# Analysis of Stunting Prevalence in Indonesia: A Comparative Study Using K-Means Clustering, Support Vector Machine, and *Decision Tree* Algorithms

## Rosa Aulia Wangi

Information System, Faculty of Engineering & Information

Multimedia Nusantara University, Tangerang Banten 15810, Indonesia.

rosa.aulia@student.umn.ac.id,

***Abstract*–** Stunting, a condition of impaired growth and development in children due to malnutrition in the first 1000 days of life, poses a significant challenge for Indonesia, with a prevalence of 21.6% in 2022 and a reduction target of 14% by 2024. This study addresses the issue of stunting by employing machine learning techniques, specifically K-Means Clustering, Support Vector Machine (SVM), and Decision Tree algorithms. The predictor variables were chosen based on the target demographics of the stunting reduction program and adapted to reflect a multisectoral approach. Data sourced from the Central Statistics Agency and Survey Status Gizi Indonesia (SSGI) 2022 were analyzed. K-Means Clustering was used to identify regional patterns related to stunting, providing crucial insights for targeted interventions. To further explore these patterns, we then performed classification analysis using SVM and Decision Tree models, which helped in investigating the factors influencing stunting more deeply. The analysis revealed that low birth weight (LBW) is significantly positively correlated with stunting prevalence, while variables such as "FASYANKES Delivery," "K4," "Complete KF," "mCPR," and "A 6-11" showed significant negative correlations. SVM achieved an accuracy of 88%, particularly with a linear kernel, whereas the Decision Tree model demonstrated perfect accuracy of 100%, indicating superior performance. The integration of clustering and classification methods in this research highlights effective strategies for stunting reduction in Indonesia.

***Keywords:*** *Decision Tree, K-means, Malnutrition, Low Birth Weight, Stunting, Support Vector Machine*

1. **INTRODUCTION**
2. **Background and Objectives**

Stunting, which is a condition of growth failure in children due to malnutrition during the first 1000 days of life, is a critical issue that threatens Indonesia's future [1]. The long-term impacts of stunting include cognitive decline, psychomotor barriers, difficulties in academic and sports performance, and the risk of degenerative diseases in adulthood [2]. The prevalence rate of stunting in Indonesia has reached 21.6% and the reduction target set in the 2020-2024 National Medium Term Development Plan (RPJMN) is 14%, making accelerated efforts in handling stunting very important [3].

Detecting stunting in the Indonesian community is usually done at Posyandu by measuring the length or height of children and comparing it to the WHO 2006 standards. However, this is often not done for several reasons, including limitations of cadres in conducting measurements, difficulty in determining the age of children, and assessing or converting measurement results according to the WHO 2006 standard to determine whether the child is stunted or not. Therefore, we will conduct research on the prevalence of stunting using machine learning methods.

As a step to overcome the complexity of factors that influence stunting, this research adopts an innovative approach by utilizing three machine learning methods, including K-Means Clustering, Support Vector Machine (SVM), and Decision Tree. The predictor factors selected are based on the target group of the stunting reduction program and have been adjusted to reflect a multisectoral approach. K-Means Clustering is used to identify regional patterns related to stunting, providing important insights for more focused intervention strategies. SVM and Decision Tree, on the other hand, were chosen for classification analysis and further research into stunting factors in each province. The basis for selecting this method is supported by previous literature. K-Means Clustering, as an unsupervised learning algorithm, has been proven to be effective in identifying similar regional patterns [4]. SVM showed superior accuracy performance in several classification studies [5] [6] [7], while Decision Tree was found to be superior in several other studies [8] [9] [10] [11] [12] [13]. Therefore, this research will compare the performance of the two methods to determine the best approach.

Combining data from the Central Bureau of Statistics and SSGI 2022, the main aim of this research is to provide evidence-based insights that can be used in formulating effective public health strategies. Through the combination of the predictive advantages of Decision Tree, the reliability of SVM, and the capabilities of K-Means Clustering, this research is expected to make a significant contribution to efforts to reduce the prevalence of stunting in Indonesia until 2024.

In this research, we aim to enhance understanding of the prevalence of stunting in children in Indonesia by employing an innovative machine learning approach, focusing on comparing the effectiveness of K-Means Clustering, Support Vector Machine (SVM), and Decision Tree methods. Our objective is to identify the most significant predictor factors and evaluate their impact on the levels of stunting across different regions. Unlike previous studies which mainly relied on conventional statistical methods or individual machine learning models, and typically achieved around 80% accuracy, our research adopts a multifaceted machine learning framework, pushing us to seek optimal accuracy. By integrating regional data from the Central Statistics Agency and SSGI 2022, we present a comprehensive analysis of stunting factors, thereby facilitating more targeted public health strategies tailored to the needs of specific provinces. This approach not only enhances the accuracy of stunting predictions but also provides detailed insights for targeted interventions, marking significant progress compared to previous studies that lacked such extensive comparative analyses.

