International Journal of Innovative Computing, Information and Control
Volume **19**, Number **4**, August **2023**

ICIC International ©2023 ISSN 1349-4198
pp. **1073–1086**

PREDICTING A STUNTING PREVALENCE USING SEMI-SUPERVISED LEARNING MODELS IN EAST NUSA TENGGARA

Stefanus Pieter Manongga[1], Hendry Hendry[2,*]
and Daniel Herman Fredy Manongga[2]

[1]Faculty of Public Health
Nusa Cendana University
Jl. Adisucipto Penfui, Kupang, East Nusa Tenggara 85001, Indonesia
spmanongga@staf.undana.ac.id

[2]Faculty of Information Technology
Satya Wacana Christian University
1-10 Notohamidjojo, Salatiga, Central Java 50715, Indonesia
danny.manongga@uksw.edu
*Corresponding author: hendry@uksw.edu

Received November 2022; revised March 2023

Abstract. *Stunting is a complicated problem to solve. The impact of stunting is on the long-term development of children where they may never reach their full high potential and have poor cognitive development which leads to less than optimal educational performance and decreased intellectual capacity, motor, and socioeconomic development. In the case of Indonesia, WHO includes Indonesia in countries with a high risk of stunting (30%-39%). This study aims to predict the risk of stunting using a semi-supervised learning model. However, it is necessary to explore the dominant determinants of stunting first. Unsupervised learning is superior for finding attributes that have a high correlation, while supervised learning is used to map attributes that correlate with stunting risk targets. Stunting data is recorded in each community health center for each district. There are some public health centers names such as "KAPAN", "PANTJE", "KUANFATU", "OEEKAM", and "OIMLASI". We found that the community health centers "KAPAN" and "PANTJE" form a stunting prevalence cluster with the highest values at 50% and above. "KUANFATU", "OEEKAM", and "OIMLASI" followed at 30% to 50%.*

Keywords: Stunting detection, Deep learning, Health sciences, Computer sciences

1. **Introduction.** Stunting is one of the major public health problems globally [1, 2, 3] and especially in developing countries [4, 5], including Indonesia. Stunting was defined as a height-for-age z-score less equal 2.0. Stunting in children can occur due to inadequate nutritional intake [6, 19] or infectious diseases, but also because of food insecurity, inadequate feeding and care practices, poor environmental health, and poor health services [7, 8]. The impact of stunting is on the long-term development of children where they may never reach their full high potential and have poor cognitive development which leads to less than optimal educational performance and decreased intellectual capacity, motor, and socioeconomic development [9, 10, 11]. Globally, there are around 144 million children under 5 years of age suffering from stunting. Of this number, half of the children live in Asia, and two out of five children live in Africa [12].

In the case of Indonesia, WHO includes Indonesia in countries with a high prevalence of stunting (30%-39%) [11]. This country even ranks fifth among the countries with the

DOI: 10.24507/ijicic.19.04.1073

highest burden of stunting children [13]. Based on data from the 2018 Indonesian Basic Health Research, the percentage of Indonesian children under 5 years of age who experienced stunting in 2007, 2013 and 2018 respectively was 36.8%, 37.2% and 30.8% [14]. Meanwhile, the current prevalence of stunting in Indonesia is 27.67% [15]. The high prevalence of stunting nationally which is still above 20% shows that stunting is still a serious public health problem. In addition, there is also a high disparity between provinces in Indonesia, from the province of Bali with the lowest prevalence of stunting in children under five at 14.42% to the province of East Nusa Tenggara (NTT) with the highest prevalence of stunting at 43.82% [15]. The disparity also occurs within the province of NTT, where the lowest prevalence of stunting in children under five is in East Flores Regency, at 23.4%, and the highest is in Timor Tengah Selatan District (TTS), at 48.3% [16]. The prevalence of TTS stunting is more than two times more than the WHO stunting prevalence tolerance rate of 20%.

Observing the high stunting rate in Indonesia and the negative impacts that can be caused, the Government of Indonesia has set a national target of stunting prevalence in 2024 of 14 percent. With a stunting rate of 24.4 percent in 2021, to achieve this target a 2.7 percent reduction is needed every year [17]. Anthropometric measurements are often used to calculate indices, identify stunting, wasting, BMI, head circumference for age, and acute malnutrition. In terms of affordability and accessibility, anthropometric parameters are very helpful. However, the anthropometric indicator only shows the stunting classification. Malnutrition can occur for a variety of reasons. Previous studies have shown that demographic factors have a significant impact on stunting, but demographic characteristics often have overlapping factors [18].

This research contributes as follows. First, the use of a complete data set with 7 categories of indicators and 25 features. Second, data processing uses an unsupervised and supervised learning approach. Unsupervised learning (k-Means clustering model) groups the same features in one group. Supervised learning predicts the proportion of stunting. In this method, three machine learning algorithms, linear regression, decision trees, and random forests, are compared.

2. Related Work.

Malnutrition is one of the main causes of stunting in children under five, in addition to various infectious diseases. Children under five are more likely to be malnourished than other children. For this reason, most researchers use anthropometric measures based on the WHO Zscore model [21] in determining stunting status in children under five. The characterization parameters are Weight-for-Height Z-Score (WHZ) for underweight, Height-for-Age Z-Score (HAZ) for stunting, and Weight-for-Age Z-Score (WAZ) for underweight which is used to define nutrition bad children. Among these, the WHZ is largely considered to define malnourished children. Several studies that have used HAZ, WHZ, and WAZ as indicators of malnutrition include [22, 23, 24, 25, 26].

Anthropometric indicators are a very useful method both in terms of availability and cost effectiveness. However, anthropometric indicators only show the classification of stunting. In an effort to prevent stunting, more other relevant indicators are needed. Stunting due to malnutrition in children under the age of 5 years is the result of a complex interaction of the availability, accessibility, and utilization of food and health services [27]. Several researchers have shown that there are several demographic and socioeconomic factors that also influence the occurrence of stunting [28, 29, 30].

The extraction of stunting risk factors has also been investigated using various statistical techniques. Among these, linear regression and logistic regression have been extensively explored to detect malnutrition in children aged 0 to 59 months [24, 31, 32, 33, 34, 35]. The use of regression techniques in medical research shows that this technique is a versatile technique because it can measure associations, predict outcomes, and control the effects of confounding variables. However, the use of regression techniques requires assumptions that must be met. Several logistic regression assumptions that need to be considered include the structure of the dependent variable, the independence of the observations, and the absence of multicollinearity. As for linear regression, assumptions that need to be met include linearity, multivariate normality, absence of multicollinearity and autocorrelation, and homoscedasticity [36, 37, 38].

