Unlocking potential of data: A localised data-driven approach for stunting reduction in South Kalimantan, Indonesia

F Rizkiah[1,*]

[1] Auditor, Finance and Development Supervisory Board, South Kalimantan, Indonesia

* Corresponding author's email: farahxrizkiah@gmail.com

Abstract. Stunting is a serious public health issue in Indonesia. It can affect children's physical growth, cognitive abilities, and long-term socioeconomic outcomes. The prevalence rate of stunting in South Kalimantan is higher than the national average, so targeted interventions are essential. This study takes a three-stage data-driven approach to understand the determinants of stunting, develop a predictive model for forecasting prevalence rate, and cluster districts/cities based on their APS trends. First, using Pooled OLS and Fixed Effect Regression models, we identified significant determinants, including the human development index, healthcare service availability, and living conditions. Second, we evaluated three machine learning algorithms — Linear Regression, Random Forest, and Support Vector Regression — and found that Linear Regression was the best algorithm for forecasting prevalence rate with an R-squared score of 0.756, Mean Absolute Error of 0.027, and Mean Squared Error of 0.001. Lastly, we used the K-means algorithm to cluster districts/cities into four distinct groups based on their prevalence rate trends, which can help allocate resources effectively for targeted interventions. This study shows the importance of using localised data to formulate effective strategies for combating stunting in South Kalimantan. By using a localised data-driven approach, this study aligns with the broader aim of reducing stunting prevalence and can help enhance public health infrastructure in South Kalimantan, Indonesia.

1. Introduction

Stunting is a condition where children fail to grow due to chronic malnutrition over an extended period, especially during the first 1.000 days of a child's life [1]. This condition is caused by various factors, including nutritional factors such as poor nutritional status in pregnant women and inadequate nutritional intake in children; environmental factors such as poor sanitation, lack of access to clean water, and limited health service accessibility; and social factors like the level of education and the low socioeconomic status of families [2, 3].

If not properly addressed, stunting can adversely affect children's health, such as inhibited physical growth, reduced cognitive abilities, and decreased immunity. Moreover, stunting can also lead to significant social and economic impacts, such as reduced productivity and quality of life, which, in the long run, can hinder economic development, increase poverty rates, and widen the gap between societal classes [4].

The 2022 Nutritional Status Survey of Indonesia (SSGI) indicated that Indonesia has a stunting prevalence rate (APS) of 21,6%. This means that about 1 in 5 children in Indonesia suffer from stunting. This rate is still above the threshold set by the World Health Organisation (WHO), which is 20% [5]. South Kalimantan Province is one of the national priority areas for stunting control. In 2021, the APS for South Kalimantan, based on SSGI, was 30% and ranked sixth highest (worst) in Indonesia. This improved in 2022 when the APS decreased to 24,6%, but it remains above the national average. The three districts/cities with the highest APS in 2022 were Barito Kuala District (33,6%), Kotabaru District (31,6%), and Hulu Sungai Tengah District (31,1%) [6].

According to a government report, one of the main issues causing the high APS in South Kalimantan is the misidentification of locus and types of stunting interventions [7, 8, 9]. The selection of intervention sites for the Community-Based Infrastructure Program (IBM) does not align with regions with the propensity of high APS. In addition, the stunting interventions conducted do not prioritise sensitive interventions that have a substantial impact on reducing the prevalence of stunted children.

The lack of alignment in determining where and what interventions to implement may hinder the achievement of the goal for accelerated reduction in stunting. Therefore, more targeted efforts are needed in setting the locus and type of stunting interventions in South Kalimantan. One way to accomplish this is through a data-driven approach, which involves mapping and analysing the factors causing stunting and identifying regions with high APS tendencies.

This research aims to assist the local government in South Kalimantan Province in developing effective and sustainable strategies for addressing stunting, taking into account the aforementioned aspects. In this study, we analyse the social, nutritional, and environmental factors that might influence stunting prevalence. We also employ machine learning methods to build a prediction model that can be used to forecast the APS of districts/cities. Additionally, we conduct clustering analysis to group districts/cities based on their APS propensity.

This study distinguishes itself by analysing a more comprehensive set of variables encompassing various factors causing stunting, including social factors, health services availability, nutritional factors, and environmental factors. In addition, we not only examine the factors influencing stunting prevalence but also develop a prediction model for APS forecasting. This prediction model can aid local governments in tailoring their stunting intervention programmes to specific regions based on the model's prediction and ensures that interventions are relevant to the local context. Furthermore, we conduct clustering analysis based on APS trends, allowing local governments to strategically allocate resources and interventions to areas with consistently high or increasing stunting rates and not just base their decision on the previous year's data. Lastly, the use of machine learning techniques in this study, both supervised and unsupervised learning, provides insights into the patterns and characteristics of stunting data and can be used to enhance government strategies for stunting reduction and prevention in the future.

2. Literature Review

Several previous studies related to the factors causing stunting have produced varied results. Torlesse et al. stated that social factors, such as poverty and uneven access to health services, could underlie stunting [10]. Furthermore, Febrina et al. identified nutritional factors such as low birth weight in infants, chronic energy deficiency in pregnant women, and insufficient complementary feeding alongside breastfeeding as contributors to stunting [11]. Meanwhile, studies on environmental factors such as sanitation and clean water access have produced differing conclusions [12, 13]. Drawing on these insights, this study investigates the roles of social, governmental, nutritional, and environmental factors in stunting prevalence within districts and cities by utilising a wide array of variables. The social factors include human development index (HDI) and poverty rates; the governmental factors measure the availability of healthcare facilities for the general public, including hospitals, clinics, community health centre (Puskesmas) and integrated health post (Posyandu) that represent the government effort in improving healthcare services; the nutritional factors include toddler's immunisation, exclusive breastfeeding, pregnant women with chronic energy deficiency, antenatal care (K4) visit, low birth weight, and children malnutrition; the environmental factors include children diarrhea cases, access to proper sanitation, access to suitable drinking water, and slums area.