1. **Study Literature**
2. K-Means algorithm

The k-means algorithm is an unsupervised approach used to categorize unlabeled data into a number of clusters. The k-means algorithm implementation process includes the following steps:

1. Selects the number of clusters (k) required to group the data.
2. Setting the cluster center in a certain way can be random initialization.
3. Classify each data into the closest cluster based on calculating the distance between the data and the cluster center. In this case, the distance is measured by the method *Euclidean Distance*.
4. Measuring the relationship of data to a particular cluster by comparing the distance of the data to the cluster center. This involves calculating the distance between each data point and each cluster center using the Euclidean Distance formula, with a variable like D(i, j) representing the distance between the data point and the cluster center as follows.

With description:

D(*i, j*) : distance from the ith data to the jth cluster center

Xto : the ith data in the kth attribute

ANDkj : the jth center point in the kth attribute

1. Modifies the center of a cluster by recalculating the average of all the data in that cluster. This method can be an average calculation or a cluster medoid option.
2. Iterates for each data considering the updated cluster center. If there is a change in the cluster center, the previous steps are repeated. This iteration continues until the cluster center stabilizes, indicating the completion of the clustering process.

Several evaluation techniques can be used to determine the most appropriate number of clusters. In this study, the silhouette score and the Calinski-Harabasz Index were used. The silhouette score serves as an evaluation indicator that measures the degree to which each sample is correctly grouped by an algorithm such as k-means. The range of this score varies between -1 to 1, closer to 1, indicating the sample is far from other groups; a score of 0 indicates that the sample falls between the group decision limits; and a negative score implies potential errors in grouping [14]. The Calinski-Harabasz Index (CHI) is also a valuable evaluation tool in assessing the efficiency of clustering algorithms. CHI, which is sometimes referred to as *Variance Ratio Criterion* (VRC), evaluates the quality of clustering by comparing between-group variation with within-group variation. In other words, the higher the CHI value, the more efficient and quality the clustering produced by the algorithm [15].

1. Algoritma Support Vector Machine

The Support Vector Machine (SVM) algorithm is a machine learning approach used for data classification, especially for two classes. In practice, SVM attempts to find a hyperplane that separates the data as optimally as possible. This hyperplane, which is also called the teacher, looks for the best position to separate the data, so that the margin or distance between the teacher and the closest data from each class is maximum. Although SVMs were originally designed for linear data, with the use of kernel functions, the data can be expanded to higher dimensions, allowing SVMs to handle non-linear data. Apart from classification, SVM can also be used in regression tasks. Following are several types of kernel functions in SVM:

1. Linear Kernel: Linear kernel function is an optimal approach when the given data has clear linear characteristics. This kernel implements linear tutoring without requiring additional data transformations, assuming the data is linearly separable. Through this approach, SVM shows significant efficiency in classifying data that shows linearity [16].
2. Polynomial Kernel: Applied to situations where the data does not allow separation with a linear model. Polynomial kernels transform data to higher dimensions, producing teachers with polynomial properties. Its main advantage lies in its flexibility to adjust the degree of the polynomial according to the complexity of the existing data [17].
3. Radial Basis Function (RBF) Kernel: Relevant when data exhibits significant non-linear properties. Through data transformation into a higher dimensional space, the RBF kernel produces a teacher capable of modeling non-linear characteristics with precision. Its ability to capture the nuances of non-linear patterns makes it a strategic choice for complex data analysis. [17].
4. Sigmoid Kernel: Recommended for data that exhibits non-linear properties that cannot be accommodated by linear or polynomial models. The sigmoid kernel modifies the data to higher dimensions, producing teachers with sigmoidal properties. This approach is particularly effective in dealing with data structures that exhibit complex non-linear relationships [17].

It is important to emphasize that the selection of the most suitable kernel should be based on an in-depth analysis of the intrinsic properties and distribution of the data to be processed [18].

1. Decision Tree Algorithm

Decision Trees are used to assist in predicting or classifying data. This tool works by taking a group of data and looking for patterns or rules in it to help in decision making. The steps to build a decision tree are as follows:

1. Take all initial data and store it in one main group.
2. Finding out what data or variables have the most influence in determining results, usually using the Attribute Selection Measure method.
3. Divide the main groups based on the variables that have been found, so that smaller but more specific groups are obtained.
4. On each newly created group, steps 2 and 3 need to be repeated. This process will continue until the program can no longer divide the data into smaller groups. These latter groups are referred to as "leaf nodes".

Decisions at each step are made based on a metric called information gain. This metric measures how well a variable can help in predicting outcomes. The higher the Information Gain, the better the variable [10].

