CLASSIFICATION OF STUNTING AND DISTRIBUTION OF REGIONAL ZONES USING SUPPORT VECTOR MACHINE AND EXTREME GRADIENT BOOST METHOD IN CIAWI DISTRICT

Bira Arya Setha[1*], Tjut Awaliyah Zuraiyah[2], Siska Andriani[3]
[1] Department of Computer Science, Pakuan University, Bogor, Indonesia
[2] Department of Computer Science, Pakuan University, Bogor, Indonesia
[3] Department of Computer Science, Pakuan University, Bogor, Indonesia

[*]E-mail: tjut.awaliyah@unpak.ac.id

Abstract. *Stunting* is one of the leading causes of morbidity and mortality in children under the age of five in most developing countries. The main purpose of this study is to design a model for the classification and distribution of stunting zones using a two-based model, namely Support Vector Machine (SVM) and EXtreme Gradient Boosting (XGBoost). This study uses data from the Citapen health center starting from February 1 2024 to February 29 2024, which has 2769 data and 33 variables. Classification is carried out using the Support Vector Machine (SVM) algorithms with 4 kernels namely RBF, Linear, Sigmoid, and Poly then EXtreme Gradient Boosting (XGBoost) algorithms. The first stage was carried out, namely the data cleaning and transformation stage, after which the division of training data and test data was carried out, namely 70:30. The model in this study was evaluated with a confusion matrix. So it is known that the Support Vector Machine (SVM) algorithm is 97.83% with its 4 kernels namely RBF is 97.83%, Linear is 99.04%, Sigmoid 97.83%, Poly is 97.83% and EXtreme Gradient Boosting (XGBoost) is 99.88%. The comparison shows that the XGBoost algorithm produces more accurate classification values than the Support Vector Machine (SVM) algorithms which are then selected to create a web application system using the Streamlit framework. Based on a series of processes that have been carried out in this study, it can be concluded that this study has succeeded in building the best model and web application for stunting prediction using the XGBoost algorithm.

1. Introduction

The main nutritional problem facing Indonesia is stunting, commonly known as Stunting Toddler. Stunting is a condition in which children experience chronic malnutrition over a long period and do not receive adequate nutrition, resulting in slow growth and height inappropriate for their age. Growth failure can begin during fetal development and becomes apparent after the child reaches 2 years. [1] Stunting affects children's lives if they do not receive fast and good treatment. [2]

Stunting can occur due to several factors, namely poor parenting, lack of maternal knowledge about health and nutrition before or after childbirth, limited health services, limited family access to nutritious food, and limited access to clean water and sanitation. [1]

The incidence of stunting will have a negative impact on toddlers if not addressed further.The short-term effects of stunting include disruption of physical and mental development, reduced intelligence, and metabolic problems. The long-term impact of stunting is reflected in decreased

cognitive abilities and decreased endurance, making the body susceptible to disease and at risk of degenerative diseases such as diabetes, cardiovascular disease, cancer, stroke, and being unhealthy. being competitive at work causes low productivity. [3]

The results of the Indonesian Nutrition Status Survey (SSGI) announced by the Ministry of Health at the BKKBN National Working Meeting show that the prevalence of stunting in Indonesia has decreased from 17.8% in 2023. According to the Indonesia.go.id website, the government aims to reduce the stunting rate to 14% by the end of 2024.

Detecting stunting in toddlers can be identified early by routinely checking the child's growth chart at the nearest Posyandu. Posyandu administrators measure the condition of toddlers, then the measurements are sent to experts or experts to find out whether the measurements show a slowdown in growth [4] The results of early identification of stunting make it easier for parents and child managers to implement efforts to prevent stunting in the early stages of growth.

Stunting checks are conducted by calculating the z-score based on the anthropometric index set by the Ministry of Health of the Republic of Indonesia with a standard deviation threshold < -3 [5], where the examination of stunting in toddlers is still quite time-consuming because it is done manually and is prone to errors. then a system is needed that can classify toddler examination data so that it can predict quickly and accurately whether the child is stunted or not quickly and accurately. Therefore, so that parents and posyandu officers can more quickly find out the condition of the child, it is necessary to build a system that uses the data mining method to be able to classify the resulting data. In order to quickly predict whether the toddler is in the stunting category or not.

Data mining can be used to extract information from large data stores so as to obtain information that can be used in predicting stunting. One of the data mining methods is classification. Classification techniques are used to group or classify data into certain classes based on specified labels. [6]

The classification process has two stages, the first stage is training, which analyzes the training data using the classification algorithm. And the second is classification, which is predicting the accuracy of classification using test data. [7]

There are several methods for classification, including the eXtreme Gradient Boosting (XGBoost) and Support Vector Machine (SVM) methods. XGBoost (Extreme Gradient Boosting) is a combination algorithm between boosting and gradient boosting. XGBoost can perform various functions such as regression, classification, and ranking [8].XGBoost is one of the tree ensembles algorithms consisting of several classification and regression trees (CART). The XGBoost method is a gradient tree boosting algorithm based on an ensemble algorithm that can effectively handle large-scale machine learning cases, which is developed with 10 times faster optimization than Gradient Boosting. The most important success factor of XGBoost is scalability in various scenarios. This scalability is due to the optimization of the previous algorithm [9]. The XGBoost method was chosen because it has several additional features that are useful for speeding up computer systems and avoiding overconfiguration.

Whereas, SVM is a model derived from statistical learning theory and will give better results than other methods. In Linear SVM, each training data is known as (xi, yi), where i=1,2,...,N and xi={xi1,xi2,...,xiq} T is the attribute for training data I, yi {-1, +1} is the Label Class [10]. SVM (Support Vector Machine) has proven to be effective in classifying the nutritional status of toddlers based on four anthropometric indicators. This method uses attributes such as gender, age, weight, height and body mass index in the classification process [11].

Information on the classification of the nutritional status of children under five is needed to determine the areas affected by the most stunting. Spatial analysis with Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost) methods, can provide information that describes the facts of the region as decision making to support information and easy to understand.

The machine learning model that has been created can be incorporated into a web-based program so that users can easily perform classification. One of the frameworks that supports the

deployment of models into web-based programs is streamlit. Streamlit is a Python-based framework that is open source. There is an urgent need for advanced analytic techniques that can deeply process stunting data with greater depth, agility, and security, allowing real-time insights through interactive web applications. The framework will be used to implement pre-tested classification algorithms. [12]

Several previous studies related to stunting classification using various algorithms have been conducted, including: (Drajana et al., 2022) which uses K-Nearest Neighbor Based on Chi Square Feature Selection to predict the Status of Stunting Patients in Toddlers in Gorontalo Province [13]. Then (Adzhima et al., 2023) classified Toddler Stunting Status with the Web-Based Support Vector Machine Method [14]. Furthermore (Damayanti et al., 2023) Classifying Toddler Stunting Status Using the Fuzzy C-Means Algorithm [15]. In addition, (Lonang et al., 2022) classifies Stunting Status in Toddlers Using K-Nearest Neighbor with Feature Selection Backward Elimination [7]. And the last one (Herliansyah et al., 2021) which predicts Stunting in Toddlers Using the Naïve Bayes Classification Algorithm [16]. Each study applies a different algorithm method for stunting classification with the aim of improving prediction accuracy and supporting health decisions.

