination (R^2) and Root Mean Square Error (RMSE). These metrics assess the model's performance in estimating stunting prevalence based on official statistics and remote sensing variables.

Coefficient of Determination (R^2)

Coefficient of determination (R^2) measures how well the model explains the variation in the target data. Its value ranges from 0 to 1, where a higher value indicates a better model fit. Equation (2) is the R^2 formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (2)$$

where n is the number of observations, y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and $\bar{y}_i$ is the mean of the actual values [24].

Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) measures the average prediction error of the model, maintaining the same unit as the original data. A lower RMSE value indicates higher prediction accuracy [25]. Equation (3) is the RMSE formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

where n is the number of observations, y_i represents the actual values, and $\hat{y}$ represents the predicted values.

2) Clustering Analysis

This study employs three internal validity metrics in cluster analysis, such as connectivity coefficient, Dunn index, and silhouette coefficient.

Connectivity Coefficient

The lower the connectivity coefficient value, the better the cluster formed. The connectivity coefficient has a range from zero (0) to infinity [1]. Equation (4) is the connectivity coefficient formula.

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}} \quad (4)$$

where $Conn(C)$: connectivity coefficient, $x_{i,nn_{i(j)}}$: zero if i and j are in the same cluster and $1/j$ otherwise, $nn_{i(j)}$: the j-th object's nearest neighbour of the i-th object, N: the number of observations, and L: the parametric number of nearest neighbours considered.

Dunn Index

Dunn's index is the ratio between the minimum distance between observations in clusters that have differences and the largest intra-cluster distance. The Dunn index has a range of values from zero to infinity. The higher the Dunn index value in a cluster, the better the cluster results [1]. Equation (5) to calculate the Dunn index.

$$D(C) = \frac{\min\limits_{C_k, C_l \in C, C_k \neq C_l} \left(\min\limits_{i \in C_k, j \in C_l} dist(i,j) \right)}{\max\limits_{C_m \in C} \left(diam(C_m) \right)} \quad (5)$$

where $D(C)$: the Dunn index and $diam(C_m)$: the maximum distance between observations in cluster C_m.

Silhouette Coefficient

The Silhouette coefficient is used as a metric to evaluate the reliability of the clustering process of an observation. The Silhouette coefficient value ranges from -1 to 1. A Silhouette coefficient value close to 1 indicates a good cluster, while one close to -1 is considered bad [1]. Equation (6) is the Silhouette coefficient formula.

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (6)$$

where $S(i)$: the Silhouette coefficient, a_i: the average distance between the i-th object and all objects in the same cluster, and b_i: the average distance between the i-th object and objects in the nearest neighbour cluster.

III. RESULTS

This section presents the results of our analysis using various machine learning approaches, including predictive modeling, clustering analysis, as well as evaluation and interpretation of findings. Each method is employed to identify patterns, key determinants, and data structures relevant to understanding the incidence of stunting in children.

A. Estimation Modelling Results with Machine Learning

Predictive modeling using machine learning regression was conducted to identify factors that significantly contribute to the incidence of stunting in children in East Nusa Tenggara Province. In this study, several regression algorithms were utilized, including Random Forest Regressor (RF), Support Vector Regression (SVR), and XGBoost Regressor (XGBR), to evaluate model performance in predicting stunting prevalence based on the available predictor variables.

To achieve optimal model performance, hyperparameter tuning was performed using the Grid Search Cross-Validation (GridSearchCV) technique. This approach allows for the exploration of the best hyperparameter combinations based on predefined evaluation metrics, namely the coefficient of determination (R^2) and Root Mean Squared Error (RMSE). Table 2 presents the optimal hyperparameter combinations obtained for each model after tuning with GridSearchCV.

TABLE II. HYPERPARAMETER OPTIMAL FROM ML REGRESSION

Algorithm	***Hyperparamater Optimal***
RFR	{'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 200}
SVR	{'C': 0.1, 'epsilon': 1, 'kernel': 'linear'}
XGBR	{'learning_rate': 0.01, 'max_depth': 7, 'n_estimators': 100}

Based on Table 2, each algorithm has an optimal combination of hyperparameters tailored to enhance model performance. Next, predictive modeling (regression) will be conducted using these optimal hyperparameters. In this study, data splitting is not performed due to sample size considerations, ensuring that all available data is utilized for model training and evaluation. The estimated prevalence of stunting in Nusa Tenggara Timur Province, based on remote sensing and official statistics variables, is presented in Fig. 2.

