Classification of Big Data Stunting Using Support Vector Regression Method at Stella Maris Medan Maternity Hospital

[1]Kelvin, [2]R. A. Fattah Adriansyah, [3]Carles Juliandy,
[4]Frans Mikael Sinaga, [5]Frederick Liko, [6]Aswin Angkasa
[1,2,4,5,6]Informatics Engineering, [3]Information Technology, Mikroskil University, Indonesia
Email: [1]kelvin.chen@mikroskil.ac.id, [2]fattah.adriansyah@mikroskil.ac.id, [3]carles.juliandy@mikroskil.ac.id,
[4]frans.sinaga@mikroskil.ac.id, [5]221113801@students.mikroskil.ac.id, [6]221113724@students.mikroskil.ac.id

Article Info

Article history:
Received Jun 27th, 2024
Revised Jul 20th, 2024
Accepted Jul 25th, 2024

Keyword:
Big Data
Medical
Stunting
Support Vector Regression
Voting Classifier

ABSTRACT

This study aims to classify big data related to stunting using the Support Vector Regression (SVR) method at Stella Maris Maternity Hospital, Medan. Stunting, a condition of impaired growth in children due to chronic malnutrition and repeated infections, affects physical and cognitive development. With increasing health data, big data processing methods are essential for accurate information. SVR was chosen for handling high-dimensional and non-linear data, providing precise results. The study uses medical information, nutritional history, and socio-economic factors collected from hospital patients. The research process includes data collection, pre-processing to address missing values and outliers, normalization, and SVR application. Final results use SVR with Voting Classifier combining Support Vector Classifier (SVC), Random Forest (RF), and Gradient Boosting (GB), achieving an accuracy of 91.67%. This approach effectively identifies main stunting factors, aiding clinical decision-making and intervention programs. The study showcases big data and machine learning's potential in healthcare, serving as a model for improving health services and monitoring children's health conditions.

Corresponding Author: (10 pt)
Frans Mikael Sinaga,
Informatics Engineering,
Mikroskil University,
Email: frans.mikael@mikroskil.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v7i2.31112

1. INTRODUCTION

Child Stunting is a condition when children are unable to develop properly before birth. In comparison, stunted children are found to have a smaller body size compared to other children of the same age [1]. Child stunting is one of the most reliable pieces of evidence in the society that distinguishes welfare between families in different levels of social status. In Indonesia, child stunting was considered high, although decreasing gradually, throughout the past decades, which was about 37% of the children nationwide in 2013 and hit approximately 30% in 2018 [2], [3]. Some of the suspected factors are bad practices of breastfeeding and complementary feeding, deterioration of nutritional status in mothers, infection, and other supporting factors [2], [4]. According to the Nutritional Status Survey of Indonesia (SSGI), the ratio level of stunting issues was targeted to decline from 24.4% to 14% in 3 years starting from 2021 [3]. To countermeasure stunting issues, year to year targets have been set separately in several regions depending on their status.

The indicator for specific stunting issues can be assessed to fetus, infants, teenagers, and even pregnant mothers [3]. One of the general indicators is the measurement of length or height of a person (for children under 2 years old, their recumbent length is taken as their measurement) [2], [5]. World Health Organization (WHO) agreed on several standards to identify whether a child is considered stunted or normal. Children with their measurement beneath -2 SDs from their growth standards median of their age and gender are categorized as stunted and severely stunted if their measurement is beneath -3 SDs [4], [5]. Other indicators like Anemia

Screening and Antenatal Care (ANC) are conducted on older individuals such as young female teenagers and pregnant mothers [3]. However, to identify the cause of child stunting and its pattern more accurately, classification towards Big Data using Support Vector Machines (SVMs) is required.

SVM is a method used for problem solving like predictions and pattern recognitions. The advantages of using SVMs are its high accuracy and less requirement of data samples to prevent overfitting [6]. There are two forms of SVM; Support Vector Regression (SVR) and Support Vector Classification (SVC). SVR is often suggested for its higher accuracy in predicting compared to other similar models. It is capable of learning from a smaller training dataset and handle many variables at once [7]. After prediction is done, the results are then compared with a reference, like the actual value, and later assessed with performance measurements like the Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and coefficient of determination (R2) [7]. Normally, people would find issue when comparing a large number of data based on their value. Thus, categorizing data is easier for them to obtain a better view for necessary comparison as it is simple and can be easily understood [7], [8]

Although categorizing data is perceived to be easier for people to conduct an “apple to apple” comparison between data, the slightest mistake in categorizing data may cause misleading in decision making. To avoid miscategorization, a framework for classifying data like the SVC is provided [7]. While SVR returns only numerical values which may lead to miscategorization, SVC can be used to help assess performances based on their classification.

This research focuses on using Big Data to classify stunting in Indonesia, particularly in North Sumatra, by examining factors such as nutritional status, environmental conditions, access to health facilities, and socio-economic indicators. The primary objective is to leverage the SVR method for Big Data analysis, aiming to uncover significant patterns and accurately predict the risk of stunting among children in local communities. By integrating these insights, the study seeks to provide a deeper understanding of how these factors interplay and contribute to stunting prevalence, thereby informing targeted interventions and policy decisions in public health.

In addition to SVMs, the study incorporates advanced machine learning techniques like Random Forest, Gradient Boosting, and SVC to enhance predictive accuracy and interpretability. Random Forest, known for its capability to handle large, complex datasets and mitigate overfitting, constructs multiple decision trees during training [11]. Gradient Boosting sequentially improves predictive performance by refining weak learners, effectively capturing intricate relationships within the data [11]. SVC, on the other hand, focuses on optimal hyperplane determination to classify different classes in high-dimensional spaces, ensuring robust outcomes even with nonlinear data [11]. In line with this approach, Kelvin et al. also utilize a combination of several machine learning techniques for data prediction [31][32]. This combination of methods demonstrates significant potential in enhancing prediction accuracy and reliability, which is relevant for various applications, including the analysis of child stunting.