Logistic regression is a statistical method, but it is included as part of the machine learning algorithm, which belongs to the supervised learning technique. The use of machine learning applications has recently increased as an alternative method in various health fields, including malnutrition prediction [20, 39, 40, 41], Anemia [42, 43], low birth weight [44, 45, 46], stunted and wasted [19, 20, 39].

For decades, conventional statistical models have been utilized to identify factors independently related to stunting in children under five [19]. This method is typically less robust when the number of covariates exceeds the number of observations and when there is multicorrelation between variables. In addition, the classical statistical technique puts rigorous constraints on data and data-generating procedures, such as error distribution and adding parameters with linear predictors, which may not be applicable in real-world scenarios [18].

Machine Learning (ML) methods are an alternative to traditional statistical methods because they can solve classification problems in a wide range of fields and are also flexible and powerful. [20, 47] emphasized the usefulness of different Machine Learning (ML) techniques (like artificial NN, SVM, decision trees, Naïve Bayes, and RF) for predicting childhood stunting in Bangladesh. This study aims to predict the risk of stunting using machine learning. However, it is necessary to explore the dominant determinants of stunting first. The data used in this study is secondary data from the results of the health survey in TTS District.

3. **Problem Definition and Preliminaries.**

Stunting is a complicated problem to solve. This study uses data from the Central Statistics Agency of East Nusa Tenggara. The dataset is secondary data that has been processed based on the attributes owned by the local government to measure the prevalence of stunting. The dataset consists of 7 categories of indicators, namely 1) Maternal and child health, 2) Counseling on nutrition, hygiene, and parental care, 3) Drinking water and sanitation, 4) PAUD, 5) Social protection, 6) Food security, and 7) Additional indicators in NTT. The target data label is the stunting rate and prevalence of stunting for each Community Health Center in NTT. The dataset consists of 278 Puskesmas spread across every sub-district and village in NTT Province.

Each category of indicators contains measurement attributes that have different amounts. The 1st category has ten attributes: a) Coverage of SEZ pregnant women receiving PMT Recovery, b) Coverage of pregnant women receiving IFA (TTD) at least 90 tablets during pregnancy, c) Coverage of under-fives who get PMT, d) Coverage of attendance at Posyandu (ratio of arrivals to total target), e) Coverage of Pregnant Women-K4, f) Coverage of children 6-59 months who receive Vitamin A, g) Coverage of children 0-11 months has complete basic immunization, h) Coverage of toddlers with diarrhea who received zinc supplementation, i) Coverage of young women getting TTD, and j) Postpartum service coverage. The 2nd category has two attributes, namely a) Coverage of pregnant women (mothers attending nutrition and health counseling), and b) Coverage of families participating in Toddler Family Development. The 3rd category has two attributes: a) Coverage of households using proper drinking water sources, and b) Coverage of households using proper sanitation. The 4th category has two attributes: a) Coverage of parents who take parenting classes, and b) Coverage of registered children aged 2-6 years (students) in PAUD. The 5th category consists of 3 attributes: a) Coverage of households participating in JKN/Jamkesda, b) Coverage of KPM PKH who receive nutrition and health FDS, and c) Family coverage of 1000 HPK poor groups as BPNT recipients. The 6th category only has 1 attribute, namely Village coverage applying KRPL. The last category consists of five attributes: a) Babies who receive exclusive breastfeeding, b) MP ASI counseling coverage, c) Malnourished children under five who are got treated, d) Toddlers with Pneumonia, and e) Pregnant with Malaria.

Data obtained from the Central Statistics Agency contains high noise, such as missing values, sparse data, and inconsistent data types. Table 1 shows the attributes of the dataset that have noise in the form of missing values. Noise must be handled by methods appropriate to the conditions. For example, missing values can be handled by filling in a value. In this study, missing values are handled by providing the value of the average or the value that appears most often from that attribute. Sparsity will be handled by deleting one line with a frequency of 5%.

Table 1. Statistics data of the dataset with missing values

Name	Mean	Median	Dispersion	Min	Max	Missing
Coverage of parents who attended parenting classes	1	0.00	9.949	0.00	100	78 (28%)
Coverage of registered children aged 2-6 years (students) in PAUD	12.431	0.00	2.291	0.00	100	78 (28%)
Random forest	0.3642	0.133	0.735	0.753	0.735	0.744
Coverage of young women getting TTD	84.096	100	0.406	0.00	100	13 (5%)
Coverage of families who follow the Toddler Family Development	30.579	0.00	1.266	0.00	100	9 (3%)
Coverage of households using proper sanitation	45.954	44.475	0.622	0.00	100	6 (2%)
Malnourished toddlers handled/ received treatment	60.451	100	0.781	0.00	100	5 (2%)
Pregnant with Malaria	0.082	0.00	12.062	0.00	14.29	4 (1%)
Coverage of households using proper drinking water sources	63.576	68.150	0.686	0.00	55.62	2 (1%)

4. **Proposed Methodology.**

In this section, we will describe the methodology to predict stunting in Timor Tengah Selatan. The system uses machine learning to learn the pattern which affects the stunting rate. Figure 1 shows the system architecture of the proposed model. The system starts by reading the dataset which is collected from "Central Bureau of Statistics" (BPS).

The proposed model consists of 2 main phases. Before the main stage, preprocessing is carried out to process dirty data into data without missing values. The preprocessing phase ensures that the data used is feasible to be processed using the system. Unsupervised learning is a method with a focus on data-driven; this is used to find the best attribute to group data points according to the closeness of the characteristics between data. We use unsupervised learning (k-Means clustering model) to find the features that have the best

