Journal of Soft Computing Exploration

Homepage: htttps://shmpublisher.com/index.php/joscex

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Comparison of supervised machine learning methods in predicting the prevalence of stunting in north sumatra province

Vinny Ramayani Saragih[1], Arnita[2], Zulfahmi Indra[3], Insan Taufik[4], Marlina Setia Sinaga[5]
[1,5]Department of Computer Science, Universitas Negeri Medan, Indonesia

Article Info

Article history:

Received Nov 23, 2024
Revised Dec 4, 2024
Accepted Dec 11, 2024

Keywords:

Stunted
Prevalence of Stunted
North Sumatra
Machine Learning
Supervised Learning

ABSTRACT

Stunting is a growth and development disorder in children caused by chronic malnutrition and repeated infections. Stunting has significant short- and long-term impacts and is one of the major health issues currently faced by Indonesia. Stunting in North Sumatra Province is 18.9%, and the provincial government aims to reduce this prevalence to 14% by 2024. Existing studies on stunting prevalence prediction often rely on a single machine learning method and limited data sources. This research compares the performance of Support Vector Regression (SVR), Decision Tree, and Random Forest models using secondary data from 2021 to 2023 across 33 districts/cities in North Sumatra. Evaluation metrics include Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). Results indicate that Random Forest provides the most accurate and consistent predictions, achieving lower MSE, MAE, RMSE, and MAPE values compared to the other models. Decision Tree performs well in some regions but shows higher errors in specific cases, while SVR exhibits more variable performance with higher prediction errors in several areas. The ensemble approach of Random Forest minimizes overfitting, ensuring stable accuracy across districts with diverse stunting patterns. It identifies critical predictors, such as low birth weight, access to safe drinking water, proper sanitation, HDI, antenatal care (K4), and postpartum vitamin A, making it an effective tool for guiding evidence-based stunting reduction policies in North Sumatra.

Corresponding Author:

Vinny Ramayani Saragih,
Department of Computer Science,
Universitas Negeri Medan,
Jl. Willem Iskandar Psr. V, Kenangan Baru, Kec. Percut Sei Tuan Kota Medan Sumatera Utara, Indonesia.
Email: vinnysaragih@mhs.unimed.ac.id
https://doi.org/10.52465/joscex.v5i4.498

1. INTRODUCTION

Presidential Regulation No. 72 of 2021 on Accelerating Stunting Reduction defines stunting as a disruption in a child's growth and development due to chronic malnutrition and recurrent infections. This condition is characterized by a height below the standards set by the Ministry of Health. Stunting not only affects a child's physical growth but also impacts cognition, learning ability, and long-term productivity [1].

The factors causing stunting are divided into direct and indirect causes. Direct causes include insufficient colostrum and exclusive breastfeeding, inadequate child feeding patterns, and frequent infections experienced by children, all of which affect their nutritional status and can lead to stunting. Meanwhile, indirect

factors include limited access and availability of food, as well as inadequate sanitation and an environment that does not support health [2].

The World Health Organization (WHO) has revealed that the global prevalence of stunting among children under five years old in 2022 is 22.3%, or 148.1 million children [3]. According to the Indonesian Ministry of Health's Nutritional Status Study report, the prevalence of stunting in Indonesia has decreased from 27.7% in 2019 to 24.4% in 2021 and further declined to 21.6% in 2022. Most of these cases occur in children aged 3–4 years (approximately 6%). Despite the improvement, this figure still does not meet the WHO standard, which sets a target below 20%. Therefore, the government is striving to further reduce the stunting rate, aiming for 17% by 2023 and 14% by 2024 [4].

The significant efforts and progress in reducing stunting in North Sumatra further justify its selection as the focus area for this study. According to the 2021 Indonesia Child Nutrition Status Survey (SSGI), the stunting prevalence in North Sumatra was 25.8%, slightly below the national average. In 2022, the rate decreased to 21.1%, and by 2023, it further dropped to 18.9%, marking a reduction of approximately 2.2% from the previous year. This aligns with the North Sumatra Provincial Government's target of reducing stunting prevalence to 18% by 2023, supported by a budget allocation of around Rp346 billion for specific and sensitive nutrition interventions. Furthermore, the provincial government has actively engaged stakeholders to collaborate in addressing stunting, aiming for a 14% reduction by 2024 [5]. In addition to the declining prevalence, North Sumatra's socio-economic and cultural diversity across districts/cities provides a rich context for exploring specific factors influencing stunting. This diversity enhances the study's applicability to designing targeted interventions. Moreover, the researcher's familiarity with the region being a resident of North Sumatra—enables a deeper understanding of local challenges and facilitates a more thorough and contextually relevant analysis. By focusing on this region, the study contributes to ongoing efforts to combat stunting and supports the achievement of both local and national development goals.