To further bolster prediction accuracy and reliability, the study adopts a Voting Classifier ensemble method, amalgamating predictions from SVM, Random Forest, and Gradient Boosting models [11]. This approach not only maximizes predictive power but also minimizes biases inherent in individual models, thereby providing more robust insights into stunting dynamics and facilitating evidence-based interventions. Previous research has demonstrated SVR's superiority in predicting uncertain variables compared to alternative methods [11]. In contrast to recent studies like [24], which explore similar themes, this research introduces a novel approach integrating Principal Component Analysis (PCA) and Synthetic Minority Over-sampling Technique (SMOTE) into its methodology. This innovative strategy aims to further enhance prediction accuracy and model resilience in capturing nuanced aspects of stunting prevalence in diverse socio-economic contexts.

2. RESEARCH METHOD

This research will be conducted within multiple stages: data collecting, data research, research flow process and stages, data analysis, model validation, performance evaluation, and lastly data testing. In the process of data collecting, there will be two approaches; qualitative and quantitative. As for the process of data analysis will be segmented into four steps; data collecting, data pre-processing, feature selection, and SVR model training.

2.1. Data Collection

Data were gathered from patients at Stella Maris Maternity Hospital, including medical records, nutritional history, and socio-economic information. This comprehensive dataset provided a robust foundation for analysis.

2.2. Data Pre-processing

1. Addressing Missing Values and Outliers: Initial steps involved cleaning the data by handling missing values and outliers to ensure the integrity of the dataset.
2. Data Normalization: The data were normalized to bring all variables to a comparable scale, enhancing the performance of the SVR method.

2.3. Support Vector Regression (SVR)

The SVR method was applied to the pre-processed data to perform the regression analysis. SVR is particularly suitable for high-dimensional and non-linear datasets, making it an ideal choice for this study [9]. The steps of the algorithm are as follows:

Step 1: Formulation of the SVR Problem:

Given a training dataset $\{(x_i, y_i)\}_{i=1}^{n}$where x_i is the input vector and $y_i$is the target value, the goal of SVR is to find a function $f(x)$ that approximates the relationship between x and y.

Step 2: Primal Problem:
The SVR aims to solve the following optimization problem:

$$\frac{min}{w,b,\xi,\xi^*}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{1}$$

Subject to:

$$y_i - (w \cdot \phi(x_i) + b) \leq \epsilon + \xi_i \tag{2}$$

$$(w \cdot \phi(x_i) + b) - y_i \leq \epsilon + \xi_i^* \tag{3}$$

$$\xi_i, \xi_i^* \geq 0 \tag{4}$$

Here, **w** is the weight vector, b is the bias term, $\xi_i \; and \; \xi_i^*$ are slack variables that allow some errors, ϵ is the margin of tolerance, C is a regularization parameter, and $\phi(x)$ is a feature transformation function.

Step 3: Dual Problem:
By introducing Lagrange multipliers α_i and α_i^* for the constraints, we can formulate the dual problem:

$$\frac{min}{\propto,\alpha^*}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) + \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \tag{5}$$

Subject to:

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \tag{6}$$

$$0 \leq \alpha_i\alpha_i^* \leq C \tag{7}$$

Here, $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ is the kernel function.

Step 4 : Decision Function :
Once the dual problem is solved, the decision function $f(x)$ can be expressed as:

$$f(x) = \sum_{i=1}^{n}(\alpha_i\alpha_i^*)K(x_i, x) + b \tag{8}$$

Here, b can be computed from the primal constraints using the support vectors (data points for which α_i and α_i^* are non-zero).

Step 5: Bias Term b:
To compute the bias term b, you can use any support vector x_i:

$$b = y_i - \sum_{j=1}^{n}(\alpha_j - \alpha_j^*)K(x_j, x_i) \tag{9}$$

This calculation can be averaged over all support vectors to improve robustness.

2.4. Model Integration

To improve the robustness and accuracy of the predictions, a Voting Classifier was implemented. This classifier combined three different models:

1. **Support Vector Classifier (SVC)**

   SVC is well-known for its effectiveness in classification tasks, especially with complex datasets [9]. According to its property, data can be classified into linearly separable data and also non-linear separable data [14].

   a. Linear Classification

   Linear Classification is ideally used for two data categories that can be separated with a single linear line, assuming the data variables are spanned on a 2-dimensional plane [14]. However, Linear Classification is also applied in different types of data classification besides the standard linear separation due to its rapid testing rate [15]. In Figure 1, the variables are separated with a single line according to their categories. Two supporting lines are drawn parallel towards the previous line to constrain the variables, called the support vectors. The distance between them are known as the margin [14]. To achieve maximum result, larger margin is more preferrable.

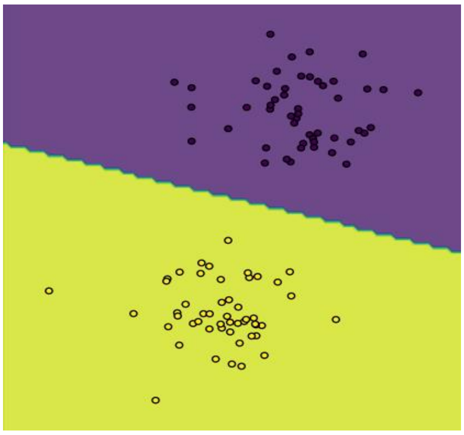

Figure 1. Two Dimensional Classification [14]

   b. Non-Linear Classification

   Non-Linear Classification is used to separate variables that are scattered unevenly (as shown in Figure 2), requiring non-linear divider to obatin a better and more accurate result [14]. Non-Linear Classification is proven to be more realistic as it is plausible to obtain analyses where two categoriies of clustered variables can be separated effortlessly with a line [14]. To calculate the margin between two supporting vectors in this case, a higher-dimensional plane is required to perform the variables separation.