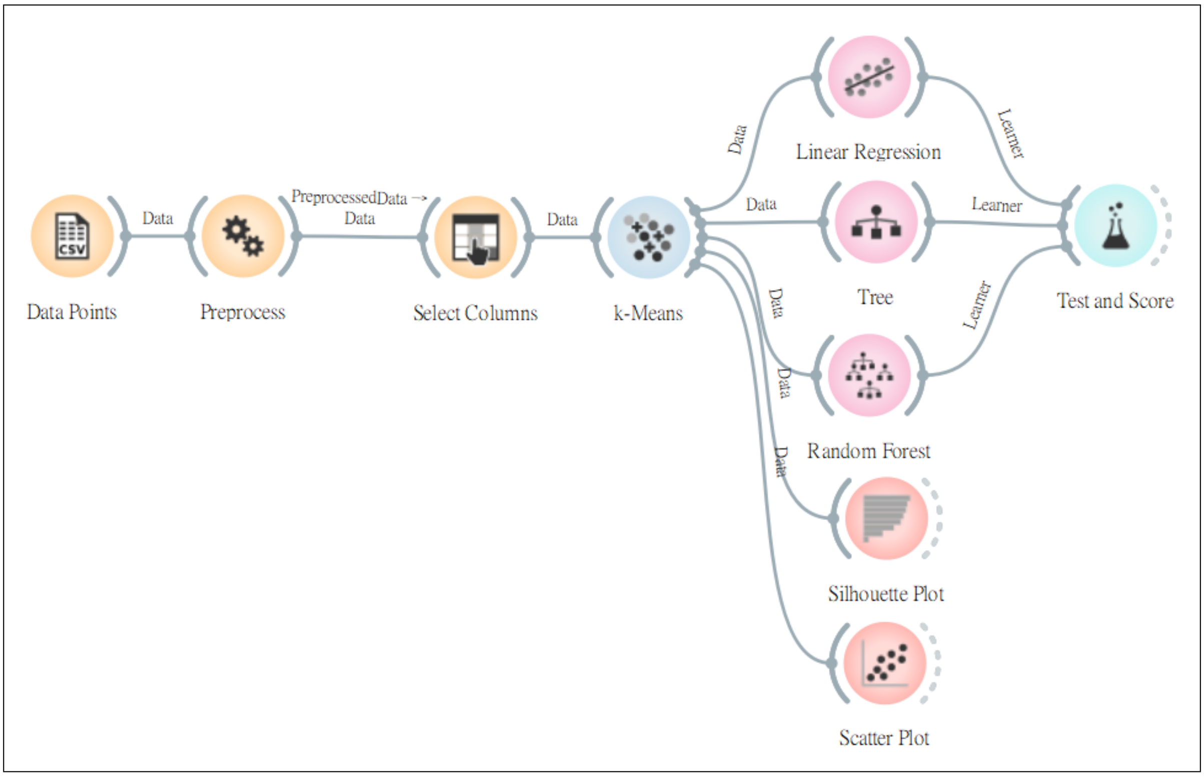


FIGURE 1. System architecture

proximity to the data points. The first main stage is the implementation of k-Means to find features that have close data characteristics for clustering. The second main phase is supervise learning which is used to predict the percentage of stunting in the province of NTT. We compare three suitable supervised learning algorithms for numerical data. The three algorithms are linear regression, decision tree, and random forest, respectively.

The final phase of the proposed model is model evaluation. In this phase, the model that has been built will look at the performance of the model using measurement metrics, i.e., Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the value of R^2.

4.1. **Clustering.** Clustering is a method for dividing data into groups or clusters so that data having high similarity will be in the same group. Clustering is used to perform an initial analysis of a problem, to make us understand the case better. For example, a public health center groups patients based on demographics. The grouping results can provide an initial description of the characteristics of the stunting in each region. This initial description can be used for deeper analysis purposes, for example, determining stunting handling models, and determining treatment.

Clustering is an unsupervised learning method, so it does not need explicit class definition in clustering. During the clustering process, all data will be used for clustering. This is different from the supervised learning method (regression and classification), where the data owned is used to build the model, so that after the model is finished, the classification process is carried out by utilizing the model. Because there is no explicit class definition, the same data can produce different groups depending on how the grouping is done.

The k-Means clustering method divides the dataset into k clusters. The process of grouping in k-Means clustering is based on the data distance to the cluster center point. Data will enter into the cluster whose central point is closest to the data. k-Means is a simple and widely known clustering method. Given a set $X = \{x_1, x_2, \ldots, x_n\}$ of n

data points. k-Means aims to group into k subsets called clusters $C = \{c_1, c_2, \ldots, c_k\}$ and calculate the cluster center for each cluster C_i to minimize Equation (1).

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 = \arg\min_{C} \sum_{i=1}^{k} |C_i| \, VarC_i \tag{1}$$

where μ_i is the mean points in C_i. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster in Equation (2).

$$\arg\min_{S} \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{x,y \in C_i} \|x - y\|^2 \tag{2}$$

The stages in k-Means clustering are as follows: a) Choose k points at random to be used as cluster centers, b) Each data is grouped into a cluster whose central point is closest to the data, c) Recalculate the center point of the cluster formed based on the average data entered into the cluster, and d) Repeat step b) until there are no more data moving clusters.

One of the critical steps of the k-Means algorithm is determining the number of clusters (k). There are several methods for determining the k value; they are the elbow method, the silhouette index method, and others. In this paper, we use the silhouette coefficient to determine the number of the clusters. The silhouette coefficient measures cluster quality by measuring how similar data in one cluster is compared to data in different clusters. Silhouette coefficient is formulated by

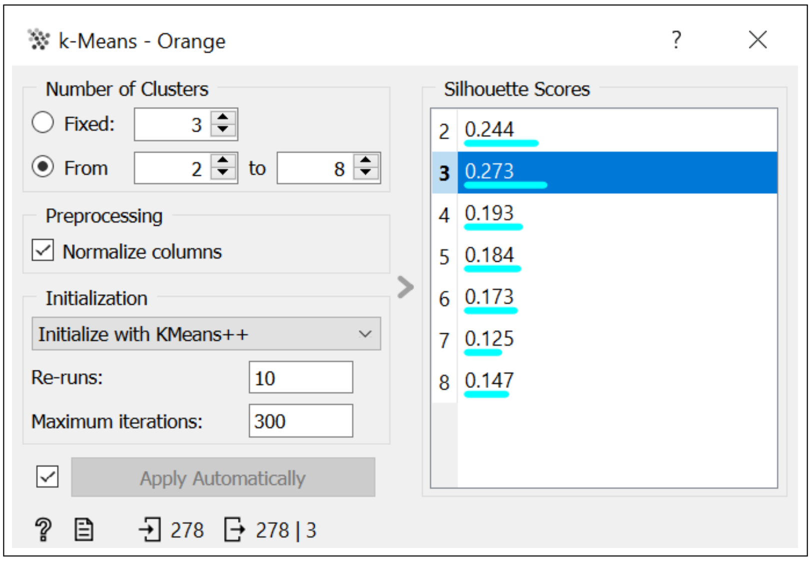
$$S = \frac{b - a}{\max(a, b)} \tag{3}$$

where a is the average distance of data with other data in the same cluster, and b is the average distance of data with other data in different clusters. The silhouette coefficient value is in the range of -1 to 1. If the silhouette coefficient value $= 1$ it shows that the cluster formed is dense and well separated. If the silhouette coefficient value $= 0$, this shows the presence of overlapping clusters. If the silhouette coefficient $= -1$ it shows that the clusters formed are not correct because the data distance in one cluster is greater than the distance between data and data in different clusters. Figure 2 shows the results of calculating the silhouette index for selecting k values in k-Means. In the figure, we found that the best k value is three clusters, based on the silhouette coeficient value.