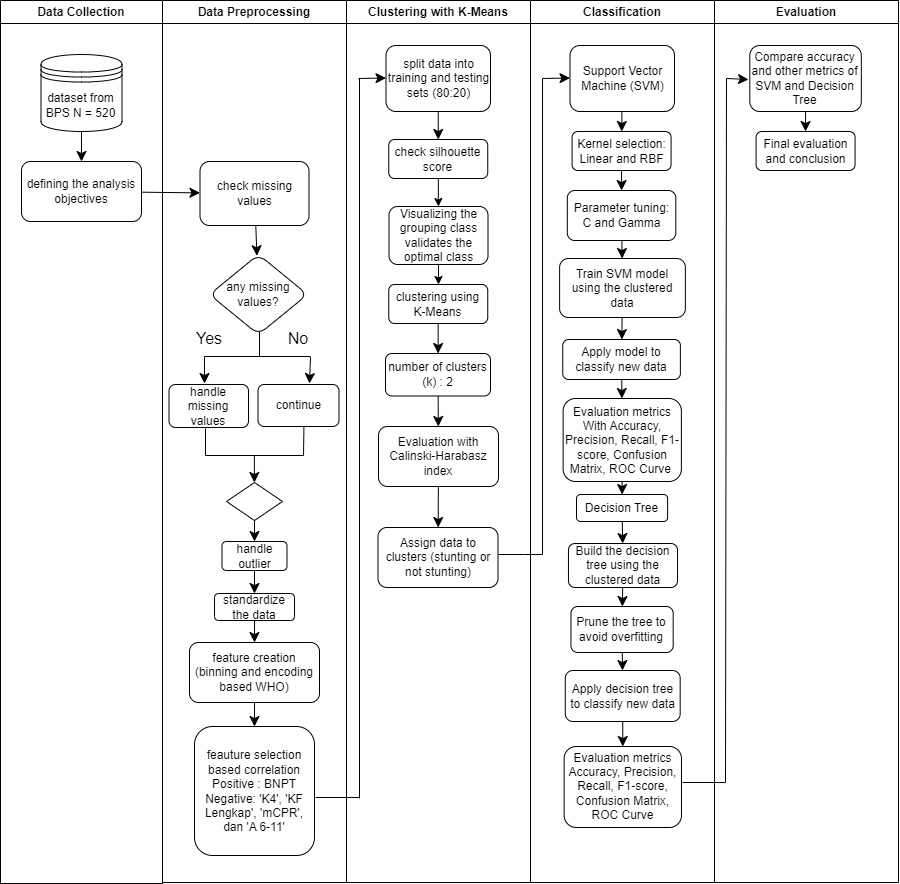
d) Receiver Operating Characteristic Curve (ROC)

Receiver Operating Characteristic is the graphical representation most frequently used in binary classification to assess classifier operation. ROC has a sensitivity value (positive level of truth) on the y-axis and a negative level of truth value of 1 – sensitivity on the x-axis. The ROC curve addresses how the classifier operates by comparing the classifier's ability to separate positive and negative classes while changing the classification threshold. The diagonal line in the ROC space shows random guessing as a representative of the classifier not being able to discriminate. This line, also known as AUC (Area Under Curve), is the area under the ROC curve. The AUC value is between 0 and 1, with a value of 1 indicating perfect performance and a value of 0 indicating performance that is no better than random. Higher AUC values ​​indicate better performance.

An ideal classifier would have a ROC curve that reaches the upper left corner, indicating high sensitivity and a low false positive rate at threshold. A comprehensive evaluation of the classifier performance across all possible threshold values ​​can be displayed on the ROC curve, thereby helping to select the optimal threshold for classification purposes. Classifiers that have curves closer to the top left corner are better than curves that are closer to the diagonal line. This curve is also resistant to class imbalance and is not influenced by the distribution of classes in the data set so it is often relied on to evaluate classifier performance. [22]

1. **METHODOLOGY**

The process carried out by researchers can be illustrated in detail through Figure 1.



*Figure 1. Research Flowchart*

In the next sub-chapter, an in-depth explanation of each stage that has been carried out will be described, providing a comprehensive understanding of the methodology applied in this research.

1. **Data collection**

This research utilizes secondary data sourced from the 2022 Central Statistics Agency publications. This data can be obtained and verified via the Central Statistics Agency's official website at https://www.bps.go.id/. The dataset that has been compiled and collected is presented in Excel file format. Furthermore, the data has been imported and integrated into the Jupyter Notebook environment for further analysis.

A screenshot of a computer

Description automatically generated

*Figure 2. Variable Information in the Dataset*

The dataset used in this research includes 23 variable columns with most of them using percentage units, except for the 'UMK' column which uses rupiah units. For more detailed information regarding each variable, it can be found in Table 1. Technically, the majority of columns have the float64 data type which indicates the numeric data type, but it should be noted that the 'Indonesian District/City' column has the object data type which indicates the data type strings.