Previous studies also have harnessed machine learning methodologies to predict APS in Indonesia. Mambang et al. employed the Linear Regression algorithm, using APS data from districts and cities spanning 2013-2020, to predict stunting prevalence in South Kalimantan [15]. Haris et al. collected data from five national surveys conducted in 2019 and used the Random Forest algorithm to predict APS in East Java [16]. Another study by Khudori et al. (2023) compared several supervised learning algorithms for APS prediction in East Java and found that the Support Vector Regression algorithm yielded the highest accuracy [17]. In this research, we utilise stunting data from districts and cities and employ three algorithms, namely Linear Regression, Random Forest, and Support Vector Regression, to build a prediction model for APS in districts/cities.

Furthermore, previous research on clustering analysis for stunting has also been conducted. Chandra et al. applied the K-means algorithm to malnourished toddlers by province, resulting in two distinct groups (clusters) characterised by varying levels of malnutrition severity [18]. In a separate study conducted by Anggraeni et al., the agglomerative hierarchical clustering method was used to categorise villages in the Tegalrejo sub-district, Yogyakarta, based on APS, resulting in three stunting prevalence levels (high, medium, and low) [19]. In this research, we utilise the K-means clustering algorithm to group districts/cities based on APS due to its computational efficiency and ease of interpretation.

This research endeavours to synergise insights from previous findings while adding several new variables previously unutilised in the social, governmental, nutritional, and environmental determinants while leveraging cutting-edge machine learning and clustering algorithms. The aim is to harness the potential of localised data in formulating strategic interventions for stunting reduction in South Kalimantan Province.

3. Methodology

3.1. Research Design

This study incorporates a three-stage research design, including analysing stunting determinant, building stunting prevalence prediction model, and conducting clustering analysis based on APS trend.

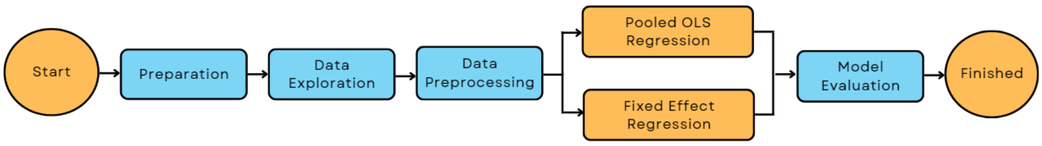

Figure 1. First stage - stunting prevalence determinant analysis.

The first stage involves analysing the factors that significantly influence stunting prevalence in districts/cities. In this initial stage, we are using two regression methodologies, Pooled Ordinary Least Squares (Pooled OLS) Regression and Fixed Effect (FE) Regression, to identify the significant factors that influence stunting prevalence. Pooled OLS Regression is suitable for this analysis as it helps to evaluate the impact of independent variables on the dependent variable across different cross-sections, while assuming a constant relationship across the different variables. On the other hand, the Fixed Effect Regression is used to account for any unobserved heterogeneity by acknowledging that each district/city may have inherent attributes that could affect stunting prevalence. The combination of Pooled OLS and Fixed Effect Regression provide a robust analytical framework to comprehensively understand the determinants of stunting prevalence both across and within the districts/cities.

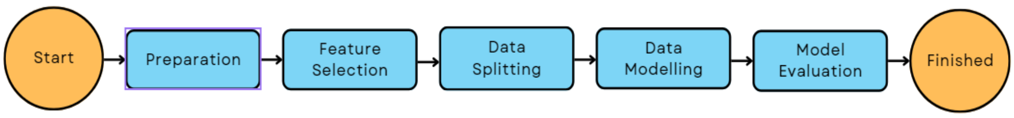

Figure 2. Second stage - building stunting prevalence rates prediction model.

In the second stage, we utilise three machine learning algorithms: Linear Regression, Random Forest, and Support Vector Regression. Each algorithm provides a unique way of understanding the data and making predictions. Linear Regression provides a simple and easy-to-understand model by assuming a linear relationship between independent and dependent variables. Random Forest is an ensemble method that combines the predictions from multiple decision trees to create a more robust and accurate model. On the other hand, Support Vector Regression works well for small to medium-sized datasets and is proficient in high-dimensional spaces.

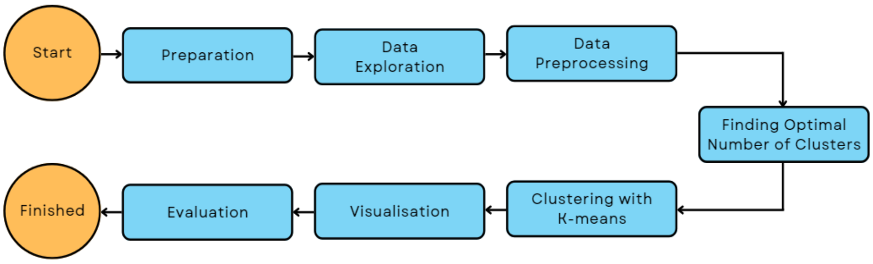

Figure 3. Third stage - clustering analysis based on APS trend.

In the final stage, we use clustering analysis to classify districts or cities based on their APS in order to understand stunting prevalence trends across different regions. We employ the K-means clustering method, which partitions a data set into distinct group of districts or cities that exhibit similar APS trends.

3.2. Data Collection

In this research, we utilised two datasets. Dataset 1 is used to analyze the determinants of stunting and predict APS. It contains social, governmental, nutritional, and environmental factors that could affect stunting prevalence, as well as APS data for districts/cities from 2019-2022. Meanwhile, Dataset 2 is used for clustering analysis. It includes historical APS data for each district/city from 2013-2022. Both datasets are secondary data that are collected from various sources as follows:

Table 1. Data descriptions and sources

No	Variable	Description	Source
1	hdi	Human Development Index (HDI) by district/city in South Kalimantan (2019-2022)	Central Bureau of Statistics (BPS) Website
2	pov	Percentage of the poor population by district/city in South Kalimantan (2019-2022)	Central Bureau of Statistics (BPS) Website
3	health_facil	Number of government healthcare facilities at level 1-3 by district/city in South Kalimantan (2019-2022)	Central Bureau of Statistics (BPS) Website
4	village_health	Number of villages with healthcare facilities by district/city in South Kalimantan (2019-2022)	Central Bureau of Statistics (BPS) Website
5	active_posyandu	Percentage of active Integrated Health Posts (Posyandu) by district/city in South Kalimantan (2019-2022)	Performance Accountability Report of the South Kalimantan Provincial Health Office 2019-2022
6	immun	Percentage of toddlers aged 0-59 months who received complete immunizations by district/city in South Kalimantan (2019-2022)	Welfare Statistics of South Kalimantan Province 2019-2022
7	breastfeed	Percentage of infants aged 0-6 months who receive exclusive breastfeeding by district/city in South Kalimantan (2019-2022)	Performance Report of the Public Health Division of the South Kalimantan Provincial Health Department 2019-2022