Based on the background described above, the researchers conducted a study that will classify stunting and the distribution of regional zones using the Support Vector Machine and Extreme Gradient Boost methods in Ciawi District.

2. Theoretical Foundations

2.1 Stunting

Stunting is a nutritional status based on the BB/U or TB/U index where in the anthropometric standard of assessing the nutritional status of children, the measurement results are at the threshold (Z-Score) <-2 SD up to -3 SD (short / stunted) and <-3 SD (very short / severely stunted). Stunting is a chronic malnutrition problem caused by insufficient nutritional intake for a long time due to feeding that is not in accordance with nutritional needs. [17]

2.2 Machine Learning

Machine Learning is the ongoing study of the concepts of pattern recognition and computational learning in artificial intelligence that uses learning algorithms such as supervised and unsupervised to predict and support automated decision making based on a set of data. (Wardhana et al., 2023). Machine learning is divided into three categories: Supervised Learning, Unsupervised Learning, Reinforcement Learning. [18]

2.3 Classification

Classification is a critical understanding of objects that involves making sequences, groups or classes, and it gives a certain meaning to a reality. [19]

2.4 Spatial

Spatial analysis is a set of techniques that can be used in GIS (geospatial and information systems) data processing. This analysis technique is used to identify morphological changes and development patterns of an area. [20]

2.5 Support Vector Machine (SVM)

The Support Vector Machine (SVM) algorithm itself is an algorithm that aims to find the maximum hyperplane, hyperplane is a function that can separate between two classes. In the process SVM will maximize the margin or distance between the training pattern and the decision boundary. There are several advantages of this algorithm, including having good performance whether it is used with small or large amounts of data, has good performance on data that has many attributes and is easy to implement. [21]

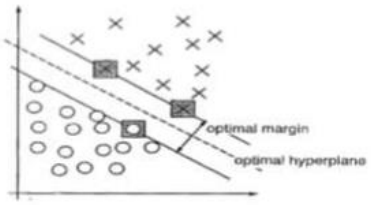

Figure 1. Conceptual Approach to the Support Vector Machine Method

The data in a dataset is given the variable $x\,i$, while the class in the dataset is given the variable $y\,i$. The SVM method divides the dataset into 2 classes. The first class separated by the hyperplane is worth 1, while the other class is worth -1. [22]

$X\,i\ .W\ + \geq 1$ of $Y\,i\ = 1$ (1)

$X\,i\ .W\ + b\ \leq$ -1 of $Y\,i\ =$ -1 (2)

Description:

$X\,i\ =$ data to -i

$W\ =$ support vector weight value perpendicular to the hyperplane

$b\ =$ bias value

$Y\,i\ =$ data class to -i

Weight vector (w) is a vector line perpendicular between the coordinate center point and the hyperplane line. Bias (b) is the coordinates of the line relative to the coordinate point. Equation 3 is the equation to calculate the value of b, while equation 4 is the equation to find the value of w.

$b\ = -\frac{1}{2}(w\ .x^{+}+w\ .\ x^{-})$ (3)

$w\ = \sum_{1\ =\ 1}^{n} \alpha\ i\ y\ i\ x\ i\ n\ i =1$ (4)

Description:

$b\ =$ bias value

w .x^+ = weight value for positive data class

w . x^- = weight value for negative data class

$w\ =$ weight vector

$\alpha\ i\ =$ weight value of i-th data

$y\ i\ =$ data class to-i

$x\ i\ =$ data to-i

H1 is the supporting hyperplane of class +1 which has the function $w\ x\ 1+b\ = +1$.

Margin =|dH1-dH2 | $= \frac{2}{||w||}$ (5)

Description:

dH1 = class support hyperplane distance +1

dH2 = distance hyprplane supporting class -1

Then to determine the optimal hyperplane of both classes using the following equation:

MinimizeJ $1[w\ ]= \frac{1}{2}\,||w||^2$ (6)

with $(x\ i\ .w\ +b\ ) - 1 \geq 0, = 1,...,$ (7)

Figure 2 shows some patterns that are members of two classes: +1 and -1. Patterns belonging to class -1 are symbolized with red squares and patterns belonging to class +1 are symbolized with yellow circles. The classification problem can be explained by finding the hyperplane line that separates the two groups. The solid line in Figure 2 shows the best hyperplane, which is located in the middle of the two classes, while the red and yellow dots inside the black circle are support vectors.

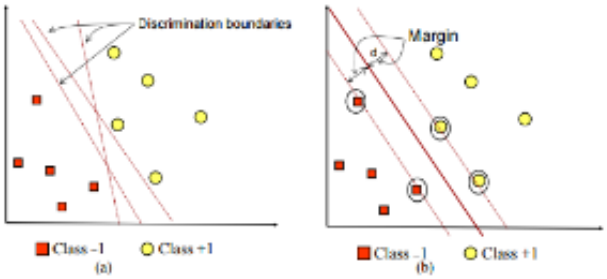

Figure 2. The best hyperplane that separates the two negative and positive classes

The Support Vector Machine (SVM) method was chosen because it has proven to have good accuracy for analyzing text. Kernel is a set of mathematical functions owned by Support Vector Machine (SVM). This kernel functions to take data as input and then convert the data into a high-level dimensional space (Kernel Space). Popular kernels in Support Vector Machine (SVM) include linear kernel, polymonical function, sigmoid, Gaussian radial basis function (RBF) [23]. The requirement for a function to be a kernel function is to fulfill the Mercer Theorem which states that the resulting kernel matrix must be positive semi-definite. The types of Support Vector Machine (SVM) kernels can be seen in the following table:

Linier

$$K(\underset{x_l}{\rightarrow},\underset{x_j}{\rightarrow}) = \underset{x_l}{\rightarrow}^{\mathrm{t}} \mathrm{j} \quad (8)$$

a. *Polynomial*

$$K(\underset{x_l}{\rightarrow},\underset{x_j}{\rightarrow}) = (\underset{x_l}{\rightarrow},\underset{x_j}{\rightarrow} + 1)^p \quad (9)$$

b. *Gaussian*

$$K(\underset{x_l}{\rightarrow},\underset{x_j}{\rightarrow}) = \exp - \frac{||\underset{x_l}{\rightarrow},\underset{x_j}{\rightarrow}||}{2\sigma^2} \quad (10)$$

c. *Sigmoid*

$$K(\underset{x_l}{\rightarrow},\underset{x_j}{\rightarrow}) = \tanh(a \underset{x_l}{\rightarrow},\underset{x_j}{\rightarrow} + \beta) \quad (11)$$

Data that cannot be classified linearly can use methods by transforming the data into a feature space. Feature spaces usually have higher dimensions than input vectors (input space). This results in computation on the feature space is very large infinite number of features and also difficult to know the right transformation function. To overcome these problems, the Support Vector Machine (SVM) uses kernel tricks. In this study using the Support Vector Machine method with (SVM) the kernel used is a linear kernel for data processing, the results of different dataset compositions are used to find the best accuracy value. [23]

2.6 Extreme Gradient Boosting

EXtreme Gradient Boosting or XGBoost is one of the boosting methods, which is a collection of DTs whose construction of the next tree will depend on the previous tree. The first tree in XGBoost will be weak in classification with an initialization probability determined by the researcher and then the weights will be updated on each tree built so as to produce a collection of strong classification trees. [24]

This method requires an objective function which is useful for assessing how well the model obtained matches the training data [. The most important characteristic of the objective function consists of 2 parts, namely the missing training value and the regularization value as in the following equation 12. [8]

$$obj(\theta) = L(\theta) + \Omega(\theta) \quad (12)$$

Where L is the missing training function, and Ω is the regularization function, and θ is the corresponding model parameter. The missing training function can be generally written as in equation 13 as follows.