4.2. **Base classifier for prediction.**

4.2.1. *Linear regression.* Regression is a method used to measure the relationship between one variable and another. Regression is used when the relationship between these variables shows a functional dependency relationship. Functional dependence is a relationship that occurs when the value of one variable will determine the value of another variable. The relationship can be linear or non-linear.

For example, suppose we have data on stunting prevalence and sanitary hygiene. Then we can do modeling using the regression method to determine the effect of sanitation hygiene on stunting prevalence. In this example, there are two variables, namely the stunting prevalence variable and the sanitation hygiene variable. The sanitation hygiene variable is a variable that can affect the prevalence of stunting. The relationship between the two variables shows a functional dependence. Variables that can determine other variables are often called independent variables or predictor variables (variables used to predict), while variables whose value depends on other variables are called dependent variables or response variables.



k-Means - Orange

Number of Clusters

○ Fixed: 3

● From 2 to 8

Preprocessing

☑ Normalize columns

Initialization

Initialize with KMeans++

Re-runs: 10

Maximum iterations: 300

☑ Apply Automatically

278 278 | 3

Silhouette Scores

2	0.244
3	0.273
4	0.193
5	0.184
6	0.173
7	0.125
8	0.147

Figure 2. Silhouette coeficient for choosing cluster number

Linear regression measures the relationship between the independent and dependent variables using a linear approach. If the number of independent variables is one, then the regression is often called simple linear regression, whereas if the number of independent variables is more than one, it is often called multiple variable linear regression. In simple linear regression, the trend in the dataset is modeled by a linear equation which can be formulated as the following Equation (4):

$$y = ax + b \tag{4}$$

where y is dependant variable, x is independant variable, a is intercept coefficient, and b is regression coefficient (related to the slope).

4.2.2. *Decision tree.* The decision tree is a classification method that uses a tree structure in its model representation. Decision trees can break down complex decision-making processes into simpler ones so that decision-making will better represent the solution to the problem. The basic idea of a decision tree is to look for attributes that have a high influence on the output. The higher the influence of these attributes, the better these attributes are in classifying the output.

In a decision tree, the ability of an attribute to identify the output class is expressed in the concepts of entropy and information gain. Entropy states the degree of randomness of the value of a variable in identifying the output class. If the value of an independent variable has low entropy, it means that the value of this variable has a low degree of randomness in determining the output class, so the value of this variable is good for predicting the output class. Meanwhile, if the value of an independent variable has high entropy, then the degree of randomness of the variable's value in determining the output class is also high, so it is not good for predicting the output class.

Consider a dataset with N classes. The entropy may be calculated using Equation (5).

$$E = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{5}$$

$$Gain = E_{parent} - E_{children} \tag{6}$$

where p_i is the probability of randomly selecting an example in class i. Information gain is a process for determining branching by utilizing entropy information. The decision tree method will calculate various possible branching schemes and then choose the best branching scheme using information gain in Equation (6). In general, the procedure in a decision tree can be described as follows: 1) Select an attribute in the dataset, 2) Calculate the entropy of the attribute, 3) Select the attribute with the best value as a node, and 4) Repeat step 1) for each branch that is formed.

4.2.3. *Random forest.* Random forest is an ensemble learning method for classifications composed of many decision trees. Ensemble learning is an algorithm composed of several algorithms to get better performance. In a random forest, the constituent algorithm is the decision tree algorithm. When carrying out the classification, each decision tree in the random forest will carry out the classification process. The output will be obtained based on the majority output from the decision tree in the random forest.

5. Experiments and Results.

In this section, we present our results and discuss the validation procedure of the proposed method.

5.1. Data collection and filtration.

Dataset is collected from the Central Bureau of Statistics (BPS). Table 2 shows the preview data of the dataset. The dataset consists of 28 features and two target variables. The features are divided into 3 attributes of categorical type and 25 attributes of numeric type. Two target labels are the number of stunting cases and the prevalence of stunting. The raw dataset contains missing values and filtered noise by filling in the missing values with the mean or most frequent value in Equation (7), and pre-processing is carried out to remove sparsity using percentage techniques.

$$\widehat{y_{mi}} = b_{r0} + \sum_{j} b_{rj} z_{mij} + \widehat{e}_{mi} \tag{7}$$

Table 2. Preview of the dataset

Subdistrict	Public health center	Village	Number of stunted children (short and very short)	% Prevalence of STUNTING	Coverage of pregnant women with KEK who receive recovery PMT	⋯	Coverage of pregnant women receiving IFA (TTD) of at least 90 tablets during pregnancy
MOLLO UTARA	KAPAN	TOMANAT	39	82.978	100	⋯	80
MOLLO UTARA	BATI	HALME	52	80	100	⋯	100
NUNKOLO	NUNKOLO	NENOAT	111	75	100	⋯	83
FATUKOPA	FATUKOPA	ELO	38	74.509	100	⋯	84
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
KIE	KIE	TESI AYOFANU	5	2.604	100	⋯	100

5.2. Parameter setting and results.

In this section, we will describe the parameters used in each model. The configuration of each model parameter setting is shown as follows.

- k-Means: We use the k subset with the number of three. The selection of k is done using the silhouette coefficients method with the highest value. The silhouette score for k equals 2 is 0.244, for k equals 3 is 0.273, for k equals 4 is 0.193, and for k is greater than 4, the silhoutte score continues to decrease. Because the silhouette value for $k = 3$ is the highest (0.273) then $k = 3$ is taken to be used as the k value in the k-Means algorithm. We use normalization preprocessing, 10 times re-runs, and 300 maximum iterations configuration.

- linear regression: We use linear regression with a ridge regression model (L2). The setting for the value of regularization strength is alpha = 0.0001, using the fit intercept.
- decision tree: The decision tree model used is an induced binary tree. The parameter setting for the minimum number of instances in leaves is 2, the tree will not be split as long as the value of the subset is less than 5, and we limit the depth of the tree to be built 100 levels down.
- random forest: The random forest model uses the same decision tree configuration. The number of ensemble trees used is ten decision trees. The formed tree will not be split if the number of subsets is less than five.

From the k-Means clustering, we get results of stunting prevalence characteristics from the dataset. Figure 3 shows the percentage of stunting prevalence in each health center in the province of NTT. From the figure, we found that the highest prevalence of stunting is in several community health centers. "KAPAN" and "PANTJE" form a stunting prevalence cluster with the highest values at 50% and above. "KUANFATU", "OEEKAM", and "OIMLASI" followed at 30% to 50%. Based on the target by the President of the Republic of Indonesia to reduce the stunting rate in Indonesia in 2024 [17] to 14%, this value is still above the target to be achieved.