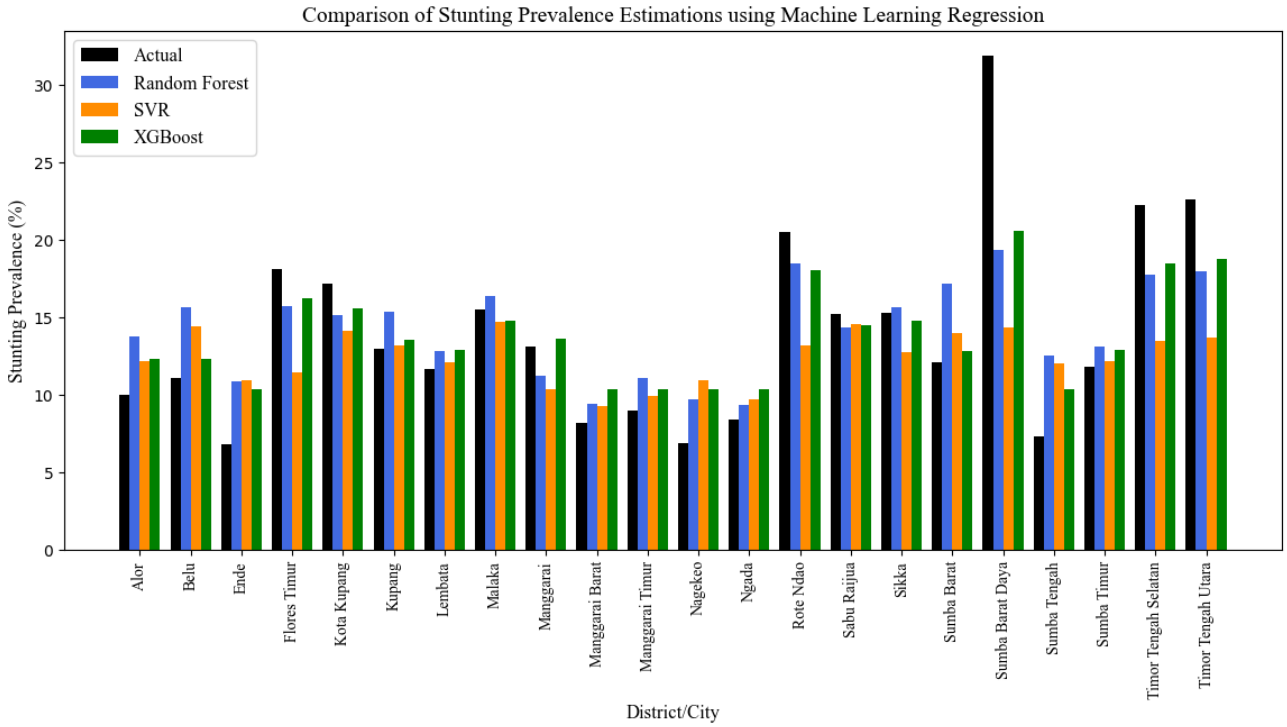


Fig. 2. Estimated stunting prevalence using four machine learning regression approaches.

Based on Figure 2, it can be seen that the XGBR algorithm is able to produce stunting prevalence estimates that are closest to the actual data compared to other models. For example, in North Central Timor District, the estimated stunting prevalence with XGBR is 18.78, while the actual data is 22.6. This difference is smaller than the estimates produced by RFR (17.96) and SVR (13.68), indicating that XGBR has a better level of accuracy in predicting stunting prevalence. These estimation results are in line with the model evaluation comparison shown in Table 3.

TABLE III. COMPARISON OF REGRESSION MODEL EVALUATION

Algorithm	***RMSE***	***R^2***

RFR	3.9682	0.5750
SVR	5.4973	0.1843
XGBR	3.2076	0.7223

Table 3 shows that the XGBoost Regressor (XGBR) outperforms other models in estimating stunting prevalence in East Nusa Tenggara Province. With the lowest RMSE (3.2076) and the highest R^2 (0.7223), XGBR demonstrates superior accuracy and predictive capability. The main advantages of XGBR over other algorithms lie in its ability to handle complex data patterns and mitigate overfitting through regularization techniques. Additionally, XGBR efficiently utilizes boosting to enhance accuracy by iteratively combining multiple decision trees. Compared to the RFR and SVR, XGBR is more adaptive to nonlinear relationships in data, more robust against outliers, and better at capturing feature interactions [26].

From the best model obtained (XGBR), a feature importance analysis was conducted to identify variables that have the greatest influence in estimating stunting prevalence. By using XGBoost Regressor (XGBR), the contribution of each variable can be evaluated based on its importance in the prediction process. Figure X presents the feature importance results from the XGBR modelling results.

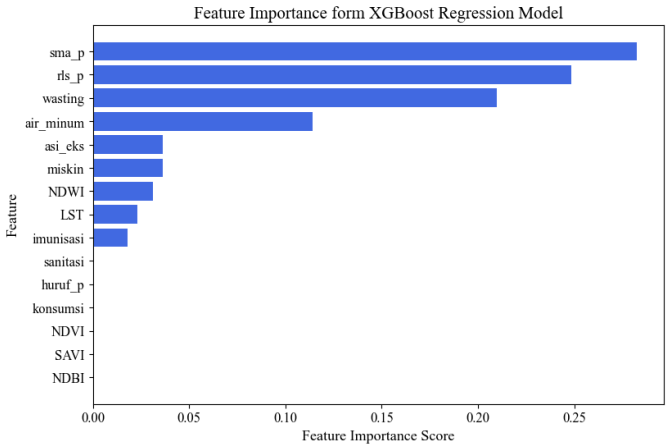


Fig. 3. Feature importance results from XGBR modelling.