No	Variable	Description	Source
8	energy_def	Percentage of pregnant women experiencing chronic energy deficiency by district/city in South Kalimantan (2019-2022)	District/City in Figures 2023
9	k4_visit	Percentage of pregnant women undergoing antenatal care K4 visits (a minimum of 4 contacts during pregnancy) by district/city in South Kalimantan (2019-2022)	District/City in Figures 2023
10	low_birth	Number of babies with low birth weight by district/city in South Kalimantan (2019-2022)	South Kalimantan Province in Figures 2020-2023
11	malnutrition	Percentage of children experiencing malnutrition by district/city in South Kalimantan (2019-2022)	Central Bureau of Statistics (BPS) Website
12	diarrhea	Number of diarrhea cases in children by district/city in South Kalimantan (2019-2022)	Central Bureau of Statistics (BPS) Website
13	sanitation	Percentage of households with access to proper sanitation by district/city in South Kalimantan (2019-2022)	Central Bureau of Statistics (BPS) Website
14	water	Percentage of households with access to suitable drinking water services by district/city in South Kalimantan (2019-2022)	South Kalimantan Province in Figures 2020-2023
15	slum	Number of uninhabitable homes by district/city in South Kalimantan (2019-2022)	Satu Data Banua
16	stunting	Stunting Prevalence Rate (APS) in South Kalimantan Province by district/city from 2013-2022	Satu Data Banua (2013-2018), Indonesian Child Nutrition Status Study (SSGBI) (2019), Community-Based Nutrition Electronic Registration and Reporting (e-PPGBM) (2020), SSGI (2021-2022)

3.3. Panel Data Regression Analysis

Panel data is a type of statistical data that consists of observations taken from different analytical units over a specified period. This approach provides a comprehensive framework to explore the dynamics that influence stunting prevalence in different districts or cities. To identify the factors that have a significant impact on stunting prevalence, we utilized two robust panel data regression methods: Pooled Ordinary Least Squares (OLS) Regression and Fixed Effect (FE) Regression.

3.3.1. Pooled OLS Regression. Pooled OLS Regression is one of the regression analysis techniques that can be used to examine how independent variables affect the dependent variable in a panel data structure. This method combines data from different cross-sections and treats it as one dataset, assuming a constant relationship across entities over time. The technique produces regression coefficient estimates, which indicate the magnitude of influence exerted by the independent variables on the dependent variable, and assesses the statistical validity of these relationships through significance testing. The model is presented as follows:

$$Y_{it} = \alpha + \beta X_{it} + u_{it}$$

Where:

Y_{it} = dependent variable (stunting prevalence) for a particular entity at a specific time

X_{it} = matrix of independent variables

α = constant term

β = vector of coefficients to be estimated

u_{it} = error term

3.3.2. Fixed Effect Regression. Fixed Effect Regression is a statistical method that takes into account the unobserved heterogeneity among different entities, such as districts or cities, by adding entity-specific intercepts. This approach is different from Pooled OLS, which assumes that all entities have the same characteristics. In contrast, Fixed Effect Regression considers that each entity may have inherent, unobservable attributes that could potentially affect the outcome being studied, such as stunting prevalence. By controlling for these unobserved, entity-specific effects, this method provides a more nuanced understanding of the relationship between the independent variables and the dependent variable within each entity. The Fixed Effect model can be expressed as follows:

$$Y_{it} = \alpha_i + \beta X_{it} + u_{it}$$

Where: α_i = the entity-specific intercepts

3.4. Predictive Modeling

The APS predictive modelling represents a supervised machine learning problem, where the data used already possesses target values that are employed to train (supervise) the algorithm to make predictions. Initially, a data preparation process is executed by identifying and selecting the variables to be incorporated as features in the model. Subsequently, the data is divided into training data and testing data in certain proportions. The model is then trained on the training data using each predetermined algorithm. Following this, an evaluation is carried out on each model by calculating the R-square value, Mean Absolute Error (MAE), and Mean Square Error (MSE). From the three models, the one with the best performance is selected to predict the stunting prevalence rate on new data.

In training the prediction model, we utilise three machine learning algorithms, namely Linear Regression, Random Forest, and Support Vector Regression.

3.4.1. Linear Regression. Linear Regression operates on the principle of establishing a linear relationship between the dependent and independent variables. Mathematically, linear regression can be represented as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Where:
Y = dependent variable
X_n = independent variables
β_0 = y-intercept
β_n = coefficients
ϵ = error term

3.4.2. Random Forest. Random Forest is an ensemble learning method that merges the predictions from multiple decision trees to produce a more accurate and robust result. It capitalises on bootstrapping and aggregation to enhance the model's generalizability and reduce overfitting. Random Forest integrates multiple trees *from the decision tree model*, *which is* represented as:

$$RF = \{Tree_1, Tree_2, \dots, Tree_n\}$$

Where each tree yields a prediction, and the mode of prediction is the final output.

3.4.3. Support Vector Regression (SVR). SVR focuses on finding a hyperplane that best fits the data *to predict* a continuous value. The primary objective of SVR is to ensure that errors do not exceed the threshold. Mathematically, SVR can be simplified as:

$$y = \sum_{i=1}^{n} \alpha_i \cdot k(x_i, x) + b$$

Where:
y = predicted output
x_i, x = data vectors

α_i = Lagrange multipliers
k = kernel function
b = bias term

3.5. Clustering Analysis

Clustering analysis, an unsupervised machine learning technique, groups data into subsets based on their similarity, ensuring that data points in the same group are more alike to each other than to those in other groups. In this study, we employed the K-means clustering method, which partitions data into 'K' distinct, non-overlapping subsets (or clusters). Mathematically, the objective of K-means is to minimise the within-cluster sum of squares. The objective function can be represented as:

$$\boldsymbol{J} = \sum_{i=1}^{K} \sum\nolimits_{x \in C_i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2$$

Where: K = number of clusters
$\boldsymbol{C_i}$ = $\boldsymbol{i}$-th cluster
$\boldsymbol{x}$ = a data point in $\boldsymbol{C_i}$
$\boldsymbol{\mu_i}$ = centroid of $\boldsymbol{C_i}$

The K-means algorithm iteratively assigns each data point to the closest centroid and then recalculates the centroid based on the mean of the data points in the cluster. This process continues until the centroids no longer change significantly.