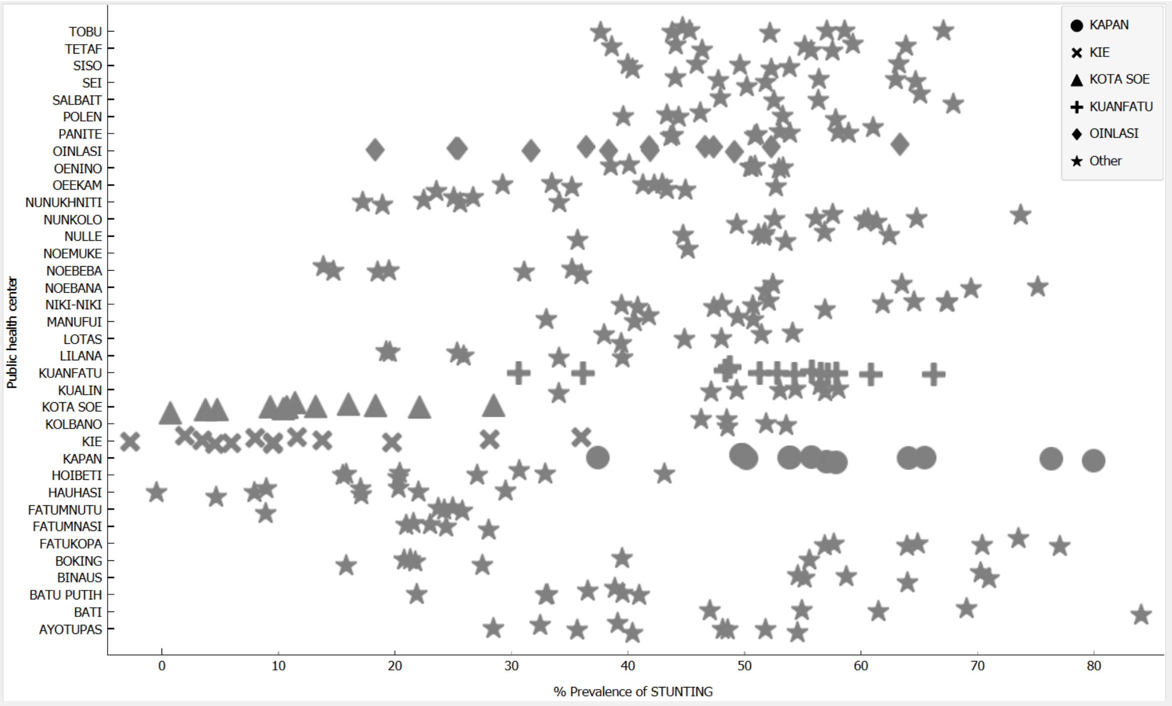
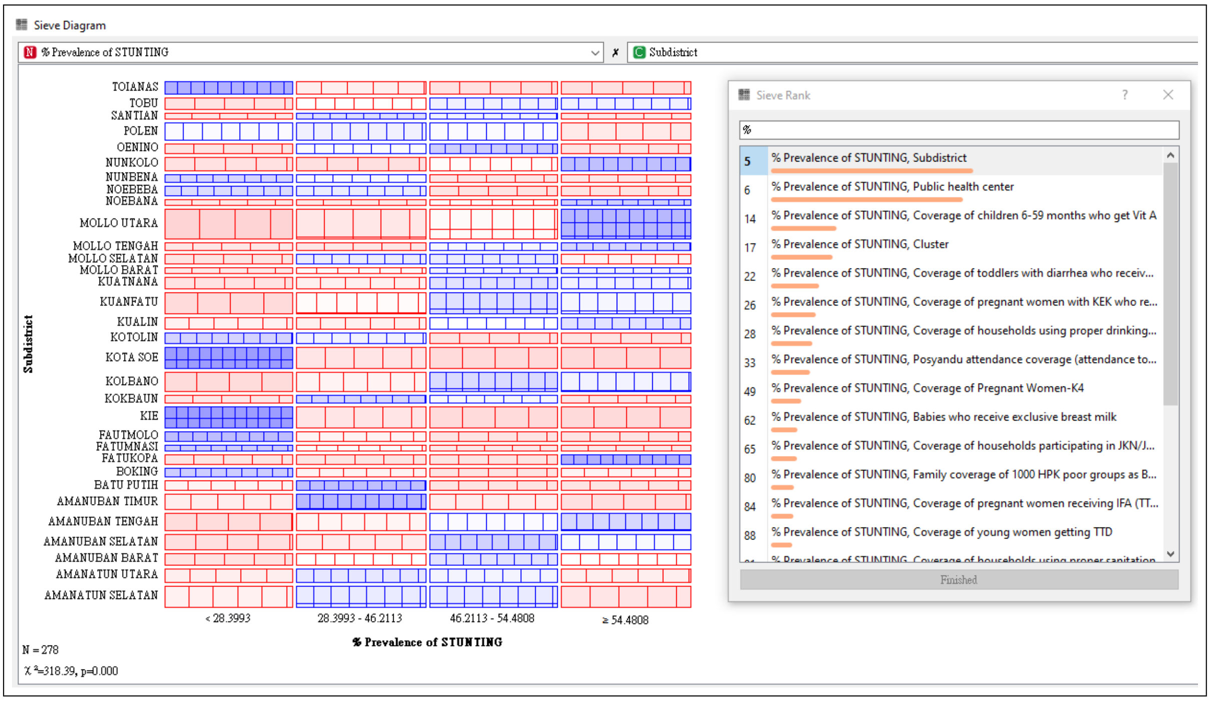


FIGURE 3. Clustering for public health center

Clustering with k-Means also produces attribute features that have a high impact on the causes of stunting prevalence. This stage is carried out to obtain feature extraction from the dataset attributes that will be used in the prediction process using a supervised learning model. Figure 4 shows the best features of the dataset that affect the prevalence of stunting in an area. From the figure, we find the most influential attributes include Subdistrict, Public health center, Coverage of children 6-59 months who get Vit A, Cluster, Coverage of toddlers with diarrhea who received zinc supplementation, Coverage of pregnant women with KEK who receive recovery PMT, and Coverage of households using