Based on Fig. 3., the variables from the official statistics that have the largest influence in estimating the prevalence of stunting are sma_p (percentage of women aged 15 years and over who have a high school diploma or equivalent), rls_p (average years of schooling for women), and wasting (percentage of children under five who are wasted). The high contribution of sma_p and rls_p indicates that the education level of women, especially mothers, plays an important role in reducing the risk of stunting. Higher education can improve mothers' understanding of nutrition, health, and good parenting, thus contributing to more optimal child growth. Meanwhile, the wasting variable also has a significant effect, reflecting that poor nutrition in children under five is closely related to the prevalence of stunting. Children who experience wasting tend to have a higher risk of experiencing stunted growth in the long term.

On the other hand, remote sensing-based variables, such as NDWI (Normalized Difference Water Index) and LST (Land Surface Temperature), suggest that environmental factors also contribute to the estimation of stunting prevalence, albeit with a lesser degree of influence than socio-economic and health variables. NDWI relates to the availability of water resources, which can affect access to clean water and sanitation, important factors in stunting prevention. Meanwhile, LST is related to ambient temperature, which can impact children's health, especially in extreme temperature conditions that can worsen nutritional status and increase the risk of infectious diseases.

B. Clustering Analysis Results with Machine Learning

An assumption that must be met before cluster analysis is checking for non-multicollinearity. Based on the correlation of each variable, it is known that all remote sensing index variables have a correlation of more than 0.80. This indicates the presence of multicollinearity in the data, so it needs to be overcome so that further analysis can be carried out in the form of clustering. Furthermore, analysis using Principal Component Analysis (PCA) is carried out by forming principal components which are linear combinations of variables in the data.

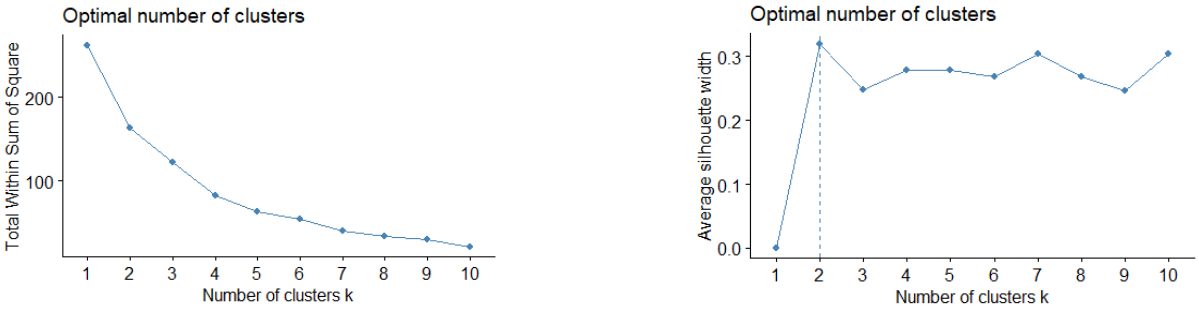
TABLE IV. DETERMINATION OF THE PRINCIPAL COMPONENT

Principal Component	Eigen Value	Variance (%)	Variance Cumulative (%)
1	5.8593	0.3906	39.0619
2	4.0405	0.2694	65.9984
3	1.4295	0.0953	75.5284
4	1.1618	0.0775	83.2737
5	0.9156	0.0610	89.3778
6	0.5710	0.0381	93.1845
7	0.3497	0.0233	95.5157
8	0.3169	0.0211	97.6281
9	0.1597	0.0106	98.6928
10	0.0735	0.0049	99.1826
11	0.0554	0.0037	99.5517
12	0.0326	0.0022	99.7690
13	0.0277	0.0018	99.9539
14	0.0052	0.0003	99.9887
15	0.0017	0.0001	100

Based on Table 4, it can be seen that the number of main components formed is four because the eigenvalue > 1 and the cumulative variance of 83.27%. The data diversity that can be explained by 83.27% is sufficient to describe the data structure. Thus, this study uses four main components for further analysis in the form of clustering.

Modelling with K-Means Clustering

The K-Means method clusters each observation (object) based on the closest centroid. First, the optimal number of clusters is determined using the Elbow and Silhouette plots. Fig. 4. shows the optimal number of clusters for the K-Means method.