In previous research, M. Syauqi et al. compared several supervised learning algorithms, such as linear regression, SVR, and random forest regression, to predict stunting in toddlers. The results show that SVR provides the best modeling because it has the lowest MAE and MSE values [6]. Another study by Kartika et al. predicts stock prices in the COVID era using multiple linear regression, support vector regression, decision tree regression, and K-nearest regression methods, of which decision tree regression provides competitive results compared to other methods [7].

We cannot undervalue the issue of stunting. Stunting has significant impacts both in the short term and long term. Short-term effects include decreased immune system strength, increased risk of various diseases, higher morbidity and mortality rates, as well as impaired intellectual and cognitive abilities in individuals experiencing stunting. Meanwhile, the long-term effects of stunting include an increased risk of degenerative diseases in adulthood and a potential hindrance to human resources [8]. Not only does it impact individuals directly, but its long-term effects also impact the country's overall growth. Each country estimates that stunting causes losses of 2-3% of the Gross Domestic Product (GDP) every year [9].

Long-term predictions are necessary to estimate future stunting conditions and prevent related diseases, given the potential impact of stunting. Machine learning can perform disease predictions. Machine learning bases its predictions on historical data, which it processes into patterns to forecast future events [10].

Machine learning, a component of artificial intelligence, enables the creation of intelligent computer systems without requiring direct human rule determination. Training the system to recognize patterns in a dataset result in models that can predict values (regression) or data groups (classification) [11]. Machine learning can analyze large datasets to discover specific patterns if there is available data as input. In machine learning, there is training data [12].

Supervised learning is one of the machine learning techniques that utilizes labeled datasets (training data) to train the machine. This allows the machine to identify input labels and make predictions or classifications using its features. Several algorithms fall under the category of supervised learning, including linear regression, K-nearest neighbor (KNN), support vector machine (SVM), naive Bayes, random forest, decision tree, and many others [13].

This research aims to compare the effectiveness of three supervised machine learning methods in the context of predicting stunting prevalence based on prevalence data and stunting indicators in North Sumatra. The indicators used in this analysis reflect a range of health, socioeconomic, and environmental factors that influence the prevalence of stunting. These include breastfeeding practices, low birth weight rates, maternal health during and after pregnancy, immunization coverage, access to health services, and socioeconomic conditions such as poverty, safe drinking water, and sanitation. Each indicator is critical in understanding the multidimensional causes of stunting. Machine learning methods, such as random forests, SVR, and decision tree regression, are well suited for this analysis as they can capture complex and non-linear relationships among variables, handle different types of data, and identify key factors. The methods used are Support Vector Machine (SVM), Decision Tree, and Random Forest, chosen for their ability to handle complex non-linear relationships in the data. Support vector regression (SVR) uses kernel tricks to map data to a higher-

dimensional feature space. This model can accurately find non-linear patterns [6]. Decision Tree Regression recursively divides the dataset based on feature values, forming a tree that represents non-linear relationships between variables [14]. Random Forest Regression uses an ensemble method to combine several decision trees to make predictions more accurate and find different non-linear patterns in the data [15]. The combination of these three methods allows for a comprehensive comparison in predicting stunting prevalence, as each has a unique way of handling data complexity.

Observational techniques require mathematical calculations to measure prediction error rates. Many calculations can be used to determine prediction accuracy, with MSE and MAPE being standard measures. With a focus on stunting prevalence, this research aims not only to identify the best method but also to deepen understanding of the impact of stunting indicators on prevalence in North Sumatra.

2. METHOD

This research uses a quantitative approach, which emphasizes numerical data and their processing using statistical methods. Figure 2.1 is the flowchart of the study.