To determine the optimal number of clusters, we combined the elbow method and silhouette analysis. The elbow method involves plotting the explained variation as a function of the number of clusters and picking the "elbow" of the curve as the number of clusters to use. The silhouette analysis gauges how similar an object is to its own cluster compared to other clusters. The silhouette coefficient ranges from -1 to 1, where a high value indicates that the object is well-matched to its cluster and poorly matched to neighbouring clusters. If most objects have a high value, then the clustering configuration is deemed appropriate.

4. Results and Discussion

4.1. Descriptive Analysis

The descriptive statistics section provides an overview of the main characteristics of the dataset.

Table 2. Descriptive statistics of Dataset 1

	count	mean	std	min	25%	50%	75%	max
hdi	52	0,71	0,04	0,66	0,69	0,70	0,71	0,80
pov	52	0,05	0,01	0,03	0,04	0,05	0,06	0,07
health_facil	52	344,62	108,82	210,00	250,00	331,50	392,00	616,00
village_health	52	73,14	27,01	27,00	55,75	65,00	86,25	141,00
active_posyandu	52	0,60	0,23	0,04	0,49	0,60	0,77	1,00
immun	52	0,61	0,10	0,39	0,54	0,62	0,68	0,77
breastfeed	52	0,63	0,13	0,19	0,61	0,66	0,70	0,88
energy_def	52	0,80	0,09	0,58	0,75	0,81	0,86	0,97
k4_visit	52	0,14	0,04	0,08	0,12	0,14	0,16	0,29
low_birth_weight	52	0,06	0,02	0,00	0,04	0,06	0,08	0,11
malnutrition	52	0,28	0,19	0,03	0,13	0,21	0,39	0,91
diarrhea	52	4.330,75	3.577,01	151,00	1.899,25	3.287,00	5.696,50	19.155,00
sanitation	52	0,81	0,11	0,47	0,77	0,82	0,89	0,96
water	52	0,72	0,13	0,49	0,63	0,70	0,83	1.00
slum	52	3.782,35	3.273,51	439,00	2.013,00	3.205,00	4.847,50	21.249,00
stunting	52	0,14	0,06	0,03	0,09	0,14	0,18	0,31

Dataset 1 covers the period from 2019 to 2022 and includes various districts and cities, providing a comprehensive view of the factors that affect child stunting. The average Human Development Index is 70.54, while the poverty rate is 7.28%. The healthcare infrastructure is represented by an average of 297.75 health facilities per district/city. However, the standard deviation of 311.22 shows significant regional disparities. On average, 91.04% of the active integrated health posts are covered, and the immunization rate is 87.71%, indicating a high standard of public health. Breastfeeding is widespread, with an average of 11.60 months, which contributes positively to child health. However, the proportion of stunted children under five is 14.65% on average, with a standard deviation of 7.17%, highlighting the persistent nutritional challenges. The dataset underscores significant variability across regions, necessitating targeted interventions for holistic development and child health improvement.

Table 3. Descriptive statistics of Dataset 2

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Balangan	8	0,356	0,106	0,158	0,317	0,358	0,409	0,521
Banjar	8	0,338	0,080	0,261	0,268	0,318	0,393	0,482
Banjarbaru	8	0,268	0,073	0,190	0,214	0,249	0,303	0,398
Banjarmasin	8	0,298	0,043	0,224	0,275	0,310	0,325	0,358
Barito Kuala	8	0,368	0,048	0,324	0,340	0,359	0,370	0,476
Hulu Sungai Selatan	8	0,366	0,102	0,203	0,290	0,375	0,436	0,500
Hulu Sungai Tengah	8	0,357	0,076	0,273	0,307	0,345	0,380	0,517
Hulu Sungai Utara	8	0,400	0,121	0,209	0,345	0,386	0,494	0,561
Kotabaru	8	0,328	0,109	0,167	0,268	0,322	0,390	0,478
Tabalong	8	0,363	0,090	0,197	0,316	0,379	0,445	0,446
Tanah Bumbu	8	0,260	0,094	0,161	0,185	0,250	0,293	0,449
Tanah Laut	8	0,338	0,055	0,266	0,300	0,334	0,388	0,407
Tapin	8	0,348	0,099	0,145	0,324	0,359	0,397	0,457

Dataset 2 includes historical APS data for each district/city from 2013-2022. The panel data reveals a range of conditions across various districts/cities, with mean stunting rates varying from 25,97% in Tanah Bumbu to 40,04% in Hulu Sungai Utara. The districts exhibit differing degrees of variability, as indicated by the standard deviations, with more stable conditions in Banjar (8,02%) and heightened variability in Hulu Sungai Utara (12,14%).

In 2022, the three districts/cities with the highest APS were Barito Kuala District (33.60%), Kotabaru District (31.60%), and Hulu Sungai Tengah District (31.10%), while the three districts/cities with the lowest APS were Tapin District (14.50%), Tanah Bumbu District (16.10%), and Tabalong District (19.70%).

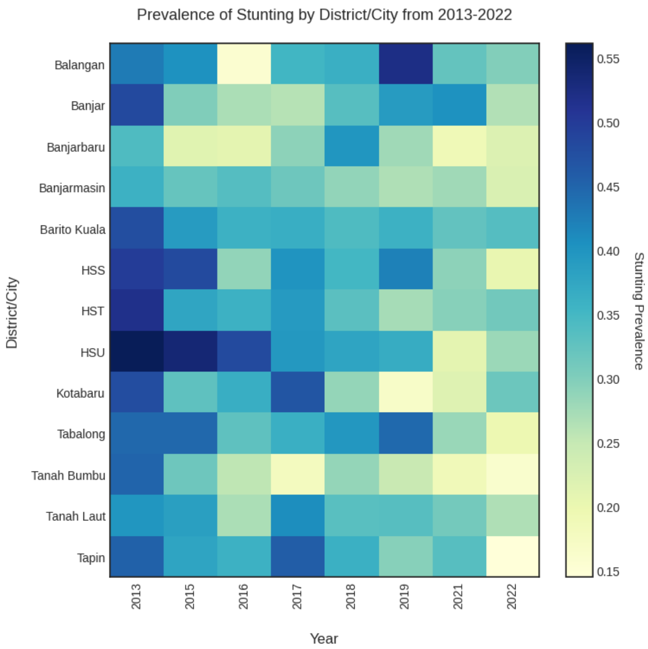

Figure 4. Heatmap of stunting prevalence in South Kalimantan from 2013-2022.