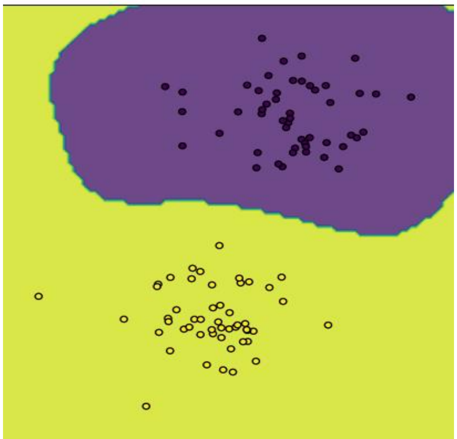

Figure 2. Multi Dimensional Classification [14]

The steps of the SVC algorithm are:

Step 1 : Decision Function and Hyperplane
The decision function for a linear SVC can be represented as:

$$f(x) = w^T x + b \tag{10}$$

In which x is the input vector (a data point), w is the weight vector and b is the bias term.
The hyperlane equation (decision boundary) is:

$$w^T x + b = 0 \tag{11}$$

Step 2 : Margin and Support Vectors
The margin γ between the two parallel hyperlanes (positive and negative class margins) is:

$$\gamma = \frac{2}{\|w\|} \tag{12}$$

Where $\|w\|$ is the Euclidean norm of the weight vector W.

Step 3 : Optimization Objective
The objective is to increase the margin to its limit while reducing classification errors, typically formulated as:

$$min_{w,b} \frac{1}{2} \|w\|^2 \tag{13}$$

Subject to:

$$y_i(w^T x_i + b) \geq 1 \, for \, all \, i \tag{14}$$

Where y_i is the class label (±1) of the i-th data point x_i.

Step 4 : Lagrangian and Dual Problem
The Lagrangian $\boldsymbol{\mathcal{L}}$**(w,b,∝)** incorporating Lagrange multipliers ∝ for the constraints:

$$\boldsymbol{\mathcal{L}}(\boldsymbol{w}, \boldsymbol{b}, \propto) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \propto_i \; [y_i(w^T x_i + b) - 1] \tag{15}$$

Step 5 : Dual Problem and Support Vector Calculation
Maximize the dual problem

$$max_{\propto}[\sum_{i=1}^{n} \propto_i - \frac{1}{2}\sum_{i,j=1}^{n} \propto_i \propto_j \; y_i y_j x_i^T x_j] \tag{16}$$

Subject to:

$$\propto_i \geq 0 \; and \; \sum_{i=1}^{n} \propto_i \; y_i = 0. \tag{17}$$

The weight vector **W** can be expressed as:

$$w = \sum_{i=1}^{n} \propto_i \; y_i x_i \tag{18}$$

Step 6 : Kernel Trick
For non-linear SVC using kernels like polynomial or RBF, transform the feature space:

$$K(x_{i,} x_j) = \phi(x_i)^T \phi(x_j) \tag{19}$$

Where $\phi(\boldsymbol{x})$ maps **x** into a higher-dimensional space

Step 7 : Prediction
To predict the class of a new data point x_{new}:

$$\hat{y} = sign\ (\sum_{i=1}^{n} \propto_i\ y_i K(x_i, x_{new}) + b \quad (20)$$

where $\propto_i$ are the support vector coefficients.

2. **Random Forest (RF)**
A versatile model that uses multiple decision trees to improve predictive performance and control overfitting [10], [16], [18]. This classifier is widely used for a high accuracy classification and speed of execution [17], [18]. Random Forest requires characteristics learning of the categories and identifications towards new uncategorized data [16], [18]. RF can be used to rate variables in order that has the highest potential in distinguishing between categories [16]. This classifier can identify data with non-linear pattern [18]. Equations for Random Forest:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x) \quad (21)$$

Where:

- $\hat{y}$ is the prediction output.
- N is the number of trees.
- $T_i(x)$ Is the prediction from tree i.

3. **Gradient Boosting (GB)**
A powerful ensemble method that builds models sequentially to correct the errors of previous models, enhancing overall prediction accuracy [10]. Gradient Boosting remains one of the most useful and competitive methods in machine learning field [19]. GB is proven to perform well in a setting with more variables compared to the samples [20]. GB is preferably opted for smaller scope of data, which requires additional steps when used for higher-dimensional data especially when dealing with anomalies or rare occurance event [20], [21]. Equations for Gradient Boosting:

$$F_m(x) = F_{m-1}(x) + h_m(x) \quad (22)$$

Where:

- $F_m(x)$ is the model at iteration m.
- $h_m(x)$ is the base learner paired with residuals from the previous model.

The Voting Classifier aggregated the strengths of these models to produce a more reliable prediction.

4. **Principal Component Analysis (PCA)**
Principal Component Analysis (PCA) is a statistical technique employed to reduce the dimensions of data while preserving maximal variance. PCA identifies principal components—directions in the data with the highest variance—and projects the data onto these components [29]. The steps involved in PCA are as follows:

Step 1 : Standardize the Data : Scale the data such that each feature has an average of 0 and a standard deviation of 1.

$$Z = \frac{X - \mu}{\alpha} \quad (23)$$

where X is the original data, μ is the mean, and $\propto$ is the standard deviation.

Step 2 : Covariance Matrix: Compute the covariance matrix to understand the relationships between features in the data.

$$C = \frac{1}{n-1} Z^T Z \quad (24)$$

Where C is the covariance matrix, Z^T is the transpose of the standardized data, and n is the number of samples.

Step 3 : Eigenvalues and Eigenvectors: Determine the eigenvalues and eigenvectors of the covariance matrix. Eigenvectors indicate the principal component directions, while eigenvalues quantify the variance along these directions.

$$C_v = \lambda_v \tag{25}$$

Step 4 : Sort and Select Principal Components: Arrange the eigenvalues in descending order and select the eigenvectors associated with the largest eigenvalues to identify the principal components for use.