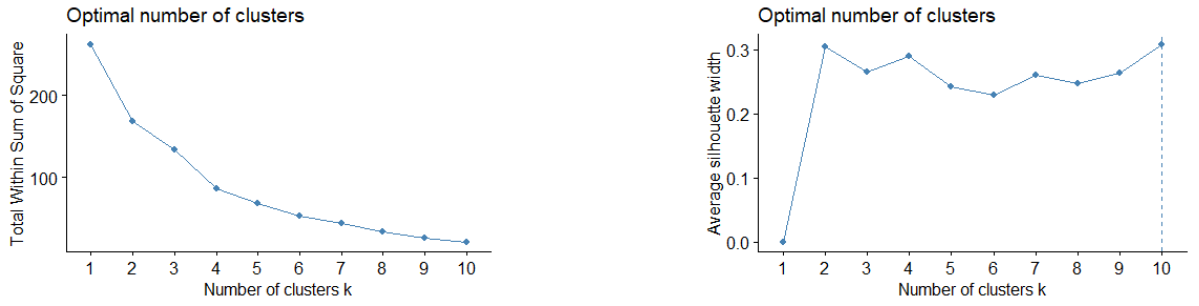
(a) (b)

Fig. 4. Determination of the optimal number of K-Means clusters (a) Elbow plot (b) Silhouette plot.

Based on Fig. 4., the optimal number of K-Means clusters formed is two clusters. Therefore, the number of clusters of two will be used for further analysis. The evaluation results in the form of internal validation obtained connectivity coefficient, Dunn index, and Silhouette coefficient of 2.9290, 0.6931, and 0.4509, respectively.

Modelling with K-Medoids Clustering

The K-Medoids method is more robust than K-Means which uses the median because it is not affected by outliers and noise. The first step in K-Medoids is the same as K-Means, which is to determine the optimal number of clusters with the Elbow and Silhouette methods. Fig. 5. shows the graph of the optimal number of clusters with the K-Medoids method.



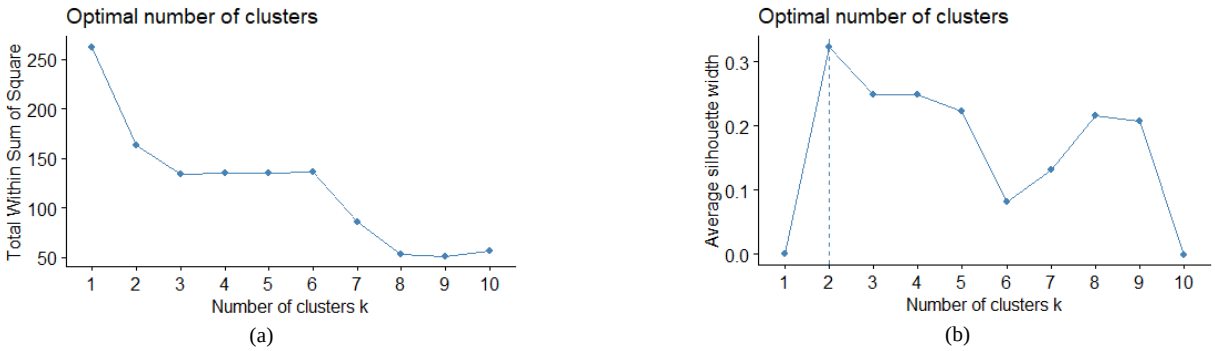
(a) (b)

Fig. 5. Determination of the optimal number of K-Medoids clusters (a) Elbow plot (b) Silhouette plot.

Based on Fig. 5., the optimal number of clusters formed is 10 clusters with the Silhouette method. This results in diverse cluster characteristics, making it difficult to interpret. However, the number of clusters of two is the second most recommended. Therefore, the number of clusters used for further analysis is two clusters. The evaluation results in the form of internal validation obtained connectivity coefficient, Dunn index, and Silhouette coefficient of 7.7012, 0.2126, and 0.3047, respectively.

Modelling with Fuzzy C-Means Clustering

The Fuzzy C-Means method applies fuzzy membership values to how much an object belongs to each existing cluster, in contrast to the hard clustering method that assigns objects strictly into one cluster. This algorithm considers each object as a member of each cluster with membership degrees varying between 0 and 1. The determination of the optimal number of clusters in this method is still the same as the previous methods, namely Elbow and Silhouette. Fig. 6. is a graph of the optimal number of clusters in Fuzzy C-Means.



(a) (b)

Fig. 6. Determination of the optimal number of Fuzzy C-Means clusters (a) Elbow plot (b) Silhouette plot.

Based on Fig. 6., the optimal number of clusters formed is two. Therefore, the number of clusters of two will be used for further analysis. The evaluation results in the form of internal validation show that the connectivity coefficient, Dunn index, and Silhouette coefficient are 8.9230, 0.1075, and 0.3226, respectively.

Model Evaluation Comparison

The evaluation of results in the form of internal validity was used to determine the best cluster analysis method based on the connectivity coefficient, Dunn index, and Silhouette coefficient. Table 5 displays the internal validity values of the three cluster analysis methods.