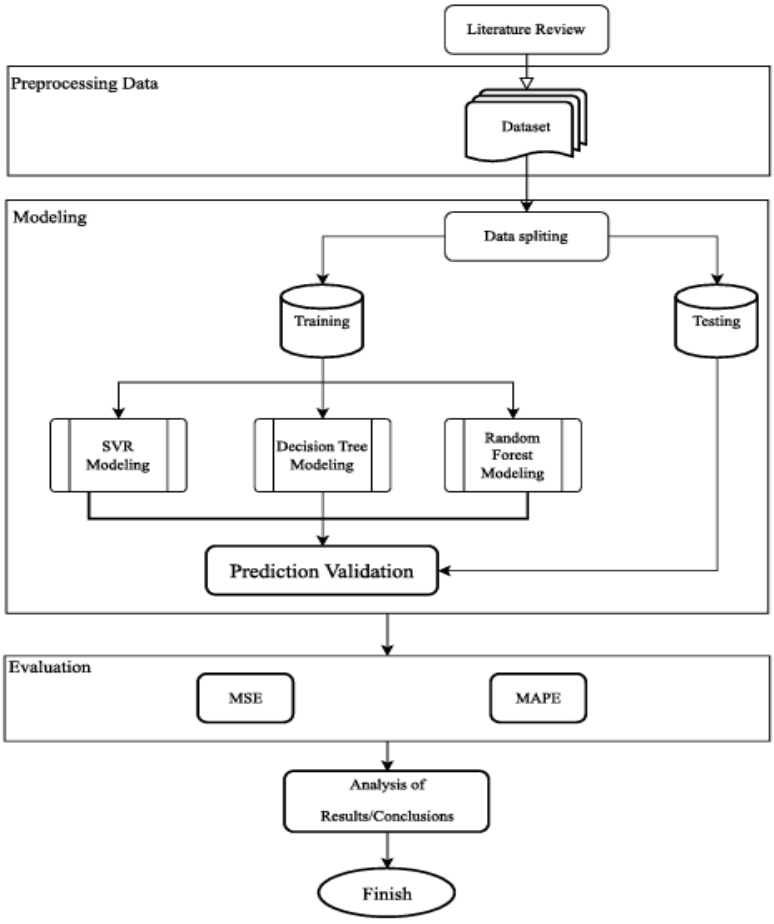

Figure 2.1 Flowchart of Research Methodology

The flowchart above illustrates the workflow of this research, starting with a literature review to explore relevant theories and methodologies. The process continues with data preprocessing to ensure the dataset's quality before analysis. The dataset is then split into training and testing sets, which are used to build and evaluate three machine learning models: support vector regression (SVR), decision tree, and random forest. The trained models are validated using testing data, and their performance is measured using mean squared error (MSE) and mean absolute percentage error (MAPE) metrics. Finally, the results are analyzed to determine the best-performing model, leading to the conclusion and recommendations for future applications.

2.1 The Data Prevalence of North Sumatra

This study utilizes secondary data, specifically the prevalence of stunting in North Sumatra, as well as data from health surveys conducted by several institutions.

1. Data on the prevalence of stunting and its indicators are available from the North Sumatra Health Office for the years 2021-2023.
2. Social demographic statistics data (North Sumatra Central Bureau of Statistics, 2021-2023).

This consists of 13 variables, one dependent variable (Y), and one independent variable (X), as listed in Table 2.1.

Table 2.1 Research Variables

Variable	Description
Y	Prevalence percentage of stunting in North Sumatra
X_1	Percentage of infants exclusively breastfed for 6 months
X_2	Percentage of Low Birth Weight (LBW) babies
X_3	Percentage of infants aged 6-59 months receiving Vitamin A
X_4	Percentage of postpartum mothers receiving Vitamin A
X_5	Percentage of pregnant mothers receiving Iron and Folic Acid Tablets (90 Tablets)
X_6	Percentage of Pregnant Women K1
X_7	Percentage of pregnant women with K4
X_8	Percentage of infants immunized
X_9	Percentage of Human Development Index (HDI)
X_{10}	Percentage of the population living in poverty
X_{11}	Percentage of households with access to safe drinking water
X_{12}	Percentage of households with access to proper sanitation

This table shows the factors that may influence the prevalence of stunting, including independent variables such as maternal and child health conditions, nutritional status, and socioeconomic factors such as poverty and the human development index. The compiled data is saved as a Comma Separated Value (CSV) file so that it can be processed using supervised machine learning tools, namely Python, using Google Colab.

2.2 Data Preprocessing

At this stage, before modeling, data preprocessing is a crucial step that helps improve the quality of the model. The first step is data cleaning, which involves checking for invalid or empty data. Next is scaling/normalization, which is important for algorithms sensitive to scale, such as SVR. This study uses standard scaling to align the range of values across variables. Standardization is a method to transform data to the same scale so that all values have a mean of 0 and a standard deviation of 1. This technique is useful when the dataset contains features with different sizes or units. The standardization formula can be found as follows [16]:

$$X_{std} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{n}(x_i - x_{mean})^2} \quad (1)$$

$$X' = \frac{X - X_{mean}}{X_{std}} \quad (2)$$

Where:
x is the original value,
X_{mean} is the average of the training data,
X_{std} is the standard deviation of the training data for each feature, and
X' is a standardized value

Subsequently, in the Detection and Handling of Outliers, outliers can significantly impact the performance of the SVR model and in some cases, also affect Decision Trees and Random Forests. This study addresses outliers through winsorizing. Winsorizing is a method to handle outliers by replacing extreme values beyond a certain threshold with that threshold value. The goal is to mitigate the impact of outliers without removing data. The upper and lower thresholds are typically determined based on percentiles, such as 5% and 95%. Values exceeding these thresholds are considered outliers and are replaced with the corresponding threshold value [17]. Furthermore, to determine the most dominant features or variables for use in data modeling, feature selection in this study is conducted by calculating the correlation between variables. Feature correlation is determined by the correlation value between independent and dependent variables. Subsequently, irrelevant or uninformative variables are removed.