Step 5 : Project the Data: Project the original data onto the selected principal components.

$$Y = ZV \tag{26}$$

where Y is the projected data, Z is the standardized data, and V is the matrix of selected eigenvectors.

5. Synthetic Minority Over-sampling Technique (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) is employed to tackle class imbalance in datasets. SMOTE generates synthetic samples of the minority class by interpolating between existing minority class instances [30]. The steps involved in SMOTE are as follows:

Step 1: Selection of Sample and Nearest Neighbours : For each minority sample x_i, choose one of its k nearest neighbors x_{in}.

Step 2: Interpolation: Create new synthetic samples by interpolating between the selected sample and its nearest neighbours.

$$x_{new} = x_i + \theta \times (x_{in} - x_i) \tag{27}$$

Where θ is a random number between 0 and 1, X_i is the original sample, and x_{in} is one of its nearest neighbors.

6. Ensemble Learning SVR with Voting Classifier (SVC, RF, dan GB)

Ensemble Learning has become important for improving prediction accuracy by combining various models. This approach is particularly relevant in addressing complex issues such as childhood stunting, as demonstrated in various studies.

a. Support Vector Regression (SVR): SVR, based on Support Vector Machines (SVM), is effective in predicting continuous outcomes such as prevalence rates, as shown in studies by Gaffar et al. (2023) and Paniagua-Tineo et al. (2011) [23], [24].

b. Voting Classifier: The Voting Classifier aggregates predictions from multiple models to make collective predictions, reducing bias and variance. It encompasses models such as Support Vector Classifier (SVC), Random Forest (RF), and Gradient Boosting (GB), each providing unique predictive strengths [25].

c. Components of Voting Classifier:
   1) Support Vector Classifier (SVC): SVC, which is an SVM model typically used for classification, can also be used as a component in the Voting Classifier to make predictions [26].
   2) Random Forest (RF): RF is an ensemble technique based on decision trees, effective for both classification and regression tasks. RF provides diverse predictions compared to SVC [27].
   3) Gradient Boosting (GB): Gradient Boosting (GB) is an ensemble learning technique where models are trained sequentially, with each subsequent model designed to refine the errors of the preceding one. This iterative process aims to continuously enhance predictive accuracy throughout the model training process. GB usually delivers accurate and robust predictions in many scenarios [28].

Benefits and Applications:

1. Improved Accuracy: By combining SVR with classifiers like RF and GB in the Voting Classifier, the ensemble approach can enhance prediction accuracy, crucial for identifying risk factors contributing to stunting [6].
2. Reduced Overfitting: Ensemble methods reduce overfitting by incorporating models with different biases and variances, ensuring strong predictions across various datasets and contexts [7].

Conclusion: Integrating SVR into the Voting Classifier enhances predictive capabilities, particularly in complex scenarios such as childhood stunting analysis. This approach leverages the strengths of individual models to provide more accurate and reliable predictions, supporting informed decision-making in public health interventions.

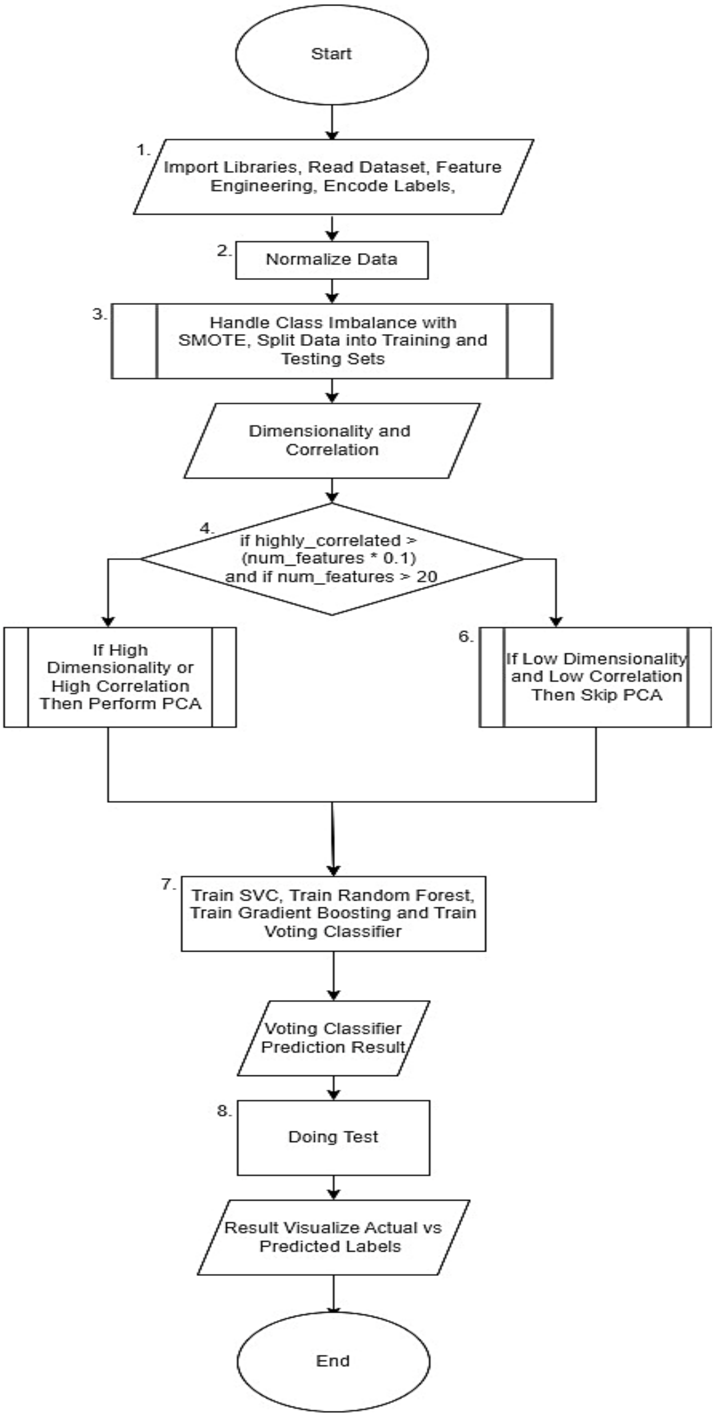
Start
1.
Import Libraries, Read Dataset, Feature Engineering, Encode Labels,
2.
Normalize Data
3.
Handle Class Imbalance with SMOTE, Split Data into Training and Testing Sets
Dimensionality and Correlation
4.
if highly_correlated > (num_features * 0.1) and if num_features > 20
5.
If High Dimensionality or High Correlation Then Perform PCA
6.
If Low Dimensionality and Low Correlation Then Skip PCA
7.
Train SVC, Train Random Forest, Train Gradient Boosting and Train Voting Classifier
Voting Classifier Prediction Result
8.
Doing Test
Result Visualize Actual vs Predicted Labels
End