The trend graph of stunting prevalence for each district/city from 2013-2022 is presented below:

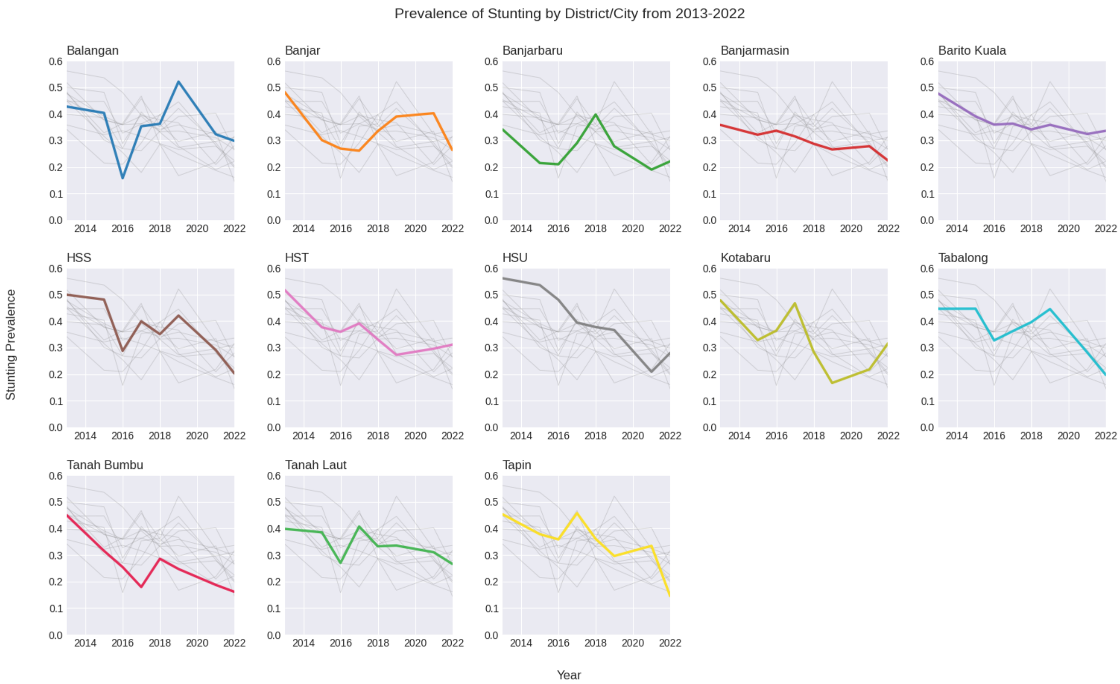

Figure 4. Stunting prevalence according to districts/cities in South Kalimantan from 2013-2022.

Throughout 2013-2022, there was a general fluctuation in APS across all districts/cities in South Kalimantan. The average APS during this period was 33,75%, with values ranging from 14,50% to 56,10%.

4.2. Analysis of Determinants of Stunting Prevalence

The results of the panel data regression analysis are presented below.

Table 4. Pooled OLS Regression and Fixed Effect Regression Result

Variables	Pooled OLS Regression				Fixed Effect Regression			
	Coefficient	Std. Err.	t-stat	p-value	Coefficient	Std. Err.	t-stat	p-value
const	1,098	0,282	3,897	0,000***	3,868	0,835	4,634	0,000***
hdi	-1,043	0,378	-2,761	0,009**	-4,609	1,302	-3,539	0,002**
pov	2,570	1,306	1,968	0,057	-0,568	2,133	-0,266	0,792
health_facil	-0,000	0,000	-0,634	0,530	0,000	0,000	0,636	0,531
village_health	-0,000	0,000	-0,118	0,906	0,000	0,000	0,282	0,780
active_posyandu	-0,101	0,043	-2,357	0,024*	0,021	0,026	0,806	0,428
immun	-0,409	0,119	-3,436	0,002**	-0,002	0,090	-0,217	0,830
breastfeed	-0,045	0,075	-0,593	0,557	-0,001	0,042	-0,228	0,821
energy_def	-0,301	0,293	-1,026	0,312	-0,331	0,180	-1,833	0,079
k4_visit	0,285	0,126	2,261	0,030*	-0,128	0,082	-1,560	0,132
low_birth_weight	-0,070	0,468	-0,149	0,883	0,127	0,338	0,376	0,710
malnutrition	0,121	0,058	2,084	0,044*	-0,040	0,036	-1,132	0,269
diarrhea	0,000	0,000	0,191	0,849	0,000	0,000	0,991	0,332
sanitation	-0,107	0,113	-0,949	0,349	-0,531	0,138	-3,857	0,001***
water	-0,114	0,091	-1,248	0,220	0,062	0,102	0,612	0,546
slum	-0,000	0,000	-2,525	0,016**	0,000	0,000	0,806	0,428
Obs.		52			52			
R-squared		0,536			0,804			
Adj. R-squared		0,342			0,804			
F-statistic		2,770			6,5423			
Prob (F-statistic)		0,006			0,000			

The Pooled OLS regression model explained around 53,60% of the variation in stunting prevalence. The F-stat test result of 2,770 with Prob (F-stat) < 0.05 indicates that the Pooled OLS Regression model is statistically significant in explaining the variation in the dependent variable. In assessing the model's fitness, various tests were performed: the Jarque-Bera test confirmed normally distributed residuals; heteroscedasticity has not been identified by the Breusch-Pagan test; no autocorrelation is found via the Durbin-Watson test; and a VIF value under 10 negated multicollinearity concerns.

The Fixed Effects regression model showed a significant increase in the explanatory power of our variables, resulting in an R-squared value of 0.804. This indicates that around 80.4% of the differences in stunting prevalence between districts and cities can be explained by the model. The F-stat result, which is 6.5423 and Prob (F-statistic): 0.000 confirms the significance of the model. This model is useful in accounting for unobserved heterogeneity across districts and cities when studying stunting determinants.