2.3 Split Train Data and Test Data

In the next stage, the data division process is carried out. The data is divided into training data and testing data. The composition is 2:1, approximately 67% for training and 33% for testing. Data from the years

2021 and 2022 are used as training data, totaling 66 data points, while data from the year 2023 serves as testing data, with 33 data points. This division aims to train the model on a subset of data and evaluate its performance on a previously unseen subset of data [18].

2.4 Supervised Learning

Linguistically, Supervised learning is directed learning. The computer or machine will learn from labeled training data in the learning process. If analogized to a student and a teacher, the computer is the student who learns, and the teacher will instruct the student to learn from problems that already have solutions and answer keys [19].

a.) Support Vector Regression (SVR)

Support Vector Regression (SVR) is a learning method used to estimate the values of continuous variables. SVR operates on a principle similar to Support Vector Machine (SVM), but its main objective is to find the line or curve that best fits the data. In SVR, the best curve is referred to as a hyperplane that can accommodate as many data points within its range as possible. Unlike other regression methods that focus on minimizing the error between predicted and actual values, SVR seeks the best curve within a specific range, known as the epsilon margin [20]:

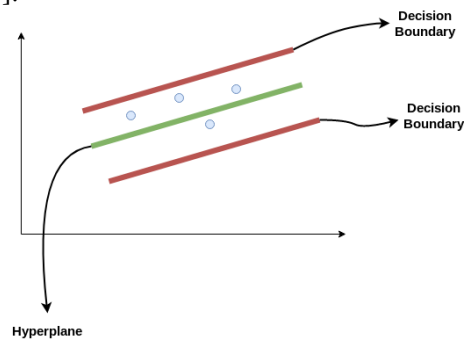

Figure 2.2 Support Vector Regression Illustration

The SVR algorithm attempts to find the best curve based on data points. Since this is a regression algorithm, SVR uses the curve to find a fit between vectors and the curve position, rather than as a decision boundary. Support Vector points help determine the optimal position of the curve to align with the data [20].

b.) Decision Tree

The Decision Tree algorithm, or decision tree, is a method used to create a decision-making model in the form of a tree based on training data attributes. Here are some popular algorithms: Classification and Regression Trees (CART), Iterative Dichotomiser 3 (ID3), C4.5, and C5.0 [19].

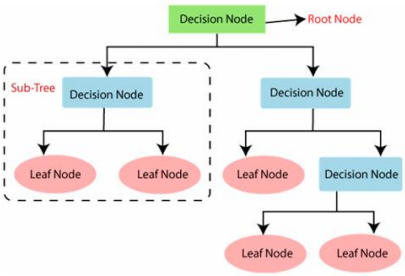

Figure 2.3 Illustration of the workflow of the Decision Tree method.

The Decision Tree process begins by selecting the most effective attribute to divide the data into smaller and uniform groups. This attribute selection is based on criteria such as Information Gain or Gini Impurity for classification, and Mean Squared Error (MSE) for regression, aiming to reduce uncertainty within each group. Each attribute is tested to determine which provides the best separation results. The MSE formula is applied at each node. After evaluating all features and splitting points, choose the one with the lowest MSE value. Once the best attribute is determined, the Decision Tree branches out, with each branch representing a different category or value of that attribute. This process continues recursively for each branch until reaching

a leaf node that represents the outcome or label. The resulting decision tree can be used to determine decision steps based on input by following the appropriate branch paths. During this process, the Decision Tree can also be pruned to prevent overfitting by removing branches that do not significantly contribute to improving the model's generalization ability. At each tested split, the data is divided into Left Node and Right Node according to the threshold. The final result is a decision tree that can be used to classify new data by following the relevant branches, enabling better and easily understandable decision-making based on the rules represented in the tree structure [21].

c) Random Forest (RF)

Random Forest (RF) is a commonly used supervised machine learning algorithm for classification and regression tasks. RF is an ensemble method, which improves accuracy by combining several classification models [22].

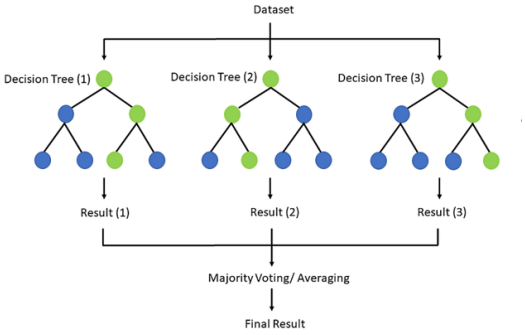

Figure 2.4 Illustration of the workflow of the Random Forest method.

Each tree within the Random Forest provides separate predictions. The final result of the Random Forest model is the average of all tree predictions.