$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y_i})$ (13)

Where $y\,i$ is the actual data value that is considered correct and $y\,i\hat{}$ is the result of the predicted value of the model, while n is the number of iterations of the value of the model. The following are the stages of finding the XGBoost value

1. Find the Residual Value of the data
2. Determining Lambda Value (ʎ)
3. Finding the similarity value

   $Ss = \frac{SR^2}{Total\ Res + ʎ}$ (14)

   Description:
   SR = amount of residual results
   ʎ = Lambda
4. Search for branch similarity values
5. Finding the gains value

   $Ssx_1 + nilai\ Ssx_2 - Ss$ (15)

   Description
   Ssx = branch similarity value

2.7 Confusion Matrix

Confusion Matrix is a matrix-shaped measurement tool used to obtain the amount of classification accuracy of the class with the algorithm used [25]. The following will present the form of Confusion Matrix in Table 1.

Table 1. Confusion Matrix

Confusion Matrix		**Actual Value**	
		True	***False***
Predicted Value	***True***	TP *(True Positive) Correct Result*	FP *(False Positive) Unexpected result*
	False	FN *(False Negative) Missing result*	TN *(True Negative) Correct absence of result*

In Table 1, the TP (true positive) and TN (true negative) values show the level of classification accuracy. Generally, the higher the TP and TN values, the better the classification level of accuracy, precision, and recall. If the predicted output label is true and the actual value is false, it is called a false positive (FP). Meanwhile, if the predicted output label is false and the actual value is true, this is called a false negative (FN). The following formulations for calculating accuracy, precision, and recall in the formation of classification models are shown in equation (16), equation (17), and equation (18).).

$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \text{X}100\%$ (16)

$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \text{X}100\%$ (17)

$Recall = \frac{\text{TP}}{\text{TP+FN}} \text{x}100\%$ (18)

The Accuracy formula in the multiclass confusion matrix this time is slightly different from the formula in the case of binary classification [26]. In this case the formula used is

$Accuracy = \frac{\text{TP}}{Ammount\ of\ Data}$ (19)

To search diagonally for the TP value contained in the cell value.

2.8 Normalization Z-Score

Normalization is the process of scaling the attribute values of the data so that they can lie within a certain range [27]. The following are some of the normalization stages carried out:

1. Calculate the average value (mean) with the equation formula 20

$$\overline{X} = \frac{x1+x2+\cdots+n}{n} \quad (20)$$

2. Calculate Standard Deviation with equation 21

$$S = \frac{\sqrt{\sum fi(xi-\overline{X})x^2}}{\sum n} \quad (21)$$

3. Calculate Z-Score using equation 22

$$z = \frac{(x-\overline{X})}{s} \quad (22)$$

Description:
S = standard deviation
xi = center value
$\overline{X}$ = mean / average value
x = data to be found average
n = number of data

2.9 Numpy

Numpy is a function library that allows you to perform many common data manipulation tasks with Python. Many of the interactions between NumPy arrays and Python are very similar to what you would do with regular Python variables. For more advanced tasks, NumPy also has functions that work with linear algebra, Fourier transforms, and matrices." [28]

2.10 Pandas

Pandas provides data structures used in the python programming language, has a library. Pandas is a library used for data manipulation and analysis. Pandas provides data structures such as Data Frames that enable efficient data processing and presentation. [29]

2.11 Matplotlib

Matplotlib is a library used for data visualization. Matplotlib provides functions and tools for creating graphs, plots, and other visualizations. [30]

2.12 Seaborn

This library was created from the Matplotlib library package. It allows application developers to create varied data visualizations according to their needs (such as mapping colors into variables). Seaborn is more integrated to work with Pandas DataFrames. [31]

2.13 Scikit-Learn

Scikit - Learn is a python module that integrates various machine learning algorithms for medium-scale supervised and unsupervised problems. This module is very efficient for data mining and data analysis. [32]

Scikit-learn or sklearn is a module for the python programming language built on NumPy, SciPy, and matplotlib. Scikit-learn functions to help process data or train data for machine-learning needs. Features that can be used in Scikitlearn include Clustering, Classification, Regression, Dimensionality, Model selection, Data Preprocessing, and Reduction. [29]

2.14 Streamlit

Streamlit is an open-source framework that allows data science and machine learning developers to create interactive web applications quickly and easily using the Python programming language. By using Streamlit, developers can create web applications that display data interactively, graphs, tables, and other interactive features easily. [33]

2.15 Method

Cross-Industry Standard Process For Data Mining (CRISP-DM) is a method that provides a standard process in data mining that is easier to apply because each stage or phase is clearly defined and structured and has a complete and well-documented data mining methodology. [34]

a. Business Understanding

This initial stage of business understanding has a focus on understanding project objectives and requirements from a business perspective, then transforms this knowledge into a problem definition and an initial data *mining* plan designed to achieve the objectives.

b. Data Understanding

The data understanding stage begins with the initial data collection and the results of the activities in order to familiarize themselves with the data, to identify data quality issues, to find the first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

c. Data Preparation

The data preparation stage includes all activities to build the final dataset (the data to be fed into modeling) from the initial raw data. Preparatory data tasks are likely to be performed multiple times and the number is unpredictable. The results of the task include tables, records and attribute selection as well as data transformation and cleaning for modeling.

d. Modeling

In this phase, various modeling techniques can be selected and applied and the model parameters are calibrated with optimal values. Usually, there are several data *mining* techniques that can solve the same problem. Some techniques have certain requirements on data forms. Therefore, stepping back to the data preparation stage is often done.

e. Evaluation

At this stage in the mining project has built a model that appears to be of high quality from the perspective of data analysis. Before proceeding to the final distribution of the model, it is important to more thoroughly evaluate the model and review the steps that have been taken to build the model. This is done to ensure that it is true that the resulting model achieves business objectives. One of the main objectives is to determine whether there are some important business issues that have not been resolved have been sufficiently considered. At the end of this stage, decisions about the use of data and *mining results* must be reached

f. Deployment

Model creation is generally not the end of the project. Even if the purpose of the model is to improve the knowledge of the data, the knowledge gained will need to be organized and presented in a way that customers can use it. It often involves and applies a "live" model in an organization's decision-making process, for example real-time personalization of Web pages or iterative scoring of marketing databases. However, depending on the requirements, the deployment phase can be as simple as possible in generating reports or making it easier to implement the data *mining* process throughout the company. In many cases it is a problem with the customer when the model is applied, not the data analyst. However, even *the deployment* phase is important for customers to understand what needs to be done to truly take advantage of the resulting model.