N = 278
χ²=318.39, p=0.000

5	% Prevalence of STUNTING, Subdistrict
6	% Prevalence of STUNTING, Public health center
14	% Prevalence of STUNTING, Coverage of children 6-59 months who get Vit A
17	% Prevalence of STUNTING, Cluster
22	% Prevalence of STUNTING, Coverage of toddlers with diarrhea who receiv...
26	% Prevalence of STUNTING, Coverage of pregnant women with KEK who re...
28	% Prevalence of STUNTING, Coverage of households using proper drinking...
33	% Prevalence of STUNTING, Posyandu attendance coverage (attendance to...
49	% Prevalence of STUNTING, Coverage of Pregnant Women-K4
62	% Prevalence of STUNTING, Babies who receive exclusive breast milk
65	% Prevalence of STUNTING, Coverage of households participating in JKN/J...
80	% Prevalence of STUNTING, Family coverage of 1000 HPK poor groups as B...
84	% Prevalence of STUNTING, Coverage of pregnant women receiving IFA (TT...
88	% Prevalence of STUNTING, Coverage of young women getting TTD
	% Prevalence of STUNTING, Coverage of households using proper sanitation

Figure 4. Dataset attributes that have a significant relation to the prevalence of stunting

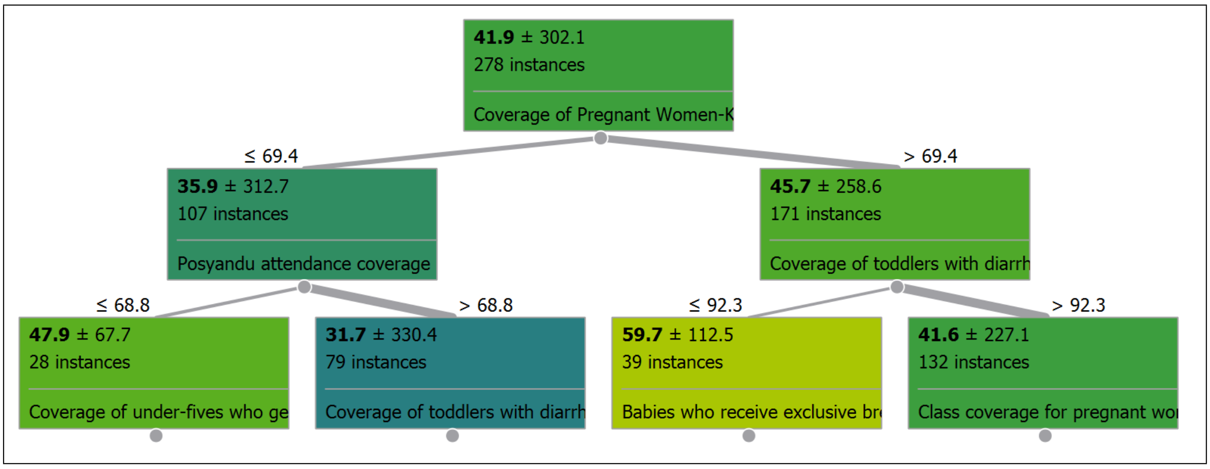

41.9 ± 302.1
278 instances
Coverage of Pregnant Women-K
≤ 69.4
> 69.4
35.9 ± 312.7
107 instances
Posyandu attendance coverage
45.7 ± 258.6
171 instances
Coverage of toddlers with diarrh
≤ 68.8
> 68.8
≤ 92.3
> 92.3
47.9 ± 67.7
28 instances
Coverage of under-fives who ge
31.7 ± 330.4
79 instances
Coverage of toddlers with diarrh
59.7 ± 112.5
39 instances
Babies who receive exclusive bre
41.6 ± 227.1
132 instances
Class coverage for pregnant wo

Figure 5. First 3 levels decision tree

proper drinking water sources. All the features will be used as input attributes in the next stage of supervised learning to predict stunting prevalence in regions.

Figure 5 shows the first 3 levels of the tree created from the entropy. From the figure, we found attribute "Coverage of Pregnant Women-K4" has the highest entropy. It becomes the root's tree to divide the value of stunting prevalence. The next level of the tree has attributes: "Posyandu attendance coverage (attendance to total target ratio)", and "Coverage of toddlers with diarrhea who received zinc supplementation" becomes the divider to determine the stunting prevalence value. Other attributes will fill the deeper levels of the tree according to the entropy value each has. These attributes can be used as a divider to predict the prevalence of stunting cases. The same ensemble tree will be used for the random forest model.

5.3. Performance evaluation.

Evaluation of the model is done by calculating the difference between the predicted value and the target value. The difference between the predicted value and the target value is often called the error. Several types of evaluation can be used for the model, among others.

- Mean Absolute Error (MAE): the average of the absolute values/absolute errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y_i}| \tag{8}$$

- Mean Squared Error (MSE): the average of the squared errors

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \tag{9}$$

- Root Mean Squared Error (RMSE): the root of MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y_i})^2} \tag{10}$$

Table 3 shows the comparison results of each base classifier. From the table, we found that the random forest outperforms other methods. The random forest gets the best MSE value at 200.220, the best RMSE value at 14.150, the best MAE value at 10.736, and the best R^2 value at 0.337, followed by the linear regression model the second best for the MSE value at 263.850 and RMSE value at 16.243, but the linear regression MAE and R^2 values are worse than the decision tree at 13.060, and 0.127, respectively. The decision tree gets an MSE value of 292.047, an RMSE value of 17.089, an MAE value of 11.445, and R^2 value of 0.033.

Table 3. Comparison of results between methods (test phase)

Methods	MSE	RMSE	MAE	R^2
Linear regression	263.850	16.243	13.060	0.127
Decision tree	292.047	17.089	11.445	0.033
Random forest	200.220	14.150	10.736	0.337

6. Conclusions.

We predict the prevalence of stunting using machine learning methods. From the experiment, we found that the community health centers “KAPAN” and “PANTJE” form a stunting prevalence cluster with the highest values at 50% and above. “KUANFATU”, “OEEKAM”, and “OIMLASI” followed at 30% to 50%. The prediction model uses supervised learning, linear regression, a decision tree, and a random forest model. The random forest model returns the best prediction error based on MSE, RMSE, MAE, and R^2 loss metrics.

Predicting the prevalence of stunting is a complicated problem. In this research, we only use the data from statistical methods. Many external factors such as demographics, climate, environmental conditions, and social and culture are not considered yet in this study.

Acknowledgment. This research is supported by DIKTI-Kedaireka Indonesia.

REFERENCES

[1] T. Beal, A. Tumilowicz, A. Sutrisna, D. Izwardy and L. M. Neufeld, A review of child stunting determinants in Indonesia, *Maternal & Child Nutrition*, vol.14, no.4, e12617, 2018.

[2] R. E. Black, C. G. Victora, S. P. Walker, Z. A. Bhutta, P. Christian, M. de Onis, M. Ezzati, S. Grantham-McGregor, J. Katz, R. Martorell et al., Maternal and child undernutrition and overweight in low-income and middle-income countries, *Lancet*, vol.382, pp.427-451, 2013.

[3] United Nations Children's Fund, World Health Organization, *Low Birthweight: Country, Regional and Global Estimates*, New York, NY, USA, 2004.