*Table 1. Variable Identification*

|  |  |
| --- | --- |
| Column Name | Definition |
| Stunting Prevalence (TB/U) % | Percentage of toddlers with nutritional conditions categorized as very short and short based on the height index for age (TB/U) with a Z score of less than -2 standard deviation in each Regency/City in Indonesia in 2022. |
| K4 | Percentage of pregnant women who undergo pregnancy checks at least 4 (four) times during the pregnancy period. This examination consists of one examination in the first trimester, one examination in the second trimester, and two examinations in the third trimester in every Regency/City in Indonesia in 2022. |
| HEALTH FACILITIES DELIVERY | Percentage of deliveries carried out at FASYANKES in each Regency/City in Indonesia in 2022. |
| Complete KF | Percentage of postnatal/postpartum mothers who receive complete postpartum visits up to 42 days after giving birth in each Indonesian Regency/City in 2022. |
| Vit A Mom | Percentage of postpartum mothers who receive Vitamin A supplements in each Regency/City of Indonesia in 2022 |
| Pregnant Women TTD (90 Tablets) | Percentage of pregnant women who received 90 Blood Supplement Tablets (TTD) during pregnancy in each Regency/City of Indonesia in 2022. |
| LBW (Low Birth Weight) | Percentage of LBW babies per Indonesian Regency/City 2022 |
| IMD (Early Initiation of Breastfeeding) | Percentage of newborns who receive IMD in each Regency/City of Indonesia 2022 |
| ASI (Mother's Milk) | Percentage of babies aged < 6 months who are exclusively breastfed in each Regency/City of Indonesia 2022 |
| CPKB (Baby Health Services Coverage) | Percentage of male and female babies receiving essential neonatal health services, including resuscitation measures, prevention of hypothermia, early and exclusive breastfeeding practices, eye, umbilical cord, skin care, immunizations, administration of vitamin K, integrated management of young toddlers (MTBM), and counseling for mothers regarding caring for neonates at home using the KIA book, in every district/city in Indonesia in 2022. |
| IDL (Complete Basic Immunization) | Percentage of babies who have received one dose of Hepatitis B vaccination, one dose of BCG vaccination, three doses of DPT-HB/DPT-HB-Hib vaccination, four doses of polio vaccination, and one dose of measles rubella vaccination in each Regency/City of Indonesia in 2022. |
| A 6-11 | Percentage of babies aged 6-11 months who receive vitamin A supplements in each district/city in Indonesia in 2022. |
| A 12-59 | Percentage of babies aged 12-59 months who receive vitamin A supplements in each district/city in Indonesia in 2022. |
| A 6-59 | Percentage of babies aged 6-59 months who receive vitamin A supplements in each district/city in Indonesia in 2022. |
| mCPR (Modern Method Active Family Planning Participants) | The percentage of new and existing family planning program participants who are still actively using contraception continuously with modern methods such as condoms, injections, pills, IUDs, MOW, MOP, Implants, and MAL for the purpose of delaying, spacing pregnancies, or ending fertility in every Regency/City in Indonesia in 2022. |
| Decent Drinking Water | Percentage of households by Regency/City in Indonesia in 2022 that have adequate access to drinking water. |
| Proper Sanitation | Percentage of households by Regency/City in Indonesia in 2022 that have access to adequate sanitation |
| IKP (Food Security Index) | The IKP percentage consists of three main aspects of food security, namely availability, affordability and utilization of food. |
| BNPT 40% | The percentage of households in quintile groups 1 and 2 (population in the lowest 40% expenditure group) in Indonesia broken down by district/city and purchase/receipt of Raskin rice/non-cash food assistance (BPNT) during the last four months of 2022 |
| KKS 40% | Percentage of households in quintile groups 1 and 2 (population in the bottom 40% expenditure group) receiving social protection cards (KPS)/prosperous family cards (KKS) in Indonesia detailed by Regency/City in 2022. |
| PAUD APK | Percentage of Gross Participation Rates of Population Aged 3-6 Years Broken Down by Regency/City in 2022 |
| UMK (Regency/City Minimum Wage) | Lowest wages (including regular allowances but excluding overtime wages) paid to employees (per type of position/job) |

1. **Data Pre-processing and Exploration**

At this stage, the data pre-processing process is carried out with the aim of preparing the data before being analyzed using the K-Means, Support Vector Machine and Decision Tree methods. The initial stage in this process is to check for missing values, remembering that the three algorithms used cannot handle data that has missing values. A more detailed description of the checking process can be seen in the illustration presented in Figure 3.

A screenshot of a computer code

Description automatically generated

*Figure 3. Checking Missing Value*

Based on the checking results, it can be concluded that the dataset used does not contain missing values. Next, the next stage is to check for outliers by focusing on evaluating non-numerical variables in the dataset. This outlier identification process refers to the interquartile value method, the details of which can be examined and understood through the illustration given in Figure 4.

A screenshot of a computer

Description automatically generated

*Figure 4. Checking for Outliers Using Interquartile*

Figure 4 indicates that the initial size of the dataset is 520 rows and 22 columns. Post outlier elimination, a number of rows have been eliminated, reducing the size of the dataset to 274 rows, although the number of columns has not changed. The next step in checking for outliers continues by using a boxplot, as depicted in Figure 5.

A graph of a box plot

Description automatically generated

(a)

A graph with lines and dots

Description automatically generated

(b)

*Figure 5. Boxplot of Numerical Variables*

Outliers were found in each variable as shown in Figure 5. Removing outliers will not be considered, as it can reduce crucial information from the dataset. Therefore, in this research, outlier handling will utilize techniques that are resistant to the impact of outlier values ​​and use data transformation. The chosen transformation technique will adapt to the distribution characteristics of each variable. As a first step, a Kernel Density Estimation (KDE) plot will be prepared to explore the data distribution pattern in more detail.

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

*Figure 6. Kernel Density Estimation (KDE) Plot*

From the evaluation of the skewness values ​​for each variable, it can be concluded that "Stunting Prevalence (TB/U) %" (Skewness: 13.017), "BPNT 40%" (Skewness: 1.64), and "KKS 40%" (Skewness: 1.09 ) shows a significant positive skewness tendency. Skewness with a significant positive number describes a data distribution with a long tail towards higher values, while most of the data is at lower values. Meanwhile, the variable "K4" with a Skewness of -1.401 indicates that there is a skew in the data distribution to the right side, where the majority of data converges at higher values. This skewness value reflects the plot of each variable.