Figure 3. Research Method Flowchart

This research methodology involves eight main stages. The initial stage entails import the libraries, reading the dataset, identifying target features, Encode Labels, and determining the dimensionality and correlation. Subsequently, in the second stage, the dataset is normalized to scale the data within a range of 0 to 1. The third stage Apply SMOTE to balance the class distribution, and split the dataset for training and testing. Following this, in the fourth stage, checking the limitation of dimensionality and correlation. Moving on to the fifth stage, High dimensionality and correlation are forecasted using PCA (Principal Component Analysis) to reduce the dimensionality and correlation of a dataset, while Low dimensionality and correlation are predicted by not using PCA. Then Train all of the model AI that we used which is SVC (support vector classifier), RF (random forest), GB (Gradient Boosting) and Voting Classifier. Subsequently, testing is conducted to generate predictions for the target features into a scatter plot at Figure 5 and Figure 6, as illustrated in the flowchart depicted in Figure 3.

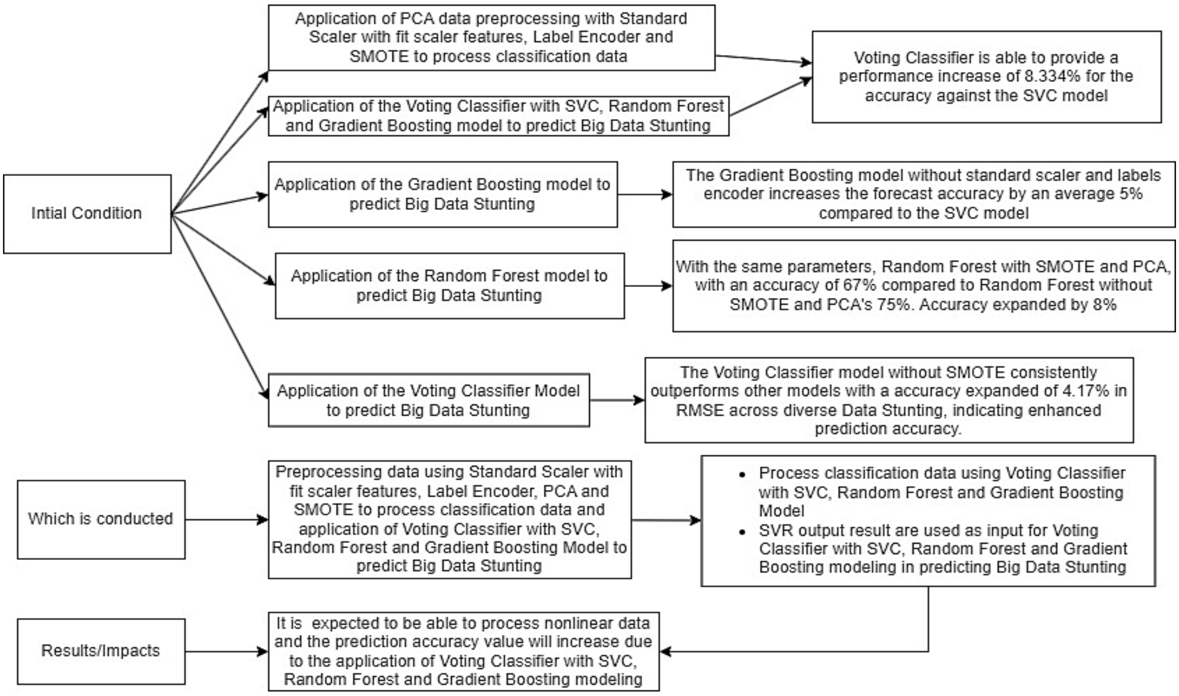
Application of PCA data preprocessing with Standard Scaler with fit scaler features, Label Encoder and SMOTE to process classification data
Voting Classifier is able to provide a performance increase of 8.334% for the accuracy against the SVC model
Application of the Voting Classifier with SVC, Random Forest and Gradient Boosting model to predict Big Data Stunting
Intial Condition
Application of the Gradient Boosting model to predict Big Data Stunting
The Gradient Boosting model without standard scaler and labels encoder increases the forecast accuracy by an average 5% compared to the SVC model
Application of the Random Forest model to predict Big Data Stunting
With the same parameters, Random Forest with SMOTE and PCA, with an accuracy of 67% compared to Random Forest without SMOTE and PCA's 75%. Accuracy expanded by 8%
Application of the Voting Classifier Model to predict Big Data Stunting
The Voting Classifier model without SMOTE consistently outperforms other models with a accuracy expanded of 4.17% in RMSE across diverse Data Stunting, indicating enhanced prediction accuracy.
Which is conducted
Preprocessing data using Standard Scaler with fit scaler features, Label Encoder, PCA and SMOTE to process classification data and application of Voting Classifier with SVC, Random Forest and Gradient Boosting Model to predict Big Data Stunting
Process classification data using Voting Classifier with SVC, Random Forest and Gradient Boosting Model
SVR output result are used as input for Voting Classifier with SVC, Random Forest and Gradient Boosting modeling in predicting Big Data Stunting
Results/Impacts
It is expected to be able to process nonlinear data and the prediction accuracy value will increase due to the application of Voting Classifier with SVC, Random Forest and Gradient Boosting modeling