[4] G. Danaei, K. G. Andrews, C. R. Sudfeld, G. Fink, D. C. McCoy, E. Peet and W. W. Fawzi, Risk factors for childhood stunting in 137 developing countries: A comparative risk assessment analysis at global, regional, and country levels, *PLoS Medicine*, vol.13, no.11, e1002164, 2016.

[5] T. Huriah and N. Nurjannah, Risk factors of stunting in developing countries: A scoping review, *Open Access Macedonian Journal of Medical Sciences*, vol.8, pp.155-160, 2020.

[6] D. J. Raiten and A. A. Bremer, Exploring the nutritional ecology of stunting: New approaches to an old problem, *Nutrients*, vol.12, no.2, 371, DOI: 10.3390/nu12020371, 2020.

[7] B. Bustami and M. Ampera, The identification of modeling causes of stunting children aged 2-5 years in Aceh Province, Indonesia (Data analysis of nutritional status monitoring 2015), *Open Access Maced. J. Med. Sci.*, vol.8, pp.657-663, https://doi.org/10.3889/oamjms.2020.4659, 2020.

[8] United Nation Children's Fund, *The State of the World's Children 2014 in Number: Every Child Counts*, New York, https://doi.org/10.18356/8504d62b-en, 2014.

[9] WHO, *Levels and Trends in Child Malnutrition*, http://www.who.int/nutgrowthdb/2018-jmebrochure.pdf?ua=1, Accessed on 4 March, 2019.

[10] S. Grantham-McGregor, Y. B. Cheung, S. Cueto, P. Glewwe, L. Richter and B. Strupp, Developmental potential in the first 5 years for children in developing countries, *Lancet*, vol.369, pp.60-70, 2007.

[11] World Health Organization, *Country Profile Indicators: Interpretation Guide*, Geneva, Switzerland, 2010.

[12] UNICEF, World Health Organization, *World Bank, Levels and Trends in Child Malnutrition UNICEF-WHO-World Bank Group Joint Child Malnutrition Estimates: Key Findings of the 2020 Edition*, World Bank Group, New York, NY, USA, https://apps.who.int/iris/bitstream/handle/10665/331621/9789240003576-eng.pdf, 2020.

[13] UNICEF, *Improving Child Nutrition: The Achievable Imperative for Global Progress*, New York, NY, USA, 2013.

[14] R. I. Kemenkes, Report on the results of Indonesia's basic health research (riskesdas) in 2018, *Basic Health Research*, pp.182-183, https://www.litbang.kemkes.go.id/laporan-riset-kesehatan-dasar-riskesdas/, 2018.

[15] R. I. Kemenkes, *Indonesian Health Profile 2020*, Ministry of Health RI, Jakarta, https://www.kemkes.go.id/downloads/resources/download/pusdatin/profil-kesehatan-indonesia/Profil-Kesehatan-Indonesia-Tahun-2020.pdf, 2021.

[16] *Pocket Book of Indonesian Nutrition Status Study Results (SSGI)*, http://www.badanpolicy.kemkes.go.id/buku-saku-hasil-studi-status-gizi-indonesia-ssgi-tahun-2021/, 2021.

[17] President of the Republic of Indonesia, *Government Targets Stunting Prevalence Rate below 14 Percent by 2024*, https://www.presidenri.go.id/siaran-pers/governmental-targetkan-angka-prevalensi-stunting-di- Bawah-14-persen-pada-2024/, 2022.

[18] O. N. Chilyabanyama, R. Chilengi, M. Simuyandi, C. C. Chisenga, M. Chirwa, K. Hamusonde, R. K. Saroj, N. T. Iqbal, I. Ngaruye and S. Bosomprah, *Performance of Machine Learning Classifiers in Classifying Stunting among Under-Five Children in Zambia*, https://doi.org/10.3390/children9071082, 2022.

[19] S. M. J. Rahman, N. A. M. F. Ahmed, M. M. Abedin, B. Ahammed, M. Ali, M. J. Rahman et al., Investigate the risk factors of stunting, wasting, and underweight among under-five Bangladeshi children and its prediction based on machine learning approach, *PLoS ONE*, vol.16, no.6, e0253172, https://doi.org/10.1371/journal.pone.0253172, 2021.

[20] M. Shahriar, M. S. Iqubal, S. Mitra and A. K. Das, A deep learning approach to predict malnutrition status of 0-59 month's older children in Bangladesh, *IEEE Int. Confer. Indus. Artifi. Intell. Commun. Technol.*, pp.145-149, 2019.

[21] World Health Organization, Use and interpretation of anthropometric indicators of nutritional status, *Bulletin of World Health Organization*, vol.64, no.6, pp.929-941, 1986.

[22] Z. Momand, P. Mongkolnam, P. Kositpanthavong and J. H. Chan, Data mining based prediction of malnutrition in Afghan children, *2020 12th International Conference on Knowledge and Smart Technology (KST)*, DOI: 10.1109/kst48564.2020.9059388, 2020.

[23] S. Das and R. M. Rahman, Application of ordinal logistic regression analysis in determining risk factors of child malnutrition in Bangladesh, *Nutr. J.*, vol.10, 124, DOI: 10.1186/1475-2891-10-124, 2011.

[24] K. R. Bhowmik and S. Das, On exploring and ranking risk factors of child malnutrition in Bangladesh using multiple classification analysis, *BMC Nutr.*, vol.3, 73, https://doi.org/10.1186/s40795-017-0194-7, 2017.

[25] T. Adhikary, A. K. Das, M. A. Razzaque, M. E. H. Chowdhury and S. Parvin, Test implementation of a sensor device for measuring soil macronutrients, *2015 International Conference on Networking Systems and Security (NSysS)*, pp.1-8, DOI: 10.1109/NSysS.2015.7042951, 2015.

[26] M. Akter, F. T. Zohra and A. K. Das, Q-MAC: QoS and mobility aware optimal resource allocation for dynamic application offloading in mobile cloud computing, *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox's Bazar, pp.803-808, 2017.

[27] I. Govender, S. Rangiah, R. Kaswa and D. Nzaumvila, Malnutrition in children under the age of 5 years in a primary health care setting, *South African Family Practice*, vol.63, no.1, DOI: 10.4102/safp.v63i1.5337, 2021.

[28] S. E. Mahgoub, M. Nnyepi and T. Bandeke, Factors affecting prevalence of malnutrition among children under three years of age in Botswana, *Afr. J. Food Agri Nutr Develo.*, vol.6, no.1, 2006.