Both the Pooled OLS Regression and Fixed Effect Regression model showed a significant and negative relationship with the human development index ('hdi'), with the Pooled OLS Regression also showed a significant and negative relationship with percentage of active integrated health posts ('active_posyandu'), toddlers immunisations rate ('immun') and the number uninhabitable homes ('slum'). The Fixed Effect regression model also showed a significant negative relationship with percentage of households with access to proper sanitation ('sanitation'). These findings highlight the crucial role of holistic human development, availability of healthcare services and access to proper living environment in reducing stunting. We also find a significant and positive relationship of stunting prevalence with antenatal care k4 visit ('k4_visit') and children malnutrition cases ('malnutrition'). One possible explanation is that the data might capture regions where health issues are more severe, hence both stunting and malnutrition cases are higher. Similarly, the higher frequency of antenatal care K4 visits could be a response to these adverse health conditions.

On the other hand, the poverty rate ('pov') did not have a significant effect in both models. This suggests that we need to look more closely at the different dimensions of poverty and how they affect child nutrition. Similarly, the number of healthcare facilities ('faskes') and the number of villages with healthcare facilities ('desa_faskes'), did not have a strong predictive power in this context. This may indicate gaps in healthcare service delivery or quality that need further investigation. Furthermore, variables related to breastfeeding practice ('breastfeed'), chronic energy deficiency in pregnant women ('energy_def'), instances of low birth weight ('low_birth_weight'), diarrhea prevalence ('diarrhea'), and access to drinking water ('water') did not have a significant effect in both models. This raises questions about their roles and other factors that could be affecting their relationships.

4.1 Predictive Modeling of Stunting Prevalence Rate

Utilising the district and cities stunting data from previous analysis, we build a predictive model for Stunting Prevalence Rate (APS). The goal is to create a reliable and accurate predictive framework for forecasting stunting prevalence. We evaluated the performance of three machine learning algorithms: Linear Regression, Random Forest, and Support Vector Regression.

We compared the efficacy and accuracy of each algorithm in predicting the APS using key performance metrics such as R-squared score, Mean Absolute Error (MAE), and Mean Squared Error (MSE). The results are as follows:

Table 5. Comparison of Model Key Performance Metrics

Algorithm	MAE	MSE	R-squared
Linear Regression	0,027	0,001	0,756
Random Forest	0,034	0,001	0,612
Support Vector Regression	0,042	0,002	0,492

The R-squared score is a key metric that indicates the amount of variability in the target variable (APS) that can be explained by the features variables. A higher R-squared score indicates a better fit of the model to the data. In this case, Linear Regression performed the best among the three algorithms, with an R-squared score of 0.756.

The Mean Absolute Error (MAE) provides insight into the average size of the errors between the predicted and actual values, irrespective of the direction. A lower MAE score indicates better predictive accuracy. Once again, the Linear Regression algorithm outperformed the other two algorithms with the lowest MAE of 0.027.

Lastly, the Mean Squared Error (MSE) is a metric that is sensitive to outliers and highlights the average of the squares of the errors. A lower MSE score indicates a better fit of the model. In this instance, both Linear Regression and Random Forest algorithms exhibited an equal MSE score of 0.001, which was better than the Support Vector Machine.

Based on these results, it can be concluded that among the three tested algorithms, the Linear Regression algorithm offers promising potential for accurately forecasting the Stunting Prevalence Rate (APS). The prediction results comparison of the three model algorithms are presented as follows:

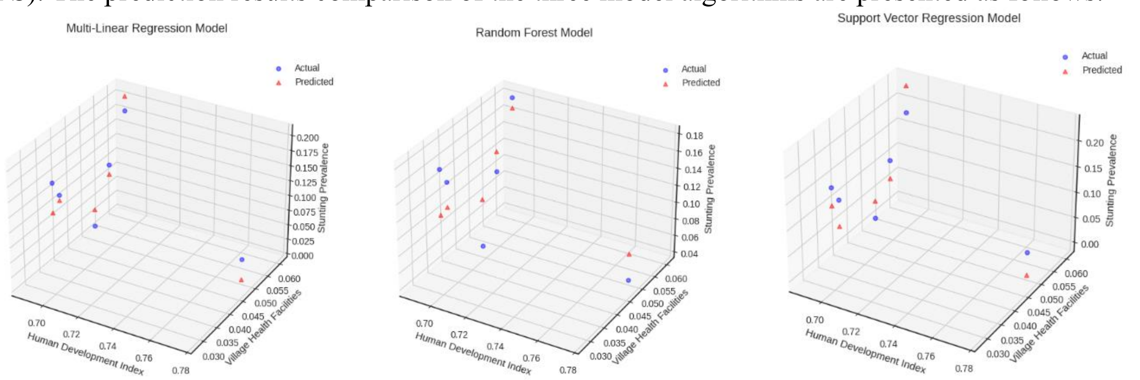

Figure 5. Comparison of prediction results from the three algorithms

This prediction model can be useful for local governments to design targeted stunting intervention programmes based on the model's predictions. By tailoring interventions to specific regions, local governments can ensure that the interventions are relevant and effective in the local context.

4.3. Clustering Analysis of APS for Districts/Cities

Clustering analysis is a useful tool for identifying natural groupings within data based on specific criteria. In this study, we used the elbow method and the silhouette method to determine the optimal number of clusters.

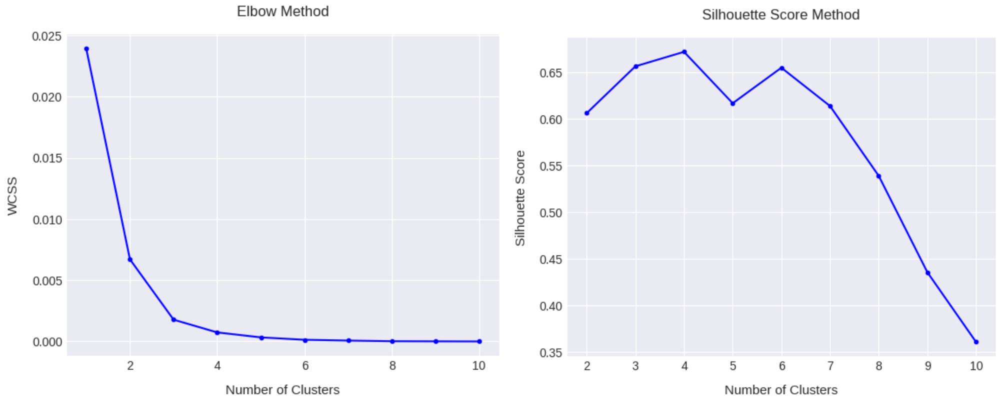

Figure 6. Determining the optimal number of clusters

The elbow method revealed a significant decrease in the inertia value at cluster counts of 3 and 4, indicating a potential fit. At the same time, the silhouette method indicated that a cluster count of 4 was the most appropriate, as it had the highest silhouette score. After considering the results of both methods, we concluded that 4 was the optimal number of clusters for our analysis.