3. Result and Discussion

3.1 Library and Package Invocation

The invocation of these libraries and packages is done so that Python, as a programming language, can execute commands for processes in machine learning. Here are the libraries used in this research: 1) numpy, as a numerical module for analyzing data in the form of arrays or matrices. 2) matplotlib and seaborn, for data visualization in the form of plots and diagrams. 3) pandas, for processing and reading data used in the program. 4) sklearn, for building machine learning models. The invocation of the library and package can be seen in figure 3.

```
!pip install pandas

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.metrics import confusion_matrix, accuracy_score, classifica
from sklearn.svm import SVC
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
```

Figure 3. Library and Package Invocation

3.2 Data Description

Data description is a type of data analysis used to describe, display, and summarize a set of data. For a description of the data in this study can be seen in Figure 4.

	Usia Saat Ukur (Bulan)	Berat	Tinggi	BB/U	ZS BB/U	TB/U	ZS TB/U	BB/TB	ZS BB/TB
count	2769.000000	2769.000000	2769.000000	2769.000000	2769.000000	2769.000000	2769.000000	2769.000000	2769.000000
mean	31.778982	12.245757	88.472120	0.081979	-0.565775	0.037920	-0.676262	3.665583	-0.271784
std	16.347504	3.028410	12.725166	0.378424	0.804551	0.280005	0.999136	1.057305	0.985676
min	0.000000	2.800000	48.000000	0.000000	-3.570000	0.000000	-5.100000	0.000000	-3.540000
25%	18.000000	10.100000	80.000000	0.000000	-1.120000	0.000000	-1.380000	4.000000	-0.950000
50%	32.000000	12.500000	90.000000	0.000000	-0.620000	0.000000	-0.840000	4.000000	-0.330000
75%	46.000000	14.500000	99.000000	0.000000	-0.110000	0.000000	-0.240000	4.000000	0.350000
max	59.000000	30.500000	120.000000	3.000000	4.510000	3.000000	5.910000	6.000000	7.050000

Figure 4. Data Description

Based on Figure 4, it is known that the data consists of 2769 samples with a total of 9 variables and consists of several classes.

Furthermore, the data will be visualized with a barchart grouping the results of the BB/U, TB/U, and BB/TB indices.

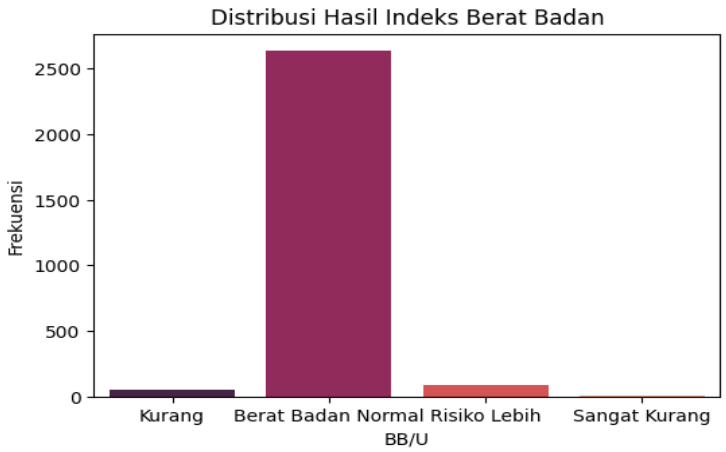

Figure 5. Barchart BB/U

On the Barchart, the normal weight index is very common compared to underweight, overweight, and severely underweight.

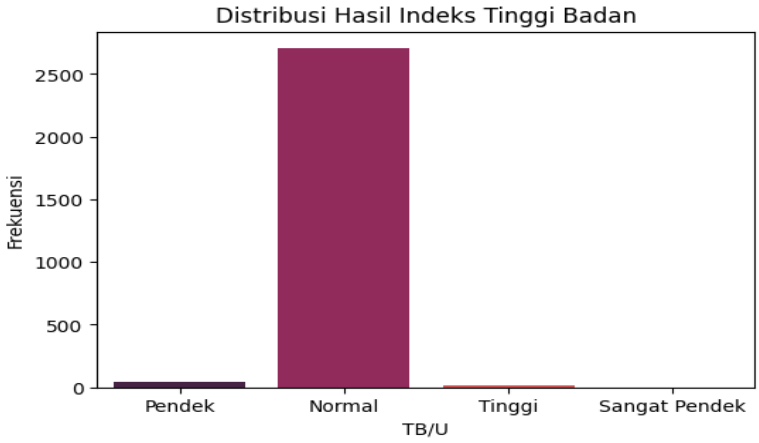

Figure 6. Barchart TB/U

On the Barchart, the normal height index is very common compared to underweight, overweight, and severely underweight.

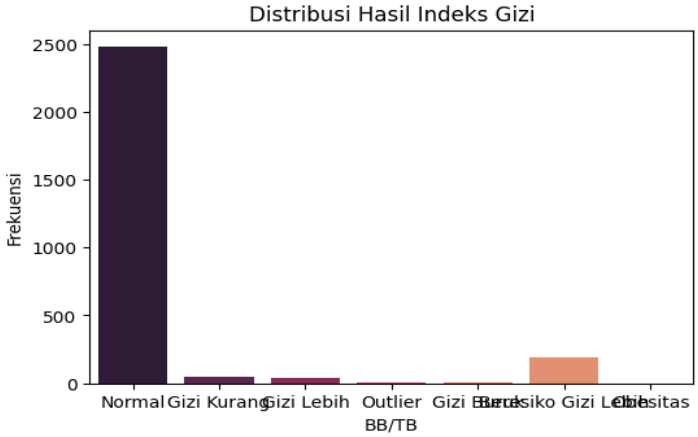

Figure 7. Barchart BB/TB

On the Barchart with normal nutrition is very much found compared to More Risky, Poor, Less, More, Obesity, Outliers.

Then the results will be correlated with Feature correlations. Feature correlations are used to determine the strength of features (variables) against the target (class), where if the variables of age at measurement (month), weight, height, BB/U, ZS BB/U/TB/U, ZS TB/U, BB/TB, and ZS BB/TB have a high value on stunting, then the variable is very influential, and vice versa, if the lower it is, the less influential. Feature correlations can be seen in Figure 8.

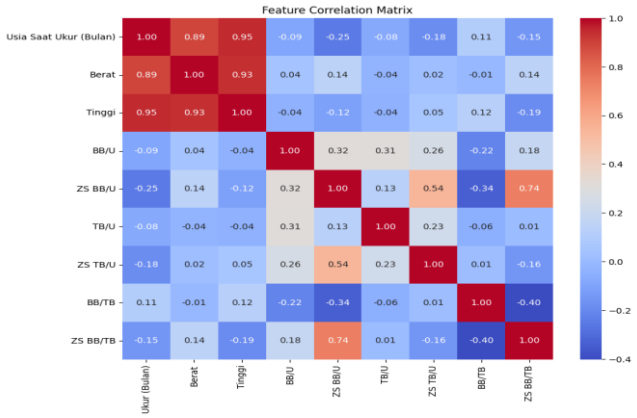

Figure 8. Feature correlations

It can be seen from Figure 8, that if the correlation result is close to the value of 1, then the correlation is said to be good, but if it is close to the value -1, the correlation is said to be bad. It can be seen from the visualization results above, which has a good correlation value is found in the variables “Weight” and “Height”, which has a correlation result value of 0.89 and 0.95, while the poor correlation value is found in the variable “BB / TB”, which has a value of -0.40, if the sideways diagonal correlation result is a variable correlation with itself.