TABLE V. INTERNAL VALIDITY

Method	***Connectivity***	***Dunn***	***Silhouette***
K-Means	29.290	0.6931	0.4509
K-Medoids (PAM)	77.012	0.2126	0.3047
Fuzzy C-Means (FCM)	89.230	0.1075	0.3226

Based on Table 5 which shows the internal validity of the K-Means, K-Medoids (PAM), and Fuzzy C-Means (FCM) methods, it can be concluded that the K-Means method provides the best cluster results. This can be seen from the lowest connectivity coefficient (2.9290), indicating a more connected and compact cluster. In addition, the highest Dunn's index value (0.6931) indicates better cluster separation, and the highest Silhouette coefficient (0.4509) indicates that objects in a cluster are more similar to their own cluster than to other clusters. In contrast, the K-Medoids and Fuzzy C-Means methods performed less well with higher values in Connectivity and lower values in Dunn's index and Silhouette coefficient. Therefore, the K-Means method can be considered more effective in producing good clusters in the evaluated dataset. The equations are an exception to the prescribed specifications

Mapping the Incidence of Toddler Stunting in East Nusa Tenggara Province

After obtaining the best clustering method in the form of K-Means as many as two clusters, the next cluster profiling is done in the form of mapping based on indicators of stunting toddlers in East Nusa Tenggara Province. Fig. 7. is the result of mapping using the K-Means clustering method.

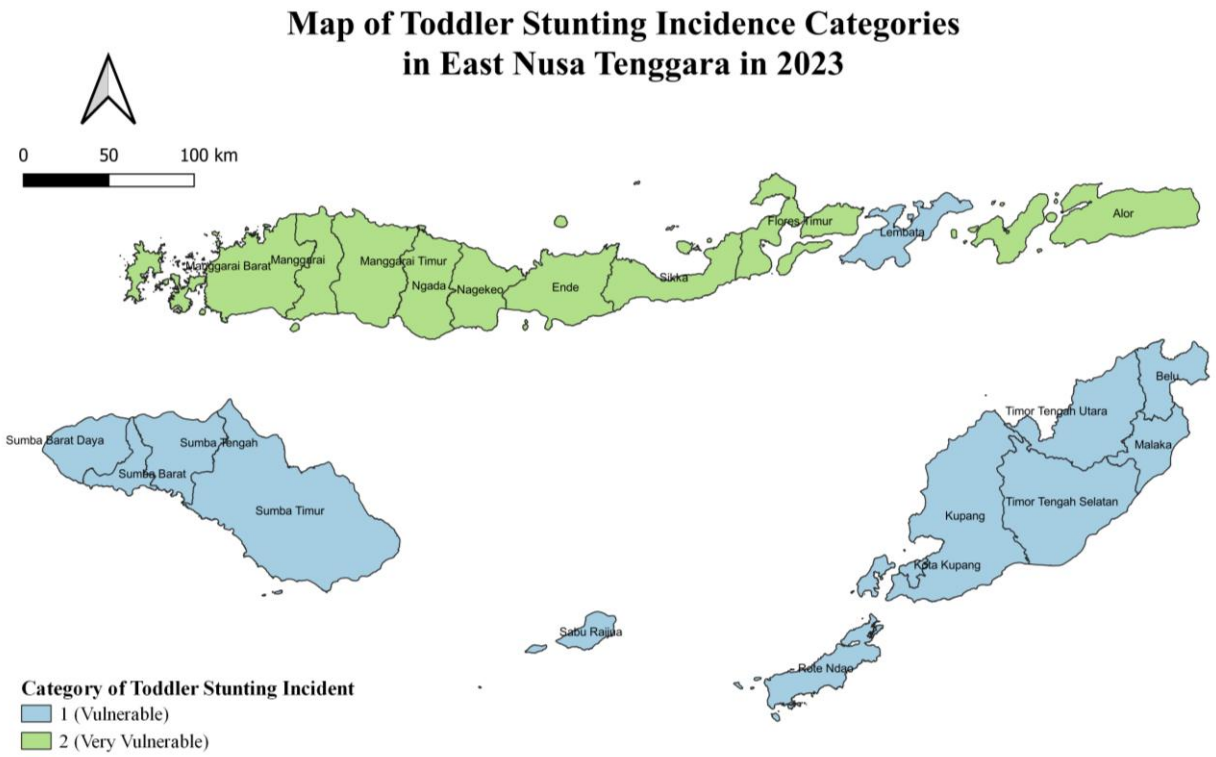


Fig. 7. Mapping of toddler stunting incidence categories in East Nusa Tenggara in 2023

Fig. 7. shows that the southern districts/cities in East Nusa Tenggara Province are vulnerable to stunting. Districts/cities in this cluster include Belu, Kota Kupang, Kupang, Lembata, Malaka, Sabu Raijua, Rote Ndao, Sumba Barat Daya, Sumba Barat, Sumba Timur, Sumba Tengah, Timor Tengah Utara, and Timor Tengah Selatan. Meanwhile, cluster two, which includes Flores Timur, Ende, Alor, Manggarai, Ngada, Nagekeo, Manggarai Timur, Manggarai Barat, and Sikka.