2.5 Evaluation

In evaluating models, several metrics can be used, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and R Squared (R^2). MAE, Root Mean Squared Error (RMSE), and MAPE values are considered good if they approach 0 [23]. According to Lewis (1982), the MAPE value can be classified into 4 categories, as shown in the following table [24].

Table 2.2 Interpretation of MAPE value

<10%	High Prediction Rate
10%-20%	Predictions are acceptable
20%-50%	Predictions can be tolerated.
>50%	Unacceptable

Testing using Mean Squared Error (MSE) aims to calculate the average possible prediction error. The MSE value is said to be good if the prediction results are close to zero [23].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \quad (3)$$

Description:

n = number of data, $\hat{x}_i$ = predicted value, x_i = true value

Measuring accuracy with MSE has two major drawbacks. First, MSE only indicates the fit of the model to the historical data, so models such as high-degree polynomials can minimize MSE but become too sensitive and less reliable for forecasting. Secondly, MSE does not take into account procedural differences between forecasting methods, so it is inappropriate to use it as the sole measure of accuracy. Therefore, the weakness of MSE can be combined with other error measurements. One of them is using MAPE or measuring error relative to the actual data. Mathematically, MAPE is expressed as [25]:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - \hat{x}_i|^2}{x_i} \times 100\% \quad (2.14)$$

MAPE is calculated by dividing the absolute error of each period by its actual value. This method is useful for evaluating prediction accuracy, showing how much the error is compared to the actual value. In choosing a forecasting method, smaller MSE and MAPE values indicate better accuracy, with prediction results that are close to actual data through error minimization.

3. RESULTS AND DISCUSSIONS

This research utilizes stunting prevalence data from 33 districts/cities in North Sumatra Province during the period from 2021 to 2023. The dataset consists of 14 predictor variables used as features for the prediction model, as well as one target variable, which is stunting prevalence. These variables encompass

factors related to socio-economic, health, and infrastructure aspects. Each district/city has data for all three years. The data from 2021 and 2022 are used as training data to train the prediction model, while the data from 2023 are used as testing data to assess the model's performance in predicting stunting prevalence for the final year. The compiled data is stored in Comma Separated Value (CSV) format, allowing processing using the Supervised Machine Learning tool, Python, through Google Colab. The following is the prediction result for each model in each district.

Table 3.1 Prediction results for the district

Districts/Cities	Actual	Predicted SVR	Decision Tree	Random forest
Asahan	1.6	1.55	1.3	1.633333
Batu Bara	9	7.95	3.5	9.433333
Dairi	13.7	14.3	14.3	14.3
Deli Serdang	0.4	0.8	0.9	0.833333
Humbang Hasundutan	10.6	17.65	16.5	17.26667
Karo	12.8	16.9	17.2	16.8
Labuhan Batu	0.8	0.7	0.2	0.866667
South Labuhan Batu	0.8	1.7	2.1	1.833333
North Labuhan Batu	1.5	1.8	1.4	1.933333
Langkat	1	2.45	2.1	2.566667
Mandailing Natal	4.8	3.300001	4.3	3.633333
Nias	14.6	19.35021	12.1	21.76667
West Nias	21.1	25.20002	21.5	26.43333
South Nias	10.7	13.95	13.3	14.16667
North Nias	3	4.55	3.9	4.333333
Padang Lawas	9.9	6.05	9.5	7.2
North Padang Lawas	5.9	6.4	6	6.266667
West Pakpak	17.9	21.7	21.8	21.66667
Samosir	11.8	10.3	7.8	11.13333
Serdang Bedagai	1.9	2.2	3.2	2.533333
Simalungun	1	1.4	1.9	1.566667
SouthTapanuli	0.9	4.05	3.5	4.233333
Middle Tapanuli	2.2	4.45002	4.3	4.5
NorthTapanuli	9.5	8.4	7	8.866667
Toba Samosir	10	8.55	8.5	8.566667
Binjai	0.9	1.1	1.1	1.1
Gunungsitoli	10.8	4.1	2.8	3.666667
Medan	0.9	0.85	0.8	0.833333
Padangsidimpuan	11.6	14.05	11.8	14.8
Pematang Siantar	2.2	2.05	0.7	1.6
Sibolga	4.7	5.859455	7	6.233333
Tanjung Balai	2.1	1.6	2.3	1.833333
Tebing Tinggi	2.2	3.2	2.9	3.1

The models used are Support Vector Regression (SVR), Decision Tree, and Random Forest. The performance of these three models is evaluated using several metrics, namely Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). The following table presents the performance comparison of each model in several districts/cities. For the SVR model evaluation, the lowest MSE of 0.0025 was obtained in Medan and Asahan cities, while the lowest MAPE of 3.125 was achieved in the Asahan district. As for the Random Forest model evaluation, the lowest MSE of 0.0011 and the lowest MAPE of 2.08 were observed in the Asahan district. Lastly, for the Decision Tree model, the lowest MSE of 0.01 and the lowest MAPE of 1.694 were found in the Padang Lawas Utara district. The following is the best evaluation result for each district and city.