Then a categorical plot visualization is carried out as in Figure 9, the plot is carried out on each variable based on TB / U so as to form a pattern of data distribution on each variable. The distribution of data for toddlers diagnosed with stunting on average has a high value on these variables. This is in line with the correlations features in Figure 8, where the variable has a large value in influencing potentially stunted toddlers.

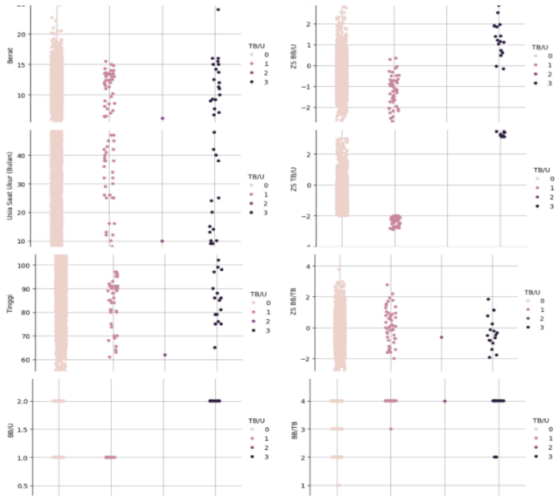

Figure 9. Categorial Plot

3.3 Preparation Data

3.3.1 Data Cleaning

In this process where the previous data that has been obtained will be carried out in the data preprocessing stage in Python. Where for this data preprocessing process, data cleaning is carried out (filling in empty data or null value data). In Figure 15, it can be seen that there is no data with null values.

```
Data.isnull().sum()
```

```
Usia Saat Ukur (Bulan)    0
Berat                     0
Tinggi                    0
BB/U                      0
ZS BB/U                   0
TB/U                      0
ZS TB/U                   0
BB/TB                     0
ZS BB/TB                  0
dtype: int64
```

Figure 10. Data Cleaning

3.3.2 Find Outlier Value

The next stage detects the value of outliers in each variable using the Z-Score method to get outlier data. The results of the outlier search can be seen in Figure 16.

	Usia Saat Ukur (Bulan)	Berat	Tinggi	BB/U	ZS BB/U	TB/U	ZS TB/U	\
0	8	6.5	65.5	1	-2.68	1	-2.55	
52	26	9.7	80.4	0	-1.75	1	-2.20	
66	16	8.5	73.7	0	-1.96	1	-2.31	
74	52	14.8	95.0	0	-1.14	1	-2.52	
234	12	7.7	70.0	1	-2.05	1	-2.12	
...	...	...	...	...	...	...	...	
2583	6	9.0	75.0	2	1.03	3	3.20	
2591	2	6.7	65.0	2	1.39	3	3.11	
2664	2	5.0	53.8	0	-0.92	1	-2.46	
2759	0	4.0	49.0	0	-0.50	1	-2.27	
2761	0	3.9	49.0	0	-0.83	1	-2.41	

	BB/TB	ZS BB/TB
0	4	-1.59
52	4	-0.70
66	4	-1.21
74	4	0.63
234	4	-1.34
...	...	...
2583	4	-0.66
2591	4	-1.00
2664	0	1.77
2759	0	2.19
2761	4	1.93

Figure 11. Outlier Value

3.3.3 Data Tranformation

The initial stage carried out in this data transformation is converting categorical values into numeric ones so that they can be processed in the future. The results can be seen in figure 17.

	Usia Saat Ukur (Bulan)	Berat	Tinggi	BB/U	ZS BB/U	TB/U	ZS TB/U	BB/TB	ZS BB/TB
0	8	6.5	65.5	1	-2.68	1	-2.55	4	-1.59
1	33	12.5	91.5	0	-0.93	0	-1.06	4	-0.56
2	38	12.8	91.4	0	-1.16	0	-1.62	4	-0.40
3	13	9.4	73.0	0	-0.50	0	-1.43	4	0.21
4	38	13.9	93.0	0	-0.23	0	-0.95	4	0.42
...	...	...	...	...	...	...	...	...	...
2764	0	3.5	49.0	0	-1.19	0	-1.95	4	0.67
2765	0	3.7	51.0	0	0.39	0	0.28	4	-0.06
2766	0	3.5	50.0	0	-0.65	0	-1.02	4	0.09
2767	0	3.0	50.0	1	-2.27	0	-1.44	4	-1.74
2768	0	2.8	50.0	0	-1.94	0	-0.71	2	-2.56

2769 rows × 9 columns

Figure 12. Categorial Value to Numeric Value

The numeric values can be seen below.

a. Indeks BB/U
   - 0 = Underweight
   - 1 = Normal
   - 2 = At Risk
   - 3 = Severely Underweight

b. Indeks TB/U
   - 0 = Short (Stunting)
   - 1 = Normal
   - 2 = Very Short (Several Stunting)
   - 3 = Tall

c. Indeks BB/TB
   - 0 = At Risk
   - 1 = Poor
   - 2 = Less
   - 3 = More
   - 4 = Normal
   - 5 = Obesity
   - 6 = Outlier

Furthermore, data transformation is carried out using z-score. It can be seen that there are different values in the data before preprocessing and after preprocessing. The difference that is very contrasting is the median value in the data before and after prepocessing. The data transformation process can be seen in Figure 13.

Usia Saat Ukur (Bulan)	Berat	Tinggi	BB/U	ZS BB/U	TB/U	ZS TB/U	BB/TB	ZS BB/TB	
24		12.0	83.0	0	-0.16	0	-1.44	4	0.83	0.000722
19		9.5	81.0	0	-0.87	0	-0.43	4	-0.92	0.000722
21		10.3	83.0	0	-0.53	0	-0.42	4	-0.47	0.000722
23		11.5	85.0	0	0.07	0	-0.34	4	0.28	0.000722
45		13.5	96.0	0	-1.05	0	-1.18	4	-0.52	0.000722
										...
24		9.4	82.0	0	-1.70	0	-1.43	4	-1.28	0.000361
		9.6	83.0	1	-2.17	0	-1.55	4	-1.92	0.000361
			85.0	0	-1.60	0	-0.43	4	-1.99	0.000361
		9.8	84.0	0	-1.91	0	-1.08	4	-1.93	0.000361
59		18.0	101.0	0	-0.06	0	-1.73	0	1.58	0.000361

Figure 13. Data Transformation

3.3.4 Data Spliting

The dataset in this study is divided into two, namely train data and test data with a spliting percentage of 70% train data and 30% test data. After data division, 1938 training data and 831 test data were obtained and input TB/U Variable to target value. Data division can be seen in Figure 14.

```
[27] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=2)

106] print(X.shape, X_train.shape, X_test.shape)

(2769, 3) (1938, 3) (831, 3)
```