Based on the feature importance results from the regression model (Fig. 3.), the three most important features from official statistics include sma_p (the percentage of women aged 15 and over with a high school diploma or equivalent), rls_p (the average years of schooling for women), and wasting (the percentage of children under five who are wasted). Meanwhile, from remote sensing data, the key features include NDWI (Normalized Difference Water Index) and LST (Land Surface Temperature). Therefore, policies to address stunting should focus on education, maternal and child health, and environmental management tailored to the specific characteristics of each cluster.

TABLE VI. INDICATORS OF TODDLER STUNTING INCIDENTS USING THE K-MEANS METHOD

Indicators	Cluster 1	Cluster 2
wasting	7.93	9.25
imunisasi	69.57	64.14
sanitasi	79.66	70.71
asi_eks	96.90	94.53
air_minum	93.58	81.97
miskin	17.79	23.01
rls_p	7.68	7.45
sma_p	14.81	17.43
huruf_p	3.48	8.65
konsumsi	1995.00	1892.75
NDVI	0.76	0.63
NDWI	-0.84	-0.77
SAVI	0.46	0.37

NDBI	-0.28	-0.13
LST	27.99	31.07

Based on Table 6, districts/cities in cluster one tend to have a low percentage of children under five with malnutrition status (wasting), as well as higher levels of female education (sma_p and rls_p) than other clusters. This suggests that maternal education plays an important role in maintaining low stunting rates, as mothers with higher education tend to have a better understanding of nutrition, health, and parenting. Therefore, policies that can be implemented as a strengthening measure include optimising women's education programmes, with a focus on improving nutrition and reproductive health literacy at the secondary school level. In addition, continuous socialisation on parenting and balanced nutrition for pregnant women and mothers with children under five also needs to be expanded through various platforms, including public health services and digital technology. In terms of environment, the cluster has low NDWI and LST, indicating that the area has limited water availability and relatively lower land surface temperatures compared to other areas. Nonetheless, access to sanitation and drinking water in this area is decent, possibly due to the existence of basic infrastructure such as clean water networks and adequate sanitation facilities. This indicates that the government's efforts to provide clean water and sanitation have gone well, despite the region's natural terrain.

However, the districts in cluster two have characteristics that are the opposite of cluster one. This cluster has a high percentage of under-five wasting and a low quality of education for women, indicating the need for comprehensive interventions in health and education. Policies that can be implemented include improving access to education for women through scholarships and compulsory education programmes (education programmes in 3T areas), as well as strengthening maternal and child health services, including under-five nutrition monitoring and supplementary feeding for vulnerable groups. Environmentally, the region has high NDWI and LST values, indicating the availability of sufficient water sources, but also hot surface temperatures. This could be because although there are water sources, their distribution and utilisation may not be optimal, or low vegetation causes temperatures to remain high. Therefore, policies that can be implemented include better management of water resources, improved access to clean water and sanitation, and afforestation efforts to lower ambient temperatures and create healthier conditions for children.

IV. Limitation

This study has several key advantages, such as the utilization of official statistics data, which has been shown in previous study to significantly influence stunting prevalence, the integration with remote sensing data, which enables a more comprehensive analysis of environmental factors affecting stunting, and the implementation of a hybrid machine learning model to enhance prediction accuracy by combining statistical and remote sensing features. However, this study also has some limitations. The availability and resolution of remote sensing data may affect the granularity of the analysis, potentially limiting the ability to capture local variations in stunting prevalence. Additionally, while the hybrid machine learning model improves predictive performance, it requires extensive computational resources and careful parameter tuning. Lastly, the study relies on existing official statistics, which may be subject to reporting biases or data collection inconsistencies across different regions. Future study could address these limitations by incorporating higher-resolution satellite data, exploring alternative machine learning techniques, and validating the findings with field-based surveys.

V. Conclusion

This study demonstrates how a hybrid machine learning model, combining predictive modeling with cluster analysis, is used to evaluate the prevalence of stunting among children under five in East Nusa Tenggara Province. By integrating remote sensing data for environmental factors and official statistics for socio-economic factors, this study provides a more comprehensive understanding of the key determinants influencing stunting prevalence. The predictive modeling results indicate that XGBR is the best-performing model for estimating stunting prevalence in this region, with an RMSE of 3.2076 and an R^2 of 0.7223. Furthermore, the feature importance results from this model are utilized for subsequent cluster analysis. The best clustering method is obtained using K-Means, which forms two clusters: Cluster 1 (vulnerable) and Cluster 2 (highly vulnerable) to stunting, with connectivity, Dunn Index, and silhouette coefficient values of 29.290, 0.6931, and 0.4509, respectively. Future study can enhance this model by incorporating additional variables, such as distance to healthcare facilities, poverty levels, and more detailed environmental factors. Additionally, utilizing higher-resolution satellite imagery can improve accuracy in analyzing environmental influences on stunting. Overall, the findings of this study can serve as a foundation for data-driven policy making to address stunting in East Nusa Tenggara Province.