Figure 4. Problem Solving Framework

The problem-solving framework is presented in Figure 4. The process commences with data preprocessing using a machine learning processing technique known as SMOTE and PCA on standard scaler features and label encoder to process input data. Subsequently, the output from preprocessing data is utilized as input for the Voting Classifier with SVC, Random Forest and Gradient Boosting model to predict Big Data Stunting. VC (SVC, RF, and GB) with SMOTE and PCA demonstrates a performance increase of 8.334% for the accuracy compared to the SVC model. Previous research on the Gradient Boosting model without standard scaler and labels encoder used for big data stunting prediction shows an average accuracy increase of 5% compared to the SVC model. The Random Forest with SMOTE and PCA model successfully improves accuracy compared to the SVC model, as evidenced by a upgrade accuracy value of 75%. The inclusion of the Attention mechanism in the model not only enhances information processing efficiency but also Balancing Class Distribution and Noise Reduction within SMOTE and PCA, making the model more robust and reliable in big data stunting prediction.

Based on these initial conditions, the intended research involves using SMOTE and PCA for data preprocessing. The resulting output from SMOTE and PCA is then used as input for VC, with the addition of the SVC, RF and GB, to predict the big data stunting. This combination of four methods has not yet been evaluated for its capacity to retain both historical and future data while capturing supplementary information that bolsters prediction outcomes. The data utilized in this study comprises secondary data sourced from Stella Maris Maternity Hospital in Medan, Indonesia. Comprising historical data on Stella Maris Maternity Hospital from 0 years old until 60 years old. The attributes used in this study are body weight (BB) and Height (TB) with its Z-score, as these attributes are highly correlated with the output results of the stunting classify. In this study, we compared the SVC model with the VC (SVC, RF, and GB) model (the proposed model). We have demonstrated that the performance of the proposed model is superior compared to the other models examined. This comparison is illustrated in Figure 4, which depicts the Problem Solving Framework.

3. RESULTS AND ANALYSIS

In this experiment, we aimed to identify the model with the highest accuracy across various datasets, using identical parameters as detailed in Table 1. The parameters that used in this experiment was

1. C: Regularization parameter that controls the trade-off between achieving a large decision margin and
2. correctly classifying data points.
3. Degree: Degree of the polynomial used in polynomial kernel.
4. Gamma: Kernel parameter that determines how far the influence of a single data point reaches.
5. Kernel: Kernel function used in SVM.
6. Criterion (RF): Measure of quality for splits in Random Forest (in this case, Gini impurity).
7. Max Depth (RF): Maximum depth of each tree in Random Forest.
8. Max Features (RF): Maximum number of features considered for the best split in Random Forest.
9. N estimators (RF): Number of trees in Random Forest.
10. Learning Rate (GB): Learning rate for Gradient Boosting.
11. Max Depth (GB): Maximum depth of each tree in Gradient Boosting.
12. N estimators (GB): Number of trees in Gradient Boosting.

The comparison results, detailed in Table 2, highlight the remarkable performance of the Voting Classifier that integrates SVC (Support Vector Classifier), Random Forest (RF), and Gradient Boosting (GB) models, achieving an impressive accuracy of **91.67%** (Figure 6). This outcome not only demonstrates the Voting Classifier's robust predictive capabilities but also underscores its potential to significantly enhance decision-making processes in addressing complex issues such as child stunting. In contrast, the SVM model utilizing SVC achieved an accuracy of 83.33% (Figure 5), indicating a notable difference in predictive accuracy favoring the Voting Classifier approach in this research context.

Table 1. Parameters used in model testing

Model	SVM with SVC	Voting Classifier (SVC, RF, GB)
Desc.		Proposed Model
C	1	1
Degree	2	2
Gamma	scale	scale
Kernel	linear	linear
Criterion (RF)	-	gini
Max Depth (RF)	-	4
Max Features (RF)	-	auto
N estimators (RF)	-	100
Learning Rate (GB)	-	0.05
Max Depth (GB)	-	4
N estimators (GB)	-	500

Table 2. Result of model testing

Model	SVM with SVC	Voting Classifier (SVC, RF, GB)
Desc.		Proposed Model
ACCURACY	83.33%	91.67%