[29] M. F. Shaka, Y. B. Woldie, H. M. Lola, K. Y. Olkamo and A. T. Anbasse, Determinants of undernutrition among children under five years old in Southern Ethiopia: Does pregnancy intention matter? A community-based unmatched case-control study, *BMC Pediatr.*, vol.20, no.1, 101, DOI: 10.1186/s12887-020-2004-7, 2020.

[30] A. Gebre, P. S. Reddy, A. Mulugeta, Y. Sedik and M. Kahssay, Prevalence of malnutrition and associated factors among under-five children in pastoral communities of AFAR Regional State, Northeast Ethiopia: A community-based cross-sectional study, *Journal of Nutrition and Metabolism*, pp.1-13, DOI: 10.1155/2019/9187609, 2019.

[31] L. Abera, T. Dejene and T. Laelago, Prevalence of malnutrition and associated factors in children aged 6-59 months among rural dwellers of Damot Gale District, South Ethiopia: Community based cross sectional study, *International Journal for Equity in Health*, vol.16, no.1, 2017.

[32] A. K. Das, T. Adhikary, M. A. Razzaque, M. Alrubaian, M. M. Hassan, Z. Uddin and B. Song, Big media healthcare data processing in cloud: A collaborative resource management perspective, *Cluster Computing*, vol.20, no.2, pp.1599-1614, 2017.

[33] A. Nshimyiryo, B. Hedt-Gauthier, C. Mutaganzwa et al., Risk factors for stunting among children under five years: A cross-sectional population-based study in Rwanda using the 2015 demographic and health survey, *BMC Public Health*, vol.19, no.1, 175, https://doi.org/10.1186/s12889-019-6504-z, 2019.

[34] L. M. Tesfaw and H. M. Fenta, Multivariate logistic regression analysis on the association between anthropometric indicators of under-five children in Nigeria: NDHS 2018, *BMC Pediatr.*, vol.21, no.1, 193, https://doi.org/10.1186/s12887-021-02657-5, 2021.

[35] H. G. Mengesha, H. Vatanparast, C. Feng and P. Petrucka, Modeling the predictors of stunting in Ethiopia: Analysis of 2016 Ethiopian demographic health survey data (EDHS), *BMC Nutr.*, vol.6, no.52, DOI: 10.1186/s40795-020-00378-z, 2020.

[36] J. C. Stoltzfus, Logistic regression: A brief primer, *Acad. Emerg. Med.*, vol.18, no.10, pp.1099-1104, DOI: 10.1111/j.1553-2712.2011.01185.x, 2011.

[37] L. Statistics, *Binomial Logistic Regression Using SPSS Statistics*, Statistical Tutorials and Software Guides, https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php, 2015.

[38] D. Schreiber-Gregory and K. Bader, *Logistic and Linear Regression Assumptions: Violation Recognition and Control*, Henry M Jackson Foundation, 2018.

[39] A. Talukder and B. Ahammed, Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh, *Nutrition*, 110861, DOI: 10.1016/j.nut.2020.110861, 2020.

[40] S. Jain, T. Khanam, A. J. Abedi and A. A. Khan, Efficient machine learning for malnutrition prediction among under-five children in India, *2022 IEEE Delhi Section Conference (DELCON)*, pp.1-10, DOI: 10.1109/DELCON54057.2022.9753080, 2022.

[41] S. Kar, S. Pratihar, S. Nayak et al., Prediction of child malnutrition using machine learning, *2021 10th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)*, pp.1-4, DOI: 10.1109/IEMECON53809.2021.9689083, 2021.

[42] M. M. Islam, M. J. Rahman, D. C. Roy, M. M. Islam, M. Tawabunnahar, N. A. M. F. Ahmed et al., Risk factors identification and prediction of anemia among women in Bangladesh using machine learning techniques, *Curr. Women's Health Rev.*, vol.17, no.1, 2021.

[43] M. Jaiswal, A. Srivastava and T. J. Siddiqui, Machine learning algorithms for anemia disease prediction, in *Recent Trends in Communication, Computing, and Electronics*, Singapore, Springer, 2019.

[44] N. Eliyati, A. Faruk, E. S. Kresnawati and I. Arifieni, Support vector machines for classification of low birth weight in Indonesia, *J. Phy.: Conf. Series*, vol.1282, no.1, 012010, 2019.

[45] U. Hange, R. Selvaraj, M. Galani and K. Letsholo, A data-mining model for predicting low birth weight with a high AUC, *Int. Conf. Comput. Inf. Sci.*, pp.109-121, 2017.

[46] N. S. Borson, M. R. Kabir, Z. Zamal and R. M. Rahman, Correlation analysis of demographic factors on low birth weight and prediction modeling using machine learning techniques, *The 4th World Con. Smart Trends System Security Sustainability*, pp.169-173, 2020.

[47] W. Wei, X. Zhang and L. Yang, Full-cycle state evaluation of S700K switch machine based on residual network and fuzzy clustering, *International Journal of Innovative Computing, Information and Control*, vol.18, no.4, pp.1203-1216, 2022.

Author Biography

Stefanus Pieter Manongga received a B.S. degree in Livestock Farming from Nusa Cendana University in 1983. He was accepted to work at the Cooperative Regional Office of East Nusa Tenggara Province, specifically at the North Central Timor Regency Cooperative Development Center. Due to his strong desire to become a lecturer, at the end of 1985, he took the entrance test for lecturers at Nusa Cendana University (Undana) Kupang. On January 1, 1986, he was appointed as a civil servant at the Undana Faculty of Agriculture, which in 1988 was transferred to the Undana Faculty of Livestock Farming. In 2005 he was transferred to the Faculty of Public Health Undana until now. His research interests are healthcare system, recommendation system, and public health.

Hendry Hendry received a B.S. degree in Information Technology from Technology School of Surabaya, in 2005, an M.S. degree in Information Technology from 10 November Institute of Technology, Surabaya, in 2009, and a Ph.D. degree in Information Management from Chaoyang University of Technology in 2018. From 2012-2014 he was the Director of the Business and Technology Incubator at Satya Wacana Christian University. He is now a Lecturer in Faculty of Information Technology, Satya Wacana Christian University, Central Java, Indonesia. His research interests include domain ontology, recommendation systems, knowledge engineering, and applications of artificial intelligence.

Daniel Herman Fredy Manongga is a Professor of Computer Science at the Faculty of Information Technology, Satya Wacana Christian University (SWCU). He earned a Bachelor's degree in Electronics from the Faculty of Electrical Engineering, SWCU, in 1980; a Master's degree in Information Technology from Queen Mary College, University of London, in 1989; and a Ph.D. degree in Management Sciences from the School of Management (formerly School of Information Systems), the University of East Anglia, Norwich, UK, 1996. His research interests include artificial intelligence application, machine learning and knowledge engineering.