Acknowledgment

The author sincerely thanks Politeknik Statistika STIS for funding support in completing this study.

References

[1] G. R. Suraya and A. W. Wijayanto, “Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting,” *Indones. J. Stat. Its Appl.*, vol. 6, no. 2, pp. 180–201, 2022, doi: 10.29244/ijsa.v6i2p180-201.

[2] D. T. Saputra, “COMPARISON OF STUNTING CLUSTERS FOR EACH PROVINCE IN INDONESIA IN 2019 AND 2020 WITH 2021 AND 2022 USING THE K-MEANS METHOD,” *J. Indones. Sos. Teknol.*, vol. 5, no. 2, pp. 384–393, 2024, doi: 10.59141/jist.v5i2.920.

[3] M. Alam, “Perlu Terobosan dan Intervensi Tepat Sasaran Lintas Sektor untuk Atasi Stunting,” *Kemenko PMK*, 2023. https://www.kemenkopmk.go.id/perlu-terobosan-dan-intervensi-tepat-

sasaran-lintas-sektor-untuk-atasi-stunting#:~:text=Secara global%2C berdasarkan data UNICEF,diantara negara-negara di Asia. (accessed Jun. 08, 2024).

[4] S. Sufri, Nurhasanah, M. Jannah, T. P. Dewi, F. Sirasa, and S. Bakri, "Child Stunting Reduction in Aceh Province: Challenges and a Way Ahead," *Matern. Child Health J.*, vol. 27, no. 5, pp. 888–901, 2023, doi: 10.1007/s10995-023-03601-y.

[5] Kemenkes, *Survei Kesehatan Indonesia dalam Angka 2023*. Jakarta: Kementerian Kesehatan RI, 2023.

[6] T. Beal, A. Tumilowicz, A. Sutrisna, D. Izwardy, and L. M. Neufeld, "A review of child stunting determinants in Indonesia," *Matern. Child Nutr.*, vol. 14, no. 4, pp. 1–10, 2018, doi: 10.1111/mcn.12617.

[7] M. A. L. Suratri *et al.*, "Risk Factors for Stunting among Children under Five Years in the Province of East Nusa Tenggara (NTT), Indonesia," *Int. J. Environ. Res. Public Health*, vol. 20, no. 2, 2023, doi: 10.3390/ijerph20021640.

[8] A. Rahmawati, T. Nurmawati, and L. Permata Sari, "Faktor yang Berhubungan dengan Pengetahuan Orang Tua tentang Stunting pada Balita," *J. Ners dan Kebidanan (Journal Ners Midwifery)*, vol. 6, no. 3, pp. 389–395, 2019, doi: 10.26699/jnk.v6i3.art.p389-395.

[9] S. F. Nashriyah, M. R. Makful, and Y. P. Devi, "Gambaran Spasial Hubungan Antara Faktor Lingkungan Dan Ekonomi Dengan Stunting Balita Di Provinsi Nusa Tenggara Timur," *J. Spat. Wahana Komun. dan Inf. Geogr.*, vol. 23, no. 2, pp. 1–8, 2023.

[10] I. Sasongko, *Pengembangan Berkelanjutan: Penyediaan Infrastrukstur Pada Kawasan Pemukiman Secara Berkelanjutan*. Surabaya: PT Muara Karya (Anggota IKAPI), 2023.

[11] T. Welas, U. Jati, M. Sukin, and A. Ultanti, "Analisis Faktor-Faktor yang Memengaruhi Stunting di Provinsi Nusa Tenggara Timur Tahun 2019-2023," *J. Stat. Terap.*, vol. 4, no. 2, pp. 83–92, 2023.

[12] M. F. Ghazali, A. Aqzela, C. Gracia, R. S. Febriningtyas, and D. Wijayanti, "Analisis Geospasial Kasus Stunting menggunakan Artificial Neural Network (ANN) di Kecamatan Gadingrejo, Pringsewu-Lampung," *Maj. Geogr. Indones.*, vol. 37, no. 1, p. 1, 2022, doi: 10.22146/mgi.70474.

[13] P. K. Arya, K. Sur, T. Kundu, S. Dhote, and S. K. Singh, "Unveiling predictive factors for household-level stunting in India: A machine learning approach using NFHS-5 and satellite-driven data," *Nutrition*, vol. 132, p. 112674, 2025, doi: 10.1016/j.nut.2024.112674.

[14] S. Das, S. A. Basher, B. Baffour, P. Godwin, A. Richardson, and S. Rashid, "Improved estimates of child malnutrition trends in Bangladesh using remote-sensed data," *J. Popul. Econ.*, vol. 37, no. 4, pp. 1–37, 2024, doi: 10.1007/s00148-024-01043-6.