In order to reduce asymmetric distribution, standardization of the data is carried out. This is done with the aim of making the data distribution more symmetrical and reducing the influence of values ​​that are outside the normal range. The standardization process is aimed at improving the ability to interpret analysis results and ensuring the reliability of statistical models that will be used subsequently.

*A screenshot of a computer

Description automatically generatedFigure 7. Standardized Data*

Next, additional variables are created that contain ordinal data for classification analysis purposes. This process is essential because in the existing dataset, no categorical variables are available. The approach to creating additional variables refers to WHO recommendations for classifying stunting prevalence, utilizing four of the five predetermined categories, namely low, medium, high and very high [19]. An illustration of this grouping process can be seen through a bar graph, as seen in Figure 8.

A graph of a number of bars

Description automatically generated

*Figure 8. Distribution of Regencies/Cities Based on Stunting Prevalence Categories*

In the very high stunting prevalence category, there are 36 regencies/cities, for example Bondowoso Regency. Meanwhile, there are 107 regencies/cities classified as high, while those classified as medium are 162, and those in the low category are 215 regencies/cities. If you look at this diagram, you can see that the majority of districts/cities in Indonesia have implemented stunting management policies well, as indicated by the large number of areas that fall into the low category. The next step involves the selection process of the variables that have the most significant influence, based on the correlation value with the prevalence of stunting in each district/city as the response variable.

A chart with red and blue squares

Description automatically generated

*Figure 9. Correlation of Numerical Variables with Response Variables*

Predictor variables that influence the percentage of "Stunting Prevalence (TB/U)" are selected based on significant correlation with the response variable. Predictor variables that showed a significant positive correlation involved “LBW,” while variables that had a significant negative correlation included “FASYANKES DELIVERY,” “K4,” “Complete KF,” “mCPR,” and “A 6-11.” Variable selection was carried out by considering a correlation value approaching ±0.5 and not less than ±0.2, indicating a strong relationship. Other variables were ignored because they had close to zero correlation, indicating the inadequacy of the relationship to explain the response variable.A screenshot of a graph

Description automatically generated

*Figure 10. Graph of distribution of numerical variables on the prevalence of stunting*

Based Figure 10 that the data points for variables that have a correlation of close to ±0.5 and at least ±0.2 show a certain distribution pattern. On the other hand, points on variables with correlations outside this interval are spread randomly. From this analysis, it can be concluded that the variable "LBW" has a significant positive relationship with the prevalence of stunting. On the other hand, the variables "FASYANKES DELIVERY", "K4", "Complete KF", "mCPR", and "A 6-11" show a significant negative relationship with the prevalence of stunting in various districts/cities.

1. **Data Sharing**

The next step involves dividing the data into two groups, namely training data and testing data. The purpose of this step is to evaluate the performance of the model that has been developed. In the context of this research, an 80:20 ratio is applied, which indicates that 80% of the dataset will be used as training data, while the remaining 20% ​​will be used as testing data. It is hoped that the distribution of this dataset can provide adequate training for machine learning models, with the hope of increasing the accuracy of research results.

1. **Data Modeling**
   * 1. K-means Clustering

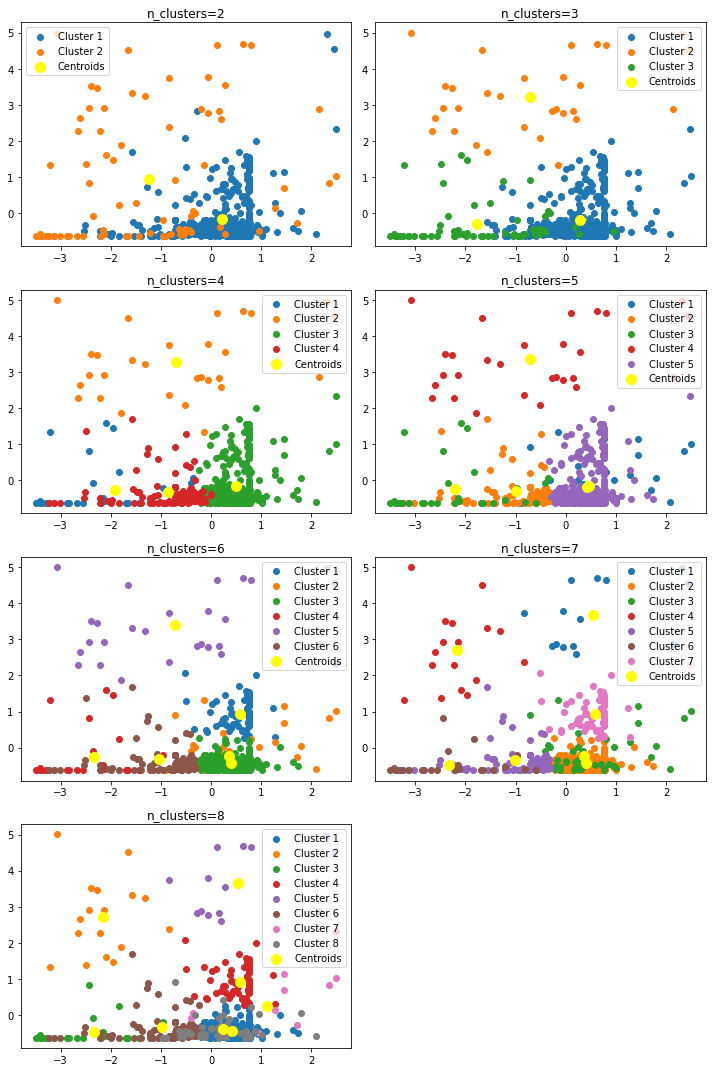
Cluster analysis techniques are used in data mining for exploration and classification of data based on the similarity of characteristics between observations [20]. Through this approach, observations in one group show more similar characteristics compared to observations in other groups. This research attempts to categorize districts/cities in Indonesia based on the similarity of predictor variables in influencing the response variable. Determining the number of clusters is based on the maximum Silhouette value, describing the optimal level of homogeneity and separation between the clusters.