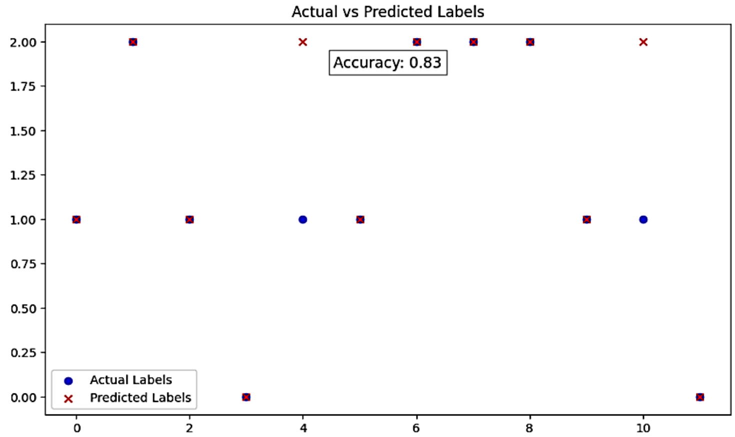

Figure 5. SVM with SVC

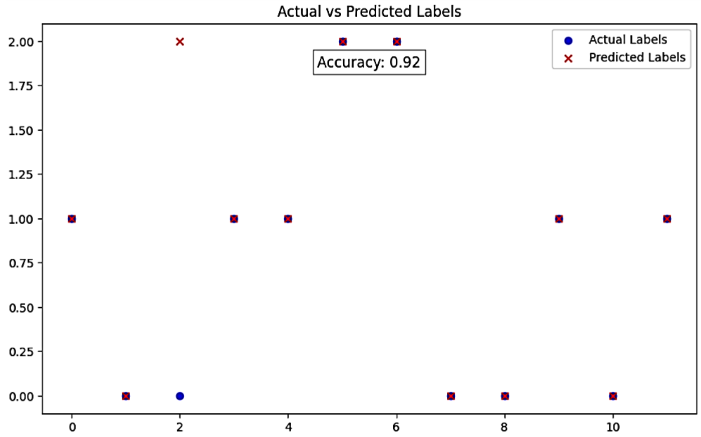

Figure 6. Voting Classifier (SVC, RF, GB)

4. CONCLUSION

This study successfully demonstrates the application of Support Vector Regression (SVR) and a Voting Classifier to classify big data related to stunting at Stella Maris Maternity Hospital, Medan. The integration of SVR with a Voting Classifier that combines Support Vector Classifier (SVC), Random Forest (RF), and Gradient Boosting (GB) models yielded a high accuracy of 91.67%. This approach effectively identified the main factors contributing to stunting, highlighting the importance of medical information, nutritional history, and socio-economic factors in understanding and addressing this condition.

The findings suggest that utilizing advanced machine learning techniques on large datasets can significantly enhance the accuracy and reliability of stunting classifications. This, in turn, supports better clinical decision-making and the development of more effective intervention programs aimed at reducing stunting prevalence among children. The successful implementation of these methods in a healthcare setting underscores the potential of big data and machine learning technologies to transform public health monitoring and service delivery. This model can be adopted by other healthcare institutions to improve health outcomes and monitor children's health conditions more effectively.

ACKNOWLEDGEMENTS

We are deeply grateful to the Directorate of Research, Technology, and Community Service and the Ministry of Education, Culture, Research, and Technology for their significant funding support.

REFERENCES

[1] S. Angriani, N. Jalil, S. Aminah, and N. Agus Salim, “Childhood Stunting: Analysis Affecting Children’s Stunting In Sulawesi,” 2021.

[2] T. Beal, A. Tumilowicz, A. Sutrisna, D. Izwardy, and L. M. Neufeld, “A review of child stunting determinants in Indonesia,” Maternal and Child Nutrition, vol. 14, no. 4. 2018. doi: 10.1111/mcn.12617.

[3] S. Processing, “Penyelenggaraan Percepatan Penurunan Stunting,” Signal Processing, 2009.

[4] M. de Onis and F. Branca, “Childhood stunting: A global perspective,” Maternal and Child Nutrition, vol. 12. 2016. doi: 10.1111/mcn.12231.

[5] T. Siswati, B. A. Paramashanti, N. Pramestuti, and L. Waris, “A POOLED DATA ANALYSIS TO DETERMINE RISK FACTORS OF CHILDHOOD STUNTING IN INDONESIA,” Journal of Nutrition College, vol. 12, no. 1, 2023, doi: 10.14710/jnc.v12i1.35413.

[6] J. T. Samudra, R. Rosnelly, and Z. Situmorang, “Comparative Analysis of SVM and Perceptron Algorithms in Classification of Work Programs,” MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 22, no. 2, 2023, doi: 10.30812/matrik.v22i2.2479.

[7] M. H. Bazrkar and X. Chu, “Development of category-based scoring support vector regression (CBS-SVR) for drought prediction,” Journal of Hydroinformatics, vol. 24, no. 1, 2022, doi: 10.2166/HYDRO.2022.104.

[8] Y. Zhang, “Support vector machine classification algorithm and its application,” in Communications in Computer and Information Science, 2012. doi: 10.1007/978-3-642-34041-3_27.

[9] C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, vol. 20, no. 3, 1995, doi: 10.1023/A:1022627411411.

[10] J. C. Platt, “Fast Training of Support Vector Machines Using Sequential Minimal Optimization,” in Advances in Kernel Methods, 2022. doi: 10.7551/mitpress/1130.003.0016.

[11] L. Breiman, “Random forests. Machine Learning,” Kluwer Academic Publishers. Manufactured in The Netherlands., vol. 45(1), 2001.

[12] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Annals of Statistics, vol. 29, no. 5, 2001, doi: 10.1214/aos/1013203451.

[13] L. I. Kuncheva, Combining Pattern Classifiers. 2004. doi: 10.1002/0471660264.

[14] H. Bhavsar and M. H. Panchal, “A Review on Support Vector Machine for Data Classification,” International Journal of Advanced Research in Computer Engineering & Technology, vol. 1, no. 10, 2012.

[15] V. K. Chauhan, K. Dahiya, and A. Sharma, “Problem formulations and solvers in linear SVM: a review,” Artificial Intelligence Review, vol. 52, no. 2. 2019. doi: 10.1007/s10462-018-9614-6.

[16] M. Belgiu and L. Drăgu, “Random forest in remote sensing: A review of applications and future directions,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 114. 2016. doi: 10.1016/j.isprsjprs.2016.01.011.

[17] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, “An assessment of the effectiveness of a random forest classifier for land-cover classification,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 67, no. 1, 2012, doi: 10.1016/j.isprsjprs.2011.11.002.

[18] A. Chaudhary, S. Kolhe, and R. Kamal, “An improved random forest classifier for multi-class classification,” Information Processing in Agriculture, vol. 3, no. 4, 2016, doi: 10.1016/j.inpa.2016.08.002.

[19] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” Artificial Intelligence Review, vol. 54, no. 3, 2021, doi: 10.1007/s10462-020-09896-5.