[15] A. Yusuf, "K-Means Clustering Based on Distance Measures: Stunting Prevalence Clustering in South Kalimantan," *2022 5th Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2022*, pp. 706–710, 2022, doi: 10.1109/ISRITI56927.2022.10052925.

[16] I. P. Sari, Al-Khowarizmi, O. K. Sulaiman, and D. Apdilah, "Implementation of Data Classification Using K-Means Algorithm in Clustering Stunting Cases," *J. Comput. Sci. Inf. Technol. Telecommun. Eng.*, vol. 4, no. 2, pp. 402–412, 2023, doi: 10.30596/jcositte.v4i2.15765.

[17] M. F. Lais, A. Atti, R. M. Pangaribuan, and R. D. Guntur, "Model Generalized Poisson Regression (GPR) Pada Kasus Stunting Di Provinsi Nusa Tenggara Timur," *J. Difer.*, vol. 5, no. 2, pp. 68–75, 2023, doi: 10.35508/jd.v5i2.11562.

[18] M. G. L. Bele, E. M. P. Hermanto, and F. Fitriani, "Pemodelan Geographically Weighted Regression pada Kasus Stunting di Provinsi Nusa Tenggara Timur Tahun 2020," *J. Stat. dan Apl.*, vol. 6, no. 2, pp. 179–191, 2022, doi: 10.21009/jsa.06204.

[19] I. K. Hasan, Nurwan, Nur Falaq, and Muhammad Rezky Friesta Payu, "Optimization Fuzzy Geographically Weighted Clustering with Gravitational Search Algorithm for Factors Analysis Associated with Stunting," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 1, pp. 120–128, 2023, doi: 10.29207/resti.v7i1.4508.

[20] L. M. Fauziah, F. N. Ihsan, R. Sisthannisa, L. V. Pangesti, T. S. Nugroho, and R. F. Putri, "Factors Affecting Stunting among Toddlers: A Case Study in West Nusa Tenggara Province," *E3S Web Conf.*, vol. 468, pp. 1–6, 2023, doi: 10.1051/e3sconf/202346806009.

[21] S. Candra Puspitasari, Sunarti, "Systematic Literature Review: Perlukah Pendekatan Spasial Dalam Penentuan Program Penanggulangan Stunting?," *J. Kesehat. Tambusai*, vol. 4, no. 3, pp. 3378–3393, 2023, [Online]. Available: http://journal.universitaspahlawan.ac.id/index.php/jkt/article/view/17911%0Ahttp://journal.universitaspahlawan.ac.id/index.php/jkt/article/download/17911/14061

[22] S. Pramana, R. Yordani, R. Kurniawan, and B. Yuniarto, *Dasar-Dasar Statistika Dengan Software R Konsep dan Aplikasi*. Bogor: IN MEDIA, 2016.

[23] S. Ndagijimana, I. H. Kabano, E. Masabo, and J. M. Ntaganda, "Prediction of Stunting among Under-5 Children in Rwanda Using Machine Learning Techniques," *J. Prev. Med. Public Heal.*, vol. 56, no. 1, pp. 41–49, 2023, doi: 10.3961/jpmph.22.388.

[24] B. Aurelia, P. Suhendi, S. W. Prayoga, and F. Kartiasih, "Comparison of ARIMA and Machine Learning Methods for Predicting Urban Land Surface Temperature in Jakarta," *J. Comput. Sci. Informatics Eng.*, vol. 8, no. 2, pp. 89–100, 2024, doi: 10.29303/jcosine.v8i2.579.

[25] S. W. Prayoga and A. W. Wijayanto, "Optimasi Prediksi Jumlah Wisatawan Nusantara ke Provinsi Bali Melalui Big Data Analytics dengan Integrasi Google Trends dan Tingkat Penghunian Kamar Hotel," *Semin. Nas. Off. Stat. 2024*, vol. 1, no. 1, pp. 571–580, 2024.

[26] M. Usman and K. Kopczewska, "Spatial and Machine Learning Approach to Model Childhood Stunting in Pakistan: Role of Socio-Economic and Environmental Factors," *Int. J. Environ. Res. Public Health*, vol. 19, no. 17, 2022, doi: 10.3390/ijerph191710967.

[27] A. Asra, A. P. Utomo, M. Asikin, and N. H. Pusponegoro, *Analisis Multivariabel Suatu Pengantar*. Bogor: In Media, 2017.

[28] S. Pramana, B. Yuniarto, I. Santoso, R. Nooraeni, and L. H. Suadaa, *Data Mining dengan R: Konsep dan Implementasi*, Edisi 2. Bogor: IN MEDIA, 2023.

[29] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed. USA: Elsevier Inc., 2012.

[30] I. H. Pamungkas and S. Pramana, "IMPROVEMENT METHOD OF FUZZY GEOGRAPHICALLY WEIGHTED CLUSTERING USING GRAVITATIONAL SEARCH ALGORITHM," *J. Ilmu Komput. dan Inf.*, vol. 11, no. 1, pp. 10–16, 2018.