A graph with blue lines and numbers

Description automatically generated

*Figure 11. Determination of the Number of Clusters*

In Figure 11, it can be seen that the Silhouette value reaches its peak when the number of clusters is equal to two, with a Silhouette value of 0.389. Therefore, regencies/cities in Indonesia will be grouped into two. Further information regarding the grouping results can be accessed in the visualization presented in Figure 12.



*Gambar 12. Selection of the Most Appropriate Number of District/City Clusters*

If Figure 11 shows the Silhouette value for each number of groups or clusters, then Figure 12 shows a visualization of the grouping. Each cluster group is marked with a different color. A good grouping is one that can clearly separate the boundaries between cluster group members, so that the number of clusters equal to two is chosen. It can be seen that group 0 is represented by a blue dot while group 1 is represented by an orange dot.

The characteristics of group 0 show a lower "Stunting Prevalence (TB/U) %" level, namely around 13.55%, in contrast to group 1 which achieved a stunting prevalence level of around 25.20%. Group 0 also showed superior performance in most health indicator variables, such as "K4", "FASYANKES Delivery", "Complete KF", "Maternal Vitamin A", "pregnant women TTD", "LBW", "IMD", " ASI”, “CPKB”, etc., with relatively higher average values.

In contrast, cluster 1 displays slightly smaller values ​​for several of these indicators. From these findings, cluster 0 can be seen as a group with more optimal public health conditions, while cluster 1 indicates a tendency to have a higher prevalence of stunting and more severe health challenges. However, the boundary between the two clusters appears blurred, marked by a mixed distribution of points, each cluster being almost the same. Thus, the use of the K-Means method in this study can be considered not to provide useful information.

*A graph with numbers and a line

Description automatically generatedFigure 13. Calinski-Harabasz index*

Evaluation of the performance of Regency/City groupings in Indonesia based on stunting prevalence predictor variables using the Calinski-Harabasz index found that the optimal number of clusters was two. Better performance quality is reflected in high index values. It should be emphasized that at least two clusters are needed to group observations, so these findings support the researcher's view that the application of the k-means clustering method may not be necessary in the context of this research.

* + 1. *Support Vector Machine* (SVM)

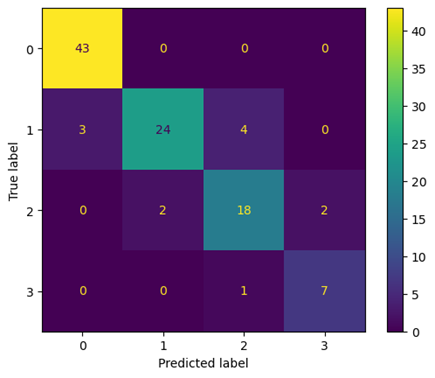
The creation of the SVM model has been carried out four times using linear, rbf, sigmoid and poly kernels. Detailed performance evaluation of each model can be referred to in Table 2.

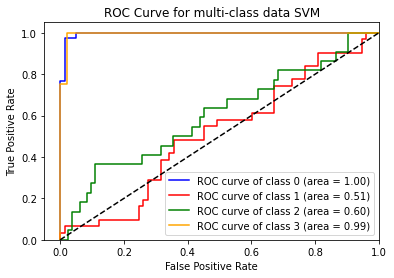
|  |
| --- |
| (kernel rbf) |
| (kernel linear) |
| (kernel sigmoid) |
| (kernel poly) |

*Table 2. Comparison of the goodness of SVM models*

After analyzing the quality of the SVM model based on the various kernels used, it was found that the model using the linear kernel showed the best performance with an accuracy of 88%. These results differ from previous findings which suggested that the rbf kernel was more effective [21]. It is possible that this difference occurs because the data structure has a linear tendency, so a linear kernel is the most appropriate option. The linear kernel is specifically designed for data that can be separated by straight lines. So far, the RBF kernel is recommended for data that requires representation in higher dimensions so that it can be separated more efficiently.

In the precision aspect, the model shows very good performance by reaching 93% for low level classification and 92% for medium level classification. However, recall in the middle category decreased slightly to 77%. However, the overall model F1-score reached 88%, indicating an optimal balance between precision and recall. Overall, the model is able to effectively identify low-level stunting cases, but there is potential to improve detection of mid-level cases. Thus, it can be concluded that this model is effective in recognizing stunting cases at the low level and shows adequate performance at the medium level.