Table 3.2 The best model in the district/city

Districts/Cities	Best Model	MSE	MAE	RMSE	MAPE
Asahan	Random Forest	0.001111	0.033333	0.033333	2.083333
Batu Bara	Random Forest	0.187778	0.433333	0.433333	4.814815

.

Dairi	Svr	0.36	0.6	0.6	4.379562
Deli Serdang	Svr	0.16	0.4	0.4	100
Humbang Hasundutan	Decision Tree	34.81	5.9	5.9	55.66038
Karo	Random Forest	16	4	4	31.25
Labuhan Batu	Random Forest	0.004444	0.066667	0.066667	8.333333
South Labuhan Batu	Svr	0.81	0.9	0.9	112.5
North Labuhan Batu	Decision Tree	0.01	0.1	0.1	6.666667
Langkat	Decision Tree	1.21	1.1	1.1	110
Mandailing Natal	Decision Tree	0.25	0.5	0.5	10.41667
Nias	Decision Tree	6.25	2.5	2.5	17.12329
West Nias	Decision Tree	0.16	0.4	0.4	1.895735
South Nias	Decision Tree	6.76	2.6	2.6	24.29907
North Nias	Decision Tree	0.81	0.9	0.9	30
Padang Lawas	Decision Tree	0.16	0.4	0.4	4.040404
North Padang Lawas	Decision Tree	0.01	0.1	0.1	1.694915
West Pakpak	Random Forest	14.18778	3.766667	3.766667	21.04283
Samosir	Random Forest	0.444444	0.666667	0.666667	5.649718
Serdang Bedagai	Svr	0.09	0.3	0.3	15.78947
Simalungun	Svr	0.16	0.4	0.4	40
SouthTapanuli	Decision Tree	6.76	2.6	2.6	288.8889
Middle Tapanuli	Decision Tree	4.41	2.1	2.1	95.45455
NorthTapanuli	Random Forest	0.401111	0.633333	0.633333	6.666667
Toba Samosir	Random Forest	2.054444	1.433333	1.433333	14.33333
Binjai	Svr	0.04	0.2	0.2	22.22222
Gunungsitoli	Svr	44.89	6.7	6.7	62.03704
Medan	Svr	0.0025	0.05	0.05	5.555556
Padangsidimpuan	Decision Tree	0.04	0.2	0.2	1.724138
Pematang Siantar	Svr	0.0225	0.15	0.15	6.818182
Sibolga	Svr	1.344336	1.159455	1.159455	24.66926
Tanjung Balai	Decision Tree	0.04	0.2	0.2	9.52381
Tebing Tinggi	Decision Tree	0.49	0.7	0.7	31.81818

Based on the analysis results, the Random Forest model demonstrates the best performance in predicting the prevalence of stunting in most districts/cities in North Sumatra Province, with lower evaluation values compared to other models. For example, in Asahan District, the Random Forest model produces MSE = 0.001111, MAE = 0.033333, RMSE = 0.033333, and MAPE = 2.083333, indicating very low prediction errors. In comparison, SVR has MSE = 0.36, MAE = 0.6, RMSE = 0.6, and MAPE = 4.379562 in Dairi District, which records significantly higher errors.

Furthermore, the Decision Tree shows good performance in some areas, such as in North Labuhan Batu District, with MSE = 0.01, MAE = 0.1, RMSE = 0.1, and MAPE = 6.666667, but also generates larger errors in other areas like Humbang Hasundutan District, with MSE = 34.81, MAE = 5.9, RMSE = 5.9, and MAPE = 55.66038. This indicates that the Decision Tree model tends to. This indicates that the Decision Tree model tends to be unstable and produces significant prediction errors in some districts.

SVR exhibits varying performance. For example, in Medan City, SVR yields MSE = 0.0025, MAE = 0.05, RMSE = 0.05, and MAPE = 5.555556, which is considered good. However, in South Labuhan Batu Regency, SVR shows very high errors with MSE = 0.81, MAE = 0.9, RMSE = 0.9, and MAPE = 112.5.

Several areas exhibit very high MAPE, especially with SVR and Decision Tree, such as in Deli Serdang Regency (100), South Labuhan Batu Regency (112.5), and South Tapanuli Regency (288.8889). This indicates that although these models may perform better in terms of MSE or MAE, their performance in terms of error percentage (MAPE) is poor.