After determining the optimal number of clusters, clustering analysis was conducted using the K-means algorithm. The resulting analysis identified four distinct clusters based on the average APS (stunting prevalence rate) levels, as follows:

Table 6. Clustering analysis results

Cluster	District/City
1	Banjarbaru Tanah Bumbu
2	Banjarmasin Banjar Kotabaru Tanah Laut
3	Hulu Sungai Tengah Balangan Barito Kuala Tapin
4	Hulu Sungai Selatan Tabalong Hulu Sungai Utara

Table 7. Evaluation results of clustering performance

Cluster	Average stunting prevalence rate	Sum of Squared Error (SSE)
1	0,250	0,000
2	0,319	0,000
3	0,355	0,000
4	0,380	0,000

The analysis resulted in four clusters based on the average APS (stunting prevalence rate). Cluster 1 consists of two districts/cities with a relatively low average APS of 25.00%. Cluster 2 is made up of four districts/cities with a medium average APS of 31.87%. Cluster 3 comprises four districts/cities with a high average APS of 35.52%. Finally, Cluster 4 consists of three districts/cities with a very high average APS of 37.99%. The visualisation of the APS clustering results in South Kalimantan Province is presented as follows:

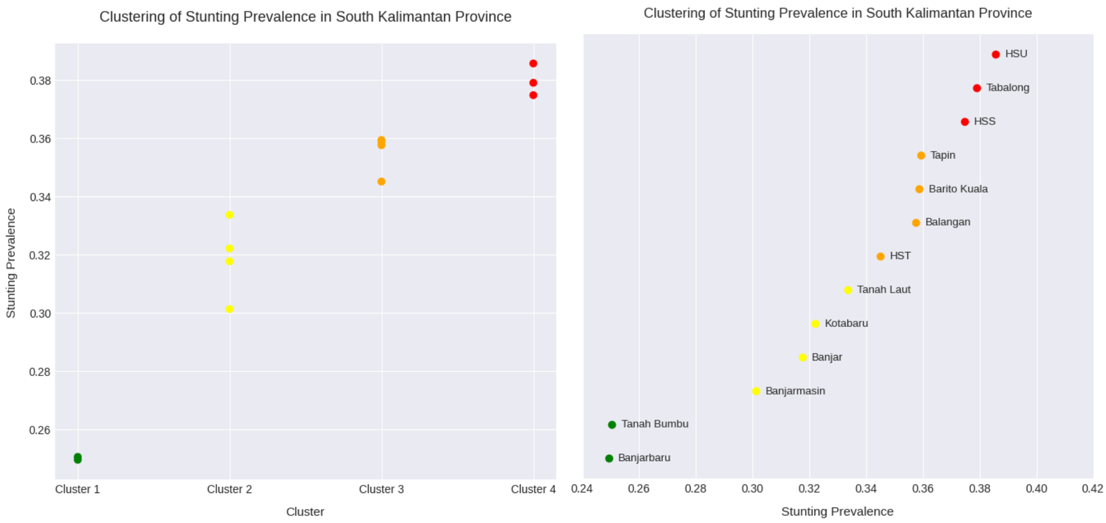

Figure 7. Clustering analysis results of stunting prevalence

The results of this clustering analysis can be used by local governments to develop more effective and efficient stunting reduction strategies. The clustering analysis, which is based on APS trends, allows local governments to allocate resources and interventions strategically to areas that have consistently high or increasing stunting rates, rather than relying solely on data from the previous year. By adjusting the stunting reduction program based on the generated clusters, access to stunting reduction programs can be prioritized for districts/cities that fall within clusters with a high average APS. This will enable local governments to target their resources more effectively and efficiently, ultimately leading to better public health outcomes.

5. Conclusions and Recommendations

5.1 Conclusions

This study highlights the potential of using localized data to develop a targeted and data-driven approach for reducing stunting in South Kalimantan, Indonesia. By analysing a broad range of variables that encompass social, governmental, nutritional, and environmental factors, we found a significant negative relationship between stunting and human development index and active integrated health posts (Posyandu). Additionally, there was a significant positive relationship with antenatal care K4 visits and cases of childhood malnutrition. These findings underscore the complex nature of stunting and highlight the need for a comprehensive approach to address it.

The predictive modeling also demonstrated the capability of machine learning algorithms, particularly Linear Regression in creating a reliable predictive framework for forecasting stunting prevalence rates (APS). The Linear Regression algorithm had the best performance with an R-squared score of 0.756, Mean Absolute Error of 0.027, and Mean Squared Error of 0.001, indicating its potential for predicting APS with accuracy. This predictive model can serve as a powerful tool for local governments to anticipate stunting trends and customize their interventions accordingly.

In the final stage, clustering analysis, categorise districts and cities into four distinct clusters based on their APS trends. The clusters ranged from a relatively low average APS of 25.00% to a very high average APS of 37.99%, highlighting the varying degrees of stunting prevalence and emphasizing the need for a targeted intervention approach. This clustering analysis not only offers a clear visualization of the APS disparities across different regions but also provides a strategic roadmap for resource allocation and targeted interventions.

5.2 Recommendations

Based on our comprehensive analysis, several recommendations are proposed for the local government in the South Kalimantan Province. First, it is crucial to design and implement stunting reduction strategies that are tailored to the specific factors found to influence stunting prevalence significantly. Emphasis should be placed on social, governmental, nutritional, and environmental factors, which include enhancing the community's access to vital health services, ensuring complete immunisation and supplements for pregnant women, and improving access to proper sanitation. At the same time, strategies should also tackle social determinants by improving the economic status and education levels of families. In addition, environmental factors, notably increasing access to appropriate sanitation and clean environment, are also important considerations.