[20] R. Blagus and L. Lusa, “Gradient boosting for high-dimensional prediction of rare events,” Computational Statistics and Data Analysis, vol. 113, 2017, doi: 10.1016/j.csda.2016.07.016.

[21] M. S. Islam Khan, N. Islam, J. Uddin, S. Islam, and M. K. Nasir, “Water quality prediction and classification based on principal component regression and gradient boosting classifier approach,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 8, 2022, doi: 10.1016/j.jksuci.2021.06.003.

[22] C. Y. Yeh, C. W. Huang, and S. J. Lee, “A multiple-kernel support vector regression approach for stock market price forecasting,” Expert Systems with Applications, vol. 38, no. 3, 2011, doi: 10.1016/j.eswa.2010.08.004.

[23] A. Paniagua-Tineo, S. Salcedo-Sanz, C. Casanova-Mateo, E. G. Ortiz-García, M. A. Cony, and E. Hernández-Martín, “Prediction of daily maximum temperature using a support vector regression algorithm,” Renewable Energy, vol. 36, no. 11, 2011, doi: 10.1016/j.renene.2011.03.030.

[24] A. W. M. Gaffar, Sugiarti, Dewi Widyawati, Andi Muhammad Kemai Arief Hidayat Paharuddin, and Andi Vania Anastasia, “Spatial Prediction of Stunting Incidents Prevalence Using Support Vector Regression Method,” Indonesian Journal of Data and Science, vol. 4, no. 2, 2023, doi: 10.56705/ijodas.v4i2.68.

[25] G. Kunapuli, Ensemble Methods for Machine Learning. 2023.

[26] A. Salini, U. Jeyapriya, S. M. College, and S. M. College, “A Majority Vote Based Ensemble Classifier for Predicting Students Academic Performance,” International Journal of Pure and Applied Mathematics, vol. 118, no. 24, 2018.

[27] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, “A survey on ensemble learning,” Frontiers of Computer Science, vol. 14, no. 2. 2020. doi: 10.1007/s11704-019-8208-z.

[28] I. D. Mienye and Y. Sun, “A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects,” IEEE Access, vol. 10. 2022. doi: 10.1109/ACCESS.2022.3207287.

[29] S. Mishra et al., “Multivariate Statistical Data Analysis- Principal Component Analysis (PCA),” International Journal of Livestock Research, vol. 7, no. 5, 2017.

[30] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, 2002, doi: 10.1613/jair.953.

[31] Kelvin, R., Purba, R., & Halim, A. (2022). Stock Price Prediction Using XCEEMDAN-Bidirectional LSTM-Spline. Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), 5(1), 1-12. https://doi.org/10.24014/ijaidm.v5i1.14424.

[32] Kelvin, Sinaga, F. M., Winardi, S., & Susmanto. (2024). Exploring New Frontiers: XCEEMDAN, Bidirectional LSTM, Attention Mechanism, and Spline in Stock Price Forecasting. Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), 7(2), 384-391. https://dx.doi.org/10.24014/ijaidm.v7i2.29649.

BIBLIOGRAPHY OF AUTHORS

Kelvin, S.Kom., M.Kom., The author is a Software Engineer and Lecturer in the Informatics Engineering, Faculty of Informatics, Mikroskil University, Medan. He completed his bachelor's degree in Informatics Engineering at STMIK Mikroskil in 2018. Then, in 2020, the author pursued postgraduate studies in Information Technology at Mikroskil University and successfully completed them in 2021. The courses he has taught include Introduction to Algorithms, Web Design, C Programming, Object-Oriented Programming, Back-End Web Development, Artificial Intelligence, and Natural Language Processing. In addition to his academic involvement, the author has over 5 years of experience as a software engineer, working for both domestic and international companies. For more information, visit the author's LinkedIn page at https://www.linkedin.com/in/kelvinchen96

Frans Mikael Sinaga, S.Kom., M.Kom., Lecturer at the Department of Informatics Engineering, Faculty of Informatics, Mikroskil University, Medan. Born in Penggalangan village on October 24, 1993. The author is the third child out of 4 siblings of Mr. Waristo and Mrs. Linda. The author completed a Bachelor's degree (S1) in Informatics Engineering and a Master's degree (S2) in Information Technology at STMIK Mikroskil Medan. The author has written several book titles such as Introduction to Computer Networks and Data Mining. In addition to writing books, the author has also conducted several research projects in the fields of Data Science and Computer Vision.

Carles Juliandy, S.Kom., M.Kom., The author is a lecturer at the Department of Information Technology, Faculty of Informatics, Universitas Mikroskil, Medan. He completed Bachelor's degree in Informatics Engineering at STMIK TIME and Master's degree in Informatics Technology at STMIK Mikroskil. The author completed several research especially in Blockchain technology and computer vision.

R. A. Fattah Adriansyah, S.Kom., M. Kom., Lecturer in the Department of Informatics Engineering, Faculty of Informatics, Mikroskil University, Medan. He graduated with a Bachelor of Computer Science at USU and a Masters in Computer Science at UNSRI, and has also completed several research studies, especially in the field of Computer Networks.

Aswin Angkasa, Student at the Faculty of Informatics, Mikroskil University. The author received his Diploma of Engineering at Macquarie University, Sydney in 2021 and is currently pursuing his Bachelor's Degree in Computer Science in the 4th semester.

Frederick Liko, Student of Informatics Engineering, Faculty of Informatics, Mikroskil University, Medan. Born in Medan, Indonesia on January 1, 2005. The author is the only child of Mr. Irsadi Pali and Mrs.Meina. The author was an Ehipassiko Buddhis Teen Organization's Leader 2022-2024. He is currently pursuing his Bachelor's Degree in Computer Science in the 4th semester. He is a private teacher for computer programming as well. He teaches Scratch, Python Programming, Arduino Electronic, and EV3 Robotic. In addition, He also teach science subject such as, mathematic, physic and chemistry for senior high school student.