Figure 14. Data Spliting

3.4 Modeling

In this stage, the previously processed data is entered into the SVM and XGBoost algorithm models, then processed to get the best classification results between the SVM and XGBoost algorithms. Modeling can be seen in Figure 15.

```
[29] model_SVC_R = SVC(kernel='rbf', random_state = 4, C=100)
     model_SVC_R.fit(X_train, y_train)

     SVC(C=100, random_state=4)

[31] model_SVC_R = SVC(kernel='sigmoid', random_state = 4, C=100)
     model_SVC_R.fit(X_train, y_train)

     SVC(C=100, kernel='sigmoid', random_state=4)

[30] model_SVC_R = SVC(kernel='linear', random_state = 4, C=100)
     model_SVC_R.fit(X_train, y_train)

     SVC(C=100, kernel='linear', random_state=4)

     model_SVC_R = SVC(kernel='poly', random_state = 4, C=100)
     model_SVC_R.fit(X_train, y_train)

     SVC(C=100, kernel='poly', random_state=4)

models = [
    ('XGBM', xgb.XGBClassifier()),
]
```

Figure 15. Modeling

3.5 Evaluation

Visualization of confusion matrix for prediction of testing data with SVM model obtained 3 classes, namely cell 1 = 813, cell 2 = 0, cell 3 = 0, cell 4 = 15, cell 5 = 0, cell 6 = 0, cell 7 = 3, cell 8 = 0, cell 9 = 0. For the XGBoost model, 3 classes are also obtained,

namely cell 1 = 812, cell 2 = 1, cell 3 = 0, cell 4 = 0, cell 5 = 15, cell 6 = 0, cell 7 = 0, cell 8 = 0, cell 9 = 3. The results of the confusion matrix visualization can be seen in Figure 16.

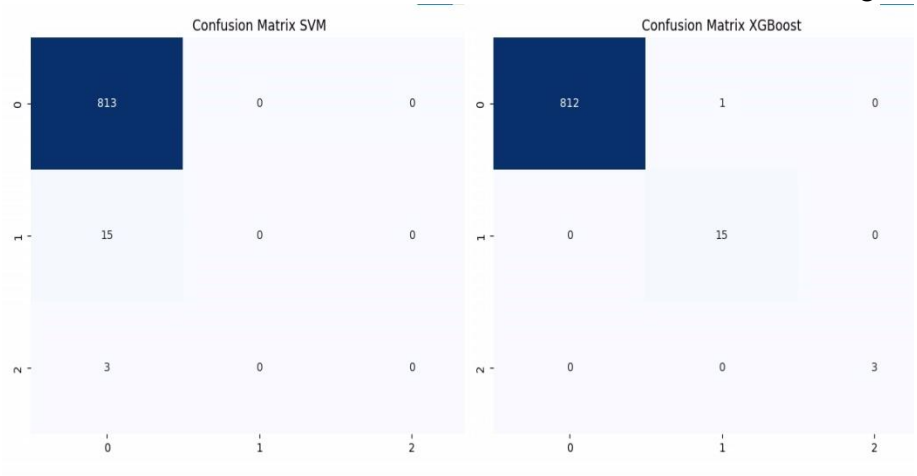

Figure 16. Confusion Matrix

Based on testing the stunting dataset at the Citapen Health Center using 2 algorithm models, namely SVM and XGBoost, the experimental research results can be summarized in table 2.

Table 2. Model Comparison

Ration	**70 : 30**					
Comparison	**SVM**	**RBF**	**Linear**	**Sigmoid**	**Poly**	**XGBoost**
Accuracy	**97.83%**	**97.83%**	**99.04%**	**97.83%**	**97.83%**	**99.88%**

From the results of the confusion matrix and accuracy XGBoost was chosen to be used to the deployment stage because it is superior to the SVM model on splitting data 70:30 has an accuracy of 99.88%.

3.6 Deployment

After the XGBoost model and data sharing with a scenario of 70% training data and 30% test data are selected as the best model, then a website application is made using the Streamlit framework with the Visual Studio Code text editor so that diabetes prediction modeling can be accessed more easily. In this application, there is a home page that refers to the initial page displayed in a web browser as shown in Figure 17.

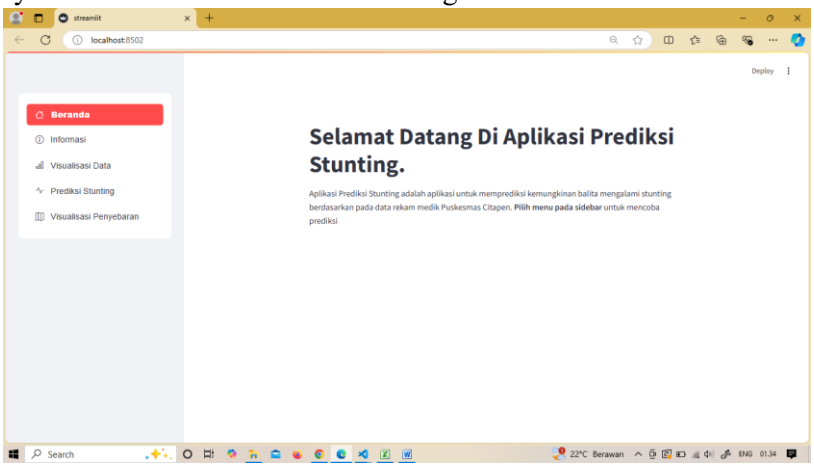

Figure 17. Homepage

On the home page there is information about the explanation of the website application that has been made and menu options located on the sidebar. On the page options there are 4 pages besides

the home page, namely the Information page, Data Visualization page, and Stunting Prediction page. The display of the information page is as shown in Figure 18.

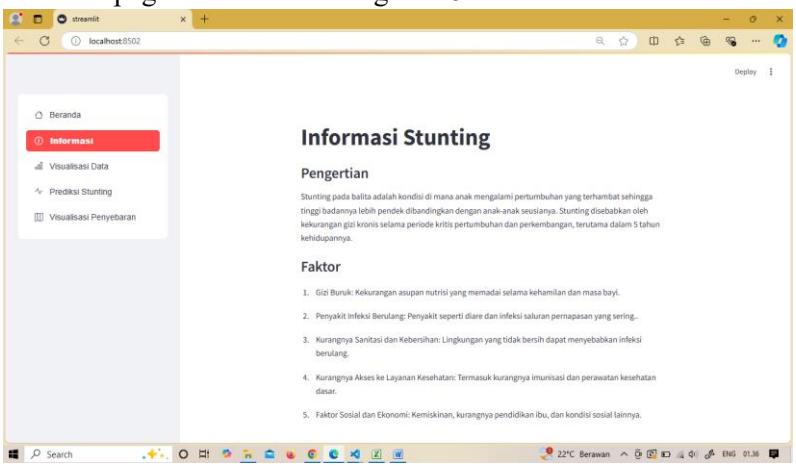

Figure 18. Information Page

The Information page explains what stunting is and what factors affect the likelihood of a toddler being stunted. The display of the Data Visualization page is as shown in Figure 19.