Secondly, in recognising the disparities in stunting prevalence across districts and cities, it is necessary to prioritize regions with a higher prevalence of stunting when selecting areas for stunting alleviation initiatives. This strategic recalibration can significantly enhance the impact of interventions.

Lastly, data-driven methodologies can provide valuable insights that empower local governments to formulate more refined and effective strategies for curtailing stunting. Predictive modeling and clustering analysis can guide the allocation of resources and interventions, which can lead to a more targeted and impactful approach towards reducing stunting in South Kalimantan.

References

[1] I. Purnamasari, F. Widiyati, and M. Sahli, “Analisis Faktor Yang Mempengaruhi Kejadian stunting Pada Balita," Jurnal Penelitian dan Pengabdian Kepada Masyarakat UNSIQ, vol. 9, no. 1, pp. 48–56, 2022.

[2] E. Noorhasanah and N. I. Tauhidah, “Hubungan Pola Asuh Ibu Dengan Kejadian stunting Anak Usia 12-59 bulan," Jurnal Ilmu Keperawatan Anak, vol. 4, no. 1, pp. 37–42, May 2021.

[3] R. Ambarwati and S. P. Hastono, “Pemodelan Faktor-Faktor Yang Mempengaruhi Balita Gizi Buruk di provinsi Jawa barat Tahun 2017 Dengan Pendekatan spatial autoregressive model (SAR)," Jurnal Ilmu Kesehatan Masyarakat, vol. 9, no. 02, pp. 121–128, May 2020.

[4] H. Wati, N. Y. Kartika, M. Efendi, and R. Rusdiansyah, “Perkembangan prevalensi balita stunting di Provinsi Kalimantan Selatan," Jurnal Geografika (Geografi Lingkungan Lahan Basah), vol. 3, no. 1, p. 33, 2022.

[5] Rokom, “Prevalensi stunting di Indonesia turun ke 21,6% dari 24,4%," Kementerian Kesehatan, 25-Jan-2023. [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20230125/3142280/prevalensi-stunting-di-indonesia-turun-ke-216-dari-244/ (accessed Sep. 03, 2023).

[6] Kementerian Kesehatan, "Buku Saku Hasil Survei Status Gizi Indonesia (SSGI) 2022". [Online]. Available: https://kesmas.kemkes.go.id/assets/uploads/contents/attachments/09fb5b8ccfdf088080f2521ff0b4374f.pdf (accessed Sep. 03, 2023).

[7] Perwakilan BPKP Provinsi Kalimantan Selatan, "Hasil Pengawasan dalam rangka Pengayaan atas Hasil Evaluasi Program Percepatan Penurunan Stunting pada Provinsi Kalimantan Selatan", PE.09.03/LHP-309/PW16/2/2022, 27-Oct-2022.

[8] I. Hanafi, “Penanganan stunting di Kalsel, belum terarah pada lokus stunting tinggi - ANTARA News Kalimantan Selatan,” ANTARA News Kalimantan Selatan, Nov. 17, 2022. https://kalsel.antaranews.com/berita/350269/penanganan-stunting-di-kalsel-belum-terarah-pada-lokus-stunting-tinggi (accessed Sep. 03, 2023).

[9] D. Susanto, “Ironi Kalsel, Daerah Kaya dengan Angka Stunting yang Tinggi,” Media Indonesia, Nov. 18, 2022. https://mediaindonesia.com/nusantara/538350/ironi-kalsel-daerah-kaya-dengan-angka-stunting-yang-tinggi (accessed Sep. 04, 2023).

[10] H. Torlesse, A. A. Cronin, S. K. Sebayang, and R. Nandy, “Determinants of stunting in Indonesian children: Evidence from a cross-sectional survey indicate a prominent role for the water, sanitation and hygiene sector in stunting reduction,” BMC Public Health, vol. 16, no. 1, 2016.
[11] Y. Febrina, “Faktor Risiko Kejadian Stunting pada Bayi Baru Lahir di RSUD Wonosari Kabupaten Gunungkidul Tahun 2016,” thesis, 2019.
[10] D. H. Ulfani, D. Martianto, and Y. F. Baliwati, “Faktor-Faktor Sosial Ekonomi dan Kesehatan Masyarakat Kaitannya dengan Masalah gizi Underweight, Stunted, dan Wasted di indonesia: Pendekatan Ekologi Gizi,” Jurnal Gizi dan Pangan, vol. 6, no. 1, p. 59, 2011.
[11] A. Widiyanto, J. T. Atmojo, and A. T. Darmayanti, “Pengaruh Faktor Kerawanan Pangan dan Lingkungan terhadap Stunting,” Jurnal Ilmu Kesehatan, vol. 8, no. 1, May 2019.
[12] M. Mambang, F. D. Marleny, and M. Zulfadhilah, “Prediction of linear model on stunting prevalence with machine learning approach,” Bulletin of Electrical Engineering and Informatics, vol. 12, no. 1, pp. 483–492, 2023.
[13] M. S. Haris, M. Anshori, and A. N. Khudori, “Prediction of Stunting Prevalence in East Java Province with Random Forest Algorithm,” Jurnal Teknik Informatika (JUTIF), vol. 4, no. 1, pp. 11–13, Feb. 2023.
[14] A. N. Khudor, M. S. Haris, and W. T. Kusuma, “Perbandingan Metode Supervised Machine Learning untuk Prediksi Prevalensi Stunting di Provinsi Jawa Timur,” Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), vol. 9, no. 7, pp. 1571–1576, Dec. 2022.
[15] M. D. Chandra, E. Irawan, I. S. Saragih, A. P. Windarto, and D. Suhendro, “Penerapan Algoritma K-means dalam Mengelompokkan Balita yang Mengalami Gizi Buruk Menurut Provinsi,” Jurnal Teknologi Informasi dan Rekayasa Komputer, vol. 2, no. 1, pp. 30–38, 2021.
[16] M. R. Anggraeni, U. Yudatama, and Maimunah, “Clustering Prevalensi Stunting Balita Menggunakan Agglomerative Hierarchical Clustering,” Jurnal Media Informatika Budidarma, vol. 7, no. 1, pp. 351–359, Jan. 2023.