****

*Gambar 14. Confusion Matrix Model SVM Kernel Linear*

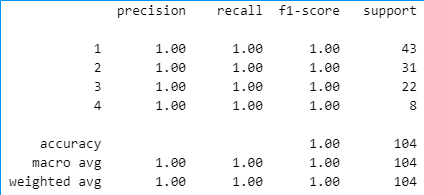
In Figure 14, the model succeeded in categorizing 43 regencies/cities into the low category, 24 regencies/cities into the middle category, 18 regencies/cities into the high category, and 7 regencies/cities into the very high category accurately. However, there were several errors in the classification, namely 3 regencies/cities which were actually in the middle category but were categorized as low, 2 regencies/cities which should have been in the high category but were classified as medium, 1 regencies/cities which should have been in the very high category but are categorized as high, 4 regencies/cities which are actually in the middle category but are classified as high, and 2 regencies/cities which should be in the high category but are categorized as very high.

The ideal ROC curve is one that approaches the top left corner of the graph. A curve like this shows that the classification model has a high true positive rate (TPR) and a low false positive rate (FPR). An area value that is closer to one indicates that the classification in that category is getting better or ideal. In the case of multi-class data, where there are four categories of stunting, a ROC curve can be plotted for each category. The ROC curve for each category can be interpreted in the same way as the ROC curve for binary data

The ROC curve in the figure shows that the SVM classification model has good performance for categories 0 (low) and 3 (very high). This finding is supported by the area value for the low category of 1 which means it reaches a perfect score and the area value of 0.99 for the very high category. Category 1 (intermediate) classification performance is random because the area value is 0.51. Meanwhile, classification performance for category 2 (high) is relatively better compared to category 1 with an area value of 0.6. This means that the classification performance for categories 1 and 2 uses a random model.

* + 1. *Decision Tree*

Evaluation of classification results based on the WHO (World Health Organization) stunting category shows that the Decision Tree model works well.

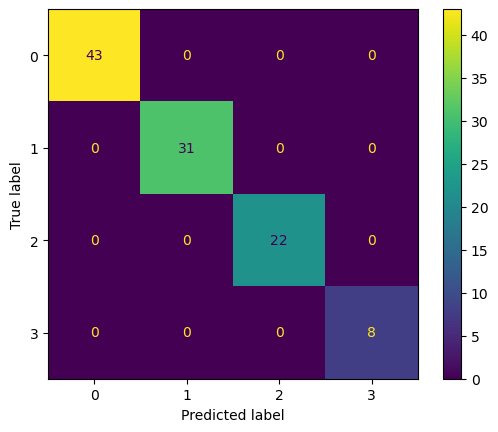


*Figure 15. Goodness of the Decision Tree Model*

Model evaluation results *Decision Tree* showed very satisfactory performance in classifying stunting factors in Indonesia into four different classes, namely low, medium, high and very high. From metrics precision*, recall*, and*f1-score* each of which reaches a value of 1.00 for each class, this indicates that the model has perfect ability to identify and differentiate between these classes without errors.

In detail, precision and recall for all classes achieve a score of 1.00. This means that all data entered into each class was identified correctly without errors. Furthermore, *F1-score* also achieved a score of 1.00 for each class, indicating that this model is good at recognizing and remembering data very well.

Overall, these results illustrate that model *Decision Tree* which was developed has high reliability for classifying stunting factors in Indonesia into four classes with an accuracy of 100%.



A white rectangular object with black lines and text

Description automatically generated with medium confidence

Figure 16*. Confusion Matrix Model Decision Tree*

This model is more reliable in analyzing and interpreting factors that contribute to the prevalence of stunting in Indonesia than the Support Vector Machine model. This statement is supported by the fact that there were no regencies/cities that were wrong in the classification process. If we review it based on the area value, it also shows that the classification performance for each category uses an ideal model with a score of one.

1. **CONCLUSIONS AND RECOMMENDATIONS**

This research assesses the success of various data analysis techniques, such as K-Means, Support Vector Machine (SVM), and Decision Tree, in identifying stunting levels in 520 districts/cities in Indonesia. The results of selecting significant predictor factors show that the variable "LBW" has a significant positive relationship with the prevalence of stunting. Meanwhile, the variables "FASYANKES Delivery", "K4", "Complete KF", "mCPR", and "A 6-11" show a significant negative relationship with the prevalence of stunting in various districts/cities. From the results, K-Means did not produce real differences between the clusters formed, so the analysis was less in-depth. In contrast, SVM achieved an accuracy rate of 88%, especially when using linear kernels. Decision Tree, on the other hand, records a perfect accuracy of 100%, and in general, the Decision Tree method outperforms SVM in classification.

Using the K-Means technique has succeeded in grouping 520 regencies/cities into two clusters. The first cluster shows a lower stunting prevalence rate than the second cluster. However, this grouping method is not optimal because there are no clear criteria to differentiate the two clusters. Therefore, it is recommended that the Indonesian government continue to implement programs that have been proven effective in reducing the prevalence of stunting, especially in districts/cities that are included in cluster 1.