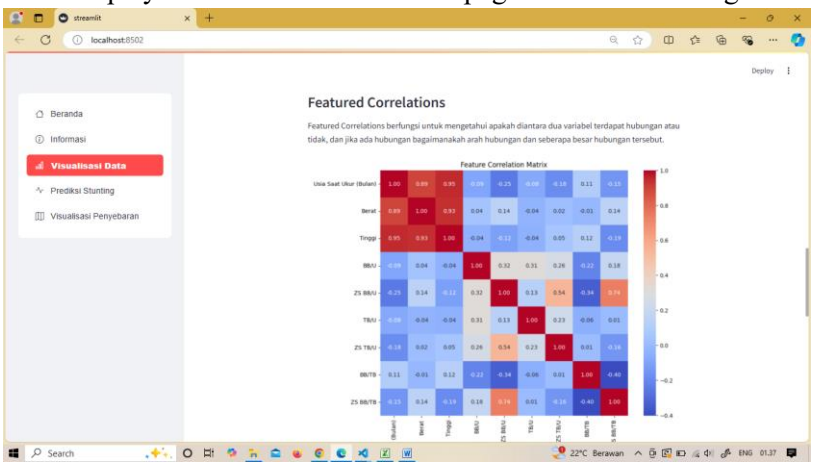

Figure 19. Data Visualization Page

On this page there is a data visualization of the previously created model, namely barchart visualization, correlations features, and categorical plot. Then there is the Stunting Prediction page as shown in Figure 20.

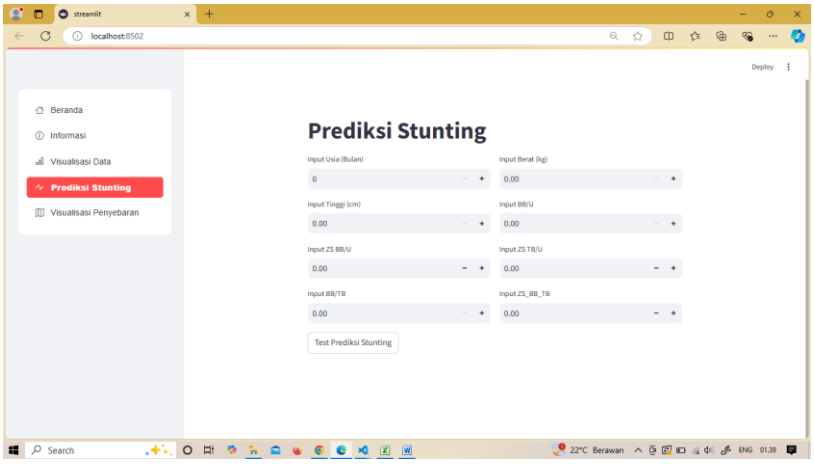

Figure 20. Stunting Prediction Page

It can be seen in Figure 20 that the Prediction page contains forms that can be filled in to predict stunting, if all forms have been filled in according to the data and click the prediction button, a diagnosis of severe stunting (Very Short), Stunting (Short), Normal, and High will appear. And the last page is the Distribution Visualization page which can be seen in Figure 21.

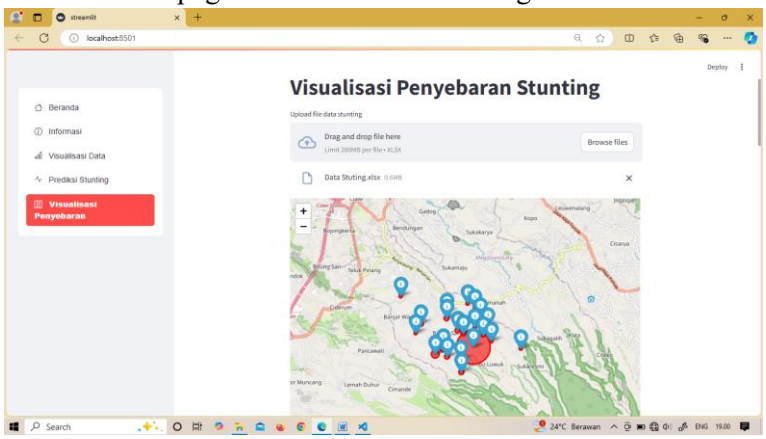

Figure 21. Stunting Distribution Page

The distribution visualization page is a page that displays the regional zone distribution of toddlers affected by stunting diagnosis. This visualization uses Latitude and Longitude data from each posyandu in the data. The display of the red-colored regional zone indicates that the regional zone is an area affected by stunting and the larger the circle the more stunted.

3.7 Testing

Tests are carried out to evaluate the effectiveness of the model on the website. In this study, trials were conducted to ensure that the model could produce the desired results on the website. The trial uses 5 random data from the dataset. Based on the test results, the model can predict the entire data correctly in accordance with the actual data. The test results can be seen in table 3.

Table 3. Trial

Usia	Berat	Tinggi	BB/U	ZS BB/U	ZS TB/U	BB/TB	ZS BB/TB	Actual	Prediction	Description
8	6.5	65.5	1	-2.68	-2.55	4	-1.59	1(short)	1(short)	Suitable
33	12.5	91.5	0	-0.93	-1.06	4	-0.56	0(Normal)	0(Normal)	Suitable
10	6.2	62	3	-3.57	-5.1	4	-0.62	2(very short)	2(very short)	Suitable
9	9.2	79	0	0.72	3.19	4	-0.82	3(tall)	3(tall)	Suitable
38	12.8	91.4	0	-1.16	-1.62	4	-0.40	0(Normal)	0(Normal)	Suitable

4. Conclusion

In Featured Correlations, it can be seen that the variables that are very influential are “Weight” and “Height”, which have a correlation result value of 0.89 and 0.95, while the poor correlation value is found in the “BB / TB” variable, which has a value of -0.40, which indicates that the variable does not really affect the diagnosis results.

Classification was performed using SVM kernel RBF, Linear, Sigmoid, and Polynomial algorithms and XGBoost. The first stage carried out is the data cleaning and transformation stage, after which the division of training data and test data is 70:30. The model in this study was evaluated with confusion matrix. So it is known that the SVM algorithm obtained the best results on the Linear kernel with an accuracy of 99.04%. In the XGBoost algorithm, the accuracy result is 99.88%. The comparison shows that the XGBoost algorithm produces a more accurate classification

value than the SVM algorithm which is then chosen to create a web application system using the Streamlit framework. Based on a series of processes that have been carried out in this research, it can be concluded that this research has succeeded in building the best model and web application for stunting prediction using the XGBoost algorithm.

In the Streamlit framework web application using the XGBoost method which is the best method compared to SVM to make predictions. All pages function as expected. The Stunting Prediction page has been tested with the value according to the data owned and for the distribution visualization page it works accordingly where the area affected by the red zone is the area affected by stunting.

References

[1] S. Lonang and D. Normawati, "Klasifikasi Status Stunting Pada Balita Menggunakan K-Nearest Neighbor Dengan Feature Selection Backward Elimination," vol. 6, pp. 49–56, 2022, doi: 10.30865/mib.v6i1.3312.