The comparison of model performance provides valuable insights into the predictive accuracy of Random Forest, SVR, and Decision Tree models for stunting prevalence in North Sumatra. However, the practical implications of these results need to be further elaborated, particularly in guiding stunting policy planning in underperforming regions. For example, the consistent performance of the Random Forest model, particularly its low error metric in districts such as Asahan, suggests that it could be a reliable tool for identifying priority areas. In contrast, the variability in the performance of Decision Tree and SVR, with high errors in districts such as Humbang Hasundutan and Labuhan Batu Selatan, highlights the need for careful application of these models in a policy context. Policymakers in areas with very high MAPE, such as Deli Serdang (100) or Tapanuli Selatan (288.8889), should consider using alternative models or improving data quality and feature engineering to increase prediction accuracy.

The results also underscore the importance of customizing interventions. For example, districts with consistently high prediction errors may benefit from additional data collection efforts to capture unique local factors that influence stunting. Additionally, the model's ability to identify significant predictors (e.g., maternal nutrition, poverty level, or access to clean water) may inform targeted policy measures such as strengthening maternal health programs, improving sanitation infrastructure, or expanding poverty alleviation efforts.

4. CONCLUSION

Based on the evaluation results across various districts/cities, it can be concluded that Random Forest demonstrates superior and consistent performance in predicting stunting prevalence, achieving the lowest values for MSE, MAE, RMSE, and MAPE in most tested areas. This stable performance highlights its effectiveness in handling complex datasets with wide variations, making it the most reliable model for this study. Decision Tree, while advantageous for its interpretability and computational efficiency, shows inconsistent results, with good performance in certain areas but significant errors in others. SVR, on the other hand, exhibits variability and is more suitable for simpler cases where computational speed is prioritized.

These findings have important implications for reducing stunting in North Sumatra. The Random Forest model's ability to identify key predictors of stunting, such as maternal health, sanitation, and socio-economic factors, can guide policymakers in designing targeted interventions. For example, districts with higher stunting prevalence can focus on improving maternal nutrition, expanding access to clean water, and enhancing healthcare services based on the significant predictors identified by the model.

However, this study has several limitations. The analysis was based on available data for North Sumatra, which may not capture all possible factors influencing stunting, such as cultural practices or specific dietary patterns. Expanding the dataset to include additional features, such as community-level health initiatives or detailed socio-economic indicators, could improve model accuracy. Additionally, future research could explore hybrid approaches, combining the strengths of multiple models, to enhance prediction performance in underperforming districts.

REFERENCES

[1] Kominfo, "Indonesia Cegah Stunting, Antisipasi Generasi Stunting Guna unggulan Indonesia Emas 2045," Kominfo. Accessed: Feb. 22, 2024. [Online]. Available: https://www.kominfo.go.id/content/detail/32898/indonesia-cegah-stunting-antisipasi-generasi-stunting-guna-mencapai-indonesia-emas-2045/0/artikel_gpr

[2] B. C. Rosha, A. Susilowati, N. Amaliah, and Y. Permanasari, "Penyebab Langsung dan Tidak Langsung Stunting di Lima Kelurahan di Kecamatan Bogor Tengah, Kota Bogor (Study Kualitatif Kohor Tumbuh Kembang Anak Tahun 2019)," *Bul. Penelit. Kesehat.*, vol. 48, no. 3, Oct. 2020, doi: 10.22435/bpk.v48i3.3131.

[3] WHO and UNICEF, *Levels and trends in child malnutrition*. 2023.

[4] Rokom, "Prevalensi Stunting di Indonesia Turun ke 21,6% dari 24,4%," Kementerian Kesehatan RI. Accessed: Mar. 28, 2024. [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20230125/3142280/prevalensi-stunting-di-indonesia-turun-ke-216-dari-244/

[5] Y. Pencawan, "Target Prevalensi Stunting Sumut 2023 Dipatok 18%," Media Indonesia. Accessed: Apr. 01, 2024. [Online]. Available: https://mediaindonesia.com/ekonomi/624470/target-prevalensi-stunting-sumut-2023-dipatok-18

[6] M. S. Haris, A. N. Khudori, and W. T. Kusuma, "Perbandingan Metode Supervised Machine Learning untuk Prediksi Prevalensi Stunting di Provisi Jawa Timur," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 9, no. 7, p. 1571, Dec. 2022, doi: 10.25126/jtiik.2022976744.

[7] K. M. Hindrayani, T. M. Fahrudin, R. Prismahardi Aji, and E. M. Safitri, "Indonesian Stock Price Prediction including Covid19 Era Using Decision Tree Regression," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, Dec. 2020, pp. 344–347. doi: 10.1109/ISRITI51436.2020.9315484.

[8] R. A. Saputri and J. Tumangger, "Hulu-Hilir Penanggulangan Stunting Di Indonesia," *J. Polit. Issues*, vol. 1, no. 1, pp. 1–9, Jul. 2019, doi: 10.33019/jpi.v1i1.2.