This study provides a new perspective in comparing the effectiveness of the various analytical methods used. For future research, further exploration should be carried out on alternative grouping techniques or machine learning algorithms to obtain a more in-depth interpretation. The results of this research can be a foundation for developing more specific and adaptive health intervention plans in each region.

**REFERENCE**

|  |  |
| --- | --- |
| [1] | Ministry of Health of the Republic of Indonesia, "Kesmas Warta," 2023. |
| [2] | H. Dasman, "Four Impacts of Stunting on Children and the Indonesian State,"*The Conversation,* 2019. |
| [3] | Ministry of National Development Planning/Bappenas, "Presidential Regulation Number 18 of 2020," 2020. [Online]. Available: https://jdih.bappenas.go.id/peraturan/detailperaturan/1037. |
| [4] | D. H. Manongga, S. P. Manongga and H. Hendry, "PREDICTING A STUNTING PREVALENCE,"*International Journal of Innovative Computing, Information, and Control,* vol. 19, 2023. |
| [5] | D. Septyha, K. Rahayu, S. Rabbani, V. Fitria, Rahmaddeni, Y. Irawan and R. Hayami, "Implementation of the Decision Tree Algorithm and Support Vector Machine for Lung Cancer Classification,"*MALCOM: Indonesian Journal of Machine Learning and Computer Science,* vol. 3, pp. 15-19, 2023. |
| [6] | S. Syahrial, R. Ilham, Z. F. Asikin and S. S. I. Nurdin, "Stunting Classification in Children's Measurement Data Using Machine Learning Models," *Newspaper La Multiapp,*2022. |
| [7] | D. WIdyawati, A. Faradibah and P. L. L. Belluano, "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indonesian Journal of Data and Science,* vol. 4, 2023. |
| [8] | G. V. S. Charan and N. S. Kumar, "Analysis and Comparison for Innovative Prediction Technique of COVID-19 using Decision Tree Algorithm over the Support Vector Machine Algorithm with Improved Accuracy," *Cardiometry,* 2022. |
| [9] | 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," *Children,* 2022. |
| [10] | K. N. S. S. Raj and K. Thinakaran, "Prediction of Heart Disease using Decision Tree over Logistic Regression using Machine Learning with Improved Accuracy," *Cardiometry,* no. 25, pp. 1514-1519, 2022. |
| [11] | A. Nugroho, H. L. H. S. Warnars, F. L. Gaol and T. Matsuo, "Trend of Stunting Weight for Infants and Toddlers Using Decision Tree," *IAENG: International Journal of Applied Mathematics,*2022. |
| [12] | M. Ula, A. F. Ulva, Mauliza, M. A. Ali and Y. R. Said, "APPLICATION OF MACHINE LEARNING IN DETERMINING THE CLASSIFICATION OF CHILDREN'S NUTRITION WITH DECISION TREE," *JUTIF: Journal of Information Engineering,*vol. 3, 2022. |
| [13] | A. Byna, "Comparative Analysis of Machine Learning Algorithms for Classification about Stunting Genesis," pp. 74-84, 2019. |
| [14] | Scikit Learn, "Selecting the number of clusters with silhouette analysis on KMeans clustering," 2023. [Online]. Available: https://scikit-learn.org/stable/auto\_examples/cluster/plot\_kmeans\_silhouette\_analysis.html. [Accessed 2023]. |
| [15] | D. Abdullah, Susilo, A. S. Ahmar, R. Rusli and R. Hidayat, "https://link.springer.com/article/10.1007/s11135-021-01176-w," [Online]. Available: https://link.springer.com/article/10.1007/s11135-021-01176-w. |
| [16] | N. Pratiwi and Y. Setyawan, "ACURACY ANALYSIS OF DIFFERENCES IN KERNEL AND COST FUNCTIONS IN SUPPORT VECTOR MACHINE CASE STUDY OF RAINFALL CLASSIFICATION IN JAKARTA,"*Journal of Fundamental Mathematics and Applications (JFMA),* vol. 4, 2021. |
| [17] | H. Harafani and R. S. Wahono, "Parameter Optimization on a Support Vector Machine Based on Genetic Algorithms for Estimating Forest Fires,"*Journal of Intelligent Systems,* vol. 1, 2015. |
| [18] | MathWorks, "Support Vector Machine," 2023. [Online]. Available: https://www.mathworks.com/discovery/support-vector-machine.html. |
| [19] | World Health Organization, "Nutrition Landscape Information System (NLiS)," 2018. [Online]. Available: https://apps.who.int/nutrition/landscape/help.aspx?menu=0&helpid=391&lang=EN. |
| [20] | Columbia Mailman School of Public Health, 2023. [Online]. Available: https://www.publichealth.columbia.edu/research/population-health-methods/k-means-cluster-analysis. |
| [21] | A. Z. Praghakusma and N. Charibaldi, "Comparison of Kernel Functions of Support Vector Machine Methods for Instagram and Twitter Sentiment Analysis (Case Study: Corruption Eradication Commission),"*Journal of Undergraduate Information Engineering,*vol. 9, pp. 33-42, 2021. |
| [22] | Liu and Y. Zhu, "On the consistent estimation of optimal Receiver Operating Characteristic (ROC) Curve`," in *Neural Information Processing Systems*, 2022. |