[9] E. Galasso and A. Wagstaff, "The aggregate income losses from childhood stunting and the returns to a nutrition intervention aimed at reducing stunting," *Econ. Hum. Biol.*, vol. 34, pp. 225–238, Aug. 2019, doi: 10.1016/j.ehb.2019.01.010.

[10] E. Vianita, A. Wibowo, B. Surarso, and A. P. Widodo, "Car insurance segmentation prediction based on the most influential features using random forest and stacking ensemble learning," *J. Soft Comput. Explor.*, vol. 2, no. 2, Sep. 2021, doi: 10.52465/joscex.v2i2.39.

[11] K. Budiman and Y. N. Ifriza, "Analysis of earthquake forecasting using random forest," *J. Soft Comput. Explor.*, vol. 2, no. 2, Sep. 2021, doi: 10.52465/joscex.v2i2.51.

[12] A. F. Mulyana, W. Puspita, and J. Jumanto, "Increased accuracy in predicting student academic performance using random forest classifier," *J. Student Res. Explor.*, vol. 1, no. 2, pp. 94–103, Jul. 2023, doi: 10.52465/josre.v1i2.169.

[13] E. Retnoningsih and R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python," *BINA Insa. ICT J.*, vol. 7, no. 2, p. 156, Dec. 2020, doi: 10.51211/biict.v7i2.1422.

[14] N. Nadiah, S. Soim, and S. Sholihin, "Implementation of Decision Tree Algorithm Machine Learning in Detecting Covid-19 Virus Patients Using Public Datasets," *Indones. J. Artif. Intell. Data Min.*, vol. 5, no. 1, p. 37, Jun. 2022, doi: 10.24014/ijaidm.v5i1.17054.

[15] Farhanuddin, Sarah Ennola Karina Sihombing, and Yahfizham, "Komparasi Multiple Linear Regression dan Random Forest

.

Regression Dalam Memprediksi Anggaran Biaya Manajemen Proyek Sistem Informasi," *J. Comput. Digit. Bus.*, vol. 3, no. 2, pp. 86–97, May 2024, doi: 10.56427/jcbd.v3i2.408.

[16] A. Ambarwari, Q. Jafar Adrian, and Y. Herdiyeni, "Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 117–122, Feb. 2020, doi: 10.29207/resti.v4i1.1517.

[17] Z. Bobbitt, "How to Winsorize Data: Definition & Examples." Accessed: Nov. 16, 2024. [Online]. Available: https://www.statology.org/winsorize/

[18] S. Raheja, "Train-Test-Validation Split: A Critical Component of ML." Accessed: Nov. 12, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2023/11/train-test-validation-split/

[19] H. Sahli, "An Introduction to Machine Learning," in *TORUS 1 – Toward an Open Resource Using Services*, Wiley, 2020, pp. 61–74. doi: 10.1002/9781119720492.ch7.

[20] Trivusi, "Algoritma Support Vector Regression (SVR): Jenis SVM untuk Regresi," Trivusi. Accessed: Jul. 12, 2024. [Online]. Available: https://www.trivusi.web.id/2022/08/algoritma-svr.html

[21] S. Baladram, "Regression Tree | Towards Data Science," Towards Data Science. Accessed: Nov. 12, 2024. [Online]. Available: https://towardsdatascience.com/decision-tree-regressor-explained-a-visual-guide-with-code-examples-fbd2836c3bef

[22] R. A. Haristu and P. H. P. Rosa, "Penerapan Metode Random Forest untuk Prediksi Win Ratio Pemain Player Unknown Battleground," *MEANS (Media Inf. Anal. dan Sist.*, pp. 120–128, Oct. 2019, doi: 10.54367/means.v4i2.545.

[23] M. A. Kholik, "PERBANDINGAN METODE POLYNOMIAL REGRESSION DAN METODE SUPPORT VECTOR MACHINE DALAM MEMPREDIKSI SEBARAN COVID-19 DI INDONESIA," *J. Inform. Teknol. dan Sains*, vol. 1, no. 2, pp. 10–20, May 2023, doi: 10.56244/formateks.v1i2.631.

[24] I. Nabillah and I. Ranggadara, "Mean Absolute Percentage Error untuk Evaluasi Hasil Prediksi Komoditas Laut," *JOINS (Journal Inf. Syst.*, vol. 5, no. 2, pp. 250–255, Nov. 2020, doi: 10.33633/joins.v5i2.3900.

[25] Dini Indriyani Putri, Agung Budi Prasetijo, and Adian Fatchur Rochim, "Prediksi Harga Saham Menggunakan Metode Brown's Weighted Exponential Moving Average dengan Optimasi Levenberg-Marquardt," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 10, no. 1, pp. 11–18, Feb. 2021, doi: 10.22146/jnteti.v10i1.678.