[2] M. R. Nugroho, R. N. Sasongko, and M. Kristiawan, "Faktor-faktor yang Mempengaruhi Kejadian Stunting pada Anak Usia Dini di Indonesia," vol. 5, no. 2, pp. 2269–2276, 2021, doi: 10.31004/obsesi.v5i2.1169.

[3] R. Fitri, N. Huljannah, and T. N. Rochmah, "Program Pencegahan Stunting Di Indonesia :," vol. 17, no. 3, pp. 281–292, 2022.

[4] F. Adzim, "Klasifikasi Status Stunting Balita Menggunakan Metode C4.5 Berbasis Web," UIN Suska Riau, 2023.

[5] E. Sormin and C. Siagian, "Pelatihan Pengukuran Antropometri dan Edukasi Gizi Seimbang sebagai Upaya Revitalisasi Posyandu dalam Rangka Menurunkan Angka Stunting di Kelurahan Cawang / Jakarta Timur," vol. 4, pp. 786–794, 2022.

[6] H. I. Islam, M. K. Mulyadien, and U. Enri, "Penerapan Algoritma C4.5 dalam Klasifikasi Status Gizi Balita," vol. 8, no. July, pp. 116–125, 2022.

[7] S. Lonang and D. Normawati, "Klasifikasi Status Stunting Pada Balita Menggunakan K-Nearest Neighbor Dengan Feature Selection Backward Elimination Klasifikasi Status Stunting Pada Balita Menggunakan K-Nearest Neighbor Dengan Feature Selection Backward Elimination," no. March, 2022, doi: 10.30865/mib.v5i1.2293.

[8] S. Elina, H. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," vol. 4, no. 1, pp. 21–26, 2022.

[9] G. A. Shafila, "Implementasi Metode Extreme Gradient Boosting (Xgboost) Untuk Klasifikasi Pada Data Bioinformatika (Studi Kasus : Penyakit Ebola , GSE 122692)," 2020.

[10] D. Darwis, E. Shintya Pratiwi, A. Ferico, and O. Pasaribu, "Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia," 2020.

[11] A. W. Septyanto and H. L. Hariyanto, "Perbandingan Teknik Klasifikasi Catatan Medis untuk Indeks Antropometri Status Gizi Balita," vol. 6, no. 1, pp. 229–235, 2024.

[12] Y. Akkem, B. Kumar Saroj, and V. Aruna, "Streamlit Application for Advanced Ensemble Learning Methods in Crop Recommendation Systems – A Review," 2023.

[13] I. C. R. Drajana and A. Bode, "Prediksi Status Penderita Stunting Pada Balita Provinsi Gorontalo Menggunakan K-Nearest Neighbor Berbasis Seleksi Fitur Chi Square," vol. 5, no. 2, pp. 309–316, 2022.

[14] F. Adzhima, E. Budianita, A. Nazir, and F. Syafria, "Klasifikasi Status Stunting Balita Dengan Metode Support Vector Machine Berbasis Web," pp. 381–392, 2023.

[15] D. K. D. Damayanti and M. Jakfar, "Klasifikasi Status Stunting Balita Menggunakan Algoritma Fuzzy C-Means (Studi Kasus Posyandu Rw 01 Kelurahan Jepara Surabaya)," vol. 11, no. 03, pp. 524–533, 2023.

[16] V. Herliansyah, R. Latuconsina, A. Dinimaharawati, and U. Telkom, “Prediksi Stunting Pada Balita Dengan Menggunakan Algoritma Klasifikasi Naïve Bayes Stunting Prediction In Children Using Naïve Bayes Classification,” vol. 8, no. 5, pp. 6642–6649, 2021.
[17] K. Rahmadhita, “Jurnal Ilmiah Kesehatan Sandi Husada Permasalahan Stunting dan Pencegahannya Pendahuluan,” vol. 11, no. 1, pp. 225–229, 2020, doi: 10.35816/jiskh.v10i2.253.
[18] A. Roihan, P. A. Sunarya, and A. S. Rafika, “Pemanfaatan Machine Learning dalam Berbagai Bidang : Review paper,” vol. 5, no. April, pp. 75–82, 2020.
[19] F. Syarifudin, “Klasifikasi Artikel-Artikel Jurnal Pustakaloka,” vol. 17, no. 1, pp. 20–37, 2022.
[20] V. Tumbelaka, J. I. Kindangen, and R. Joseph, “ISSN 2442-3262 Jurnal Perencanaan Wilayah dan Kota Alam – FMIPA tahun 1998 . Sedangkan Program Pasca Sarjana didirikan pada tahun 1985 . Tahun 2009 lewat surat persetujuan DIKTI No . 212 / D / 2009 tanggal 17 Februari 2009 Fakultas Kesehatan Masyarakat re,” vol. 6, no. 1, pp. 59–67, 2019.
[21] F. Abdusyukur, “KOMPUTA : Jurnal Ilmiah Komputer dan Informatika Penerapan Algoritma Support Vector Machine (SVM) Untuk Klasifikasi Pencemaran Nama Baik Komputa : Jurnal Ilmiah Komputer dan Informatika,” vol. 12, no. 1, 2023.
[22] I. M. Parapat and M. T. Furqon, “Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak,” vol. 2, no. 10, pp. 3163–3169, 2019.
[23] S. H. Pramita, “Evaluasi Kinerja Metode Support Vector Machine (SVM), Naive Bayes Dan Decision Tree Untuk Diagnosa Penyakit Jantung,” 2023.
[24] Y. Rombe, “Penggunaan Metode Xgboost Untuk Klasifikasi Status Obesitas Di Indonesia,” 2021.
[25] L. Qadrini, A. Seppewali, and A. Aina, “Decision Tree Dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial,” vol. 2, no. 7, 2021.
[26] A. M. Khalimi, “Analisis Klasifikasi Algoritma Naive Bayes,” Pengalaman Edukasi. [Online]. Available: https://www.pengalaman-edukasi.com/2020/01/analisis-klasifikasi-algoritma-naive.html
[27] D. A. Nasution, H. H. Khotimah, and N. Chamidah, “Perbandingan Normalisasi Data Untuk Klasifikasi Wine Menggunakan Algoritma K-NN,” vol. 4, no. 1, pp. 78–82, 2019.
[28] I. N. Tri, A. Putra, K. S. Kartini, Y. K. Suyitno, and I. M. Sugiarta, “Penerapan Library Tensorflow , Cvzone , dan Numpy pada Sistem Deteksi Bahasa Isyarat Secara Real Time,” vol. 2, 2023.
[29] Z. M. E. Darmawan and A. F. Dianta, “Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM Implementation of GridSearchCV Hyperparameter Optimization in Heart Attack Prediction System Using SVM,” vol. 13, no. 1, pp. 8–15, 2023.
[30] K. Hermanto, D. Salim, B. Wu, O. R. Salim, and R. B. Gunadi, “Penggunaan Python Untuk Menganalisis Pola Penyebaran Covid-19 Di Masa Pandemi,” vol. 3, pp. 62–75, 2023.
[31] Y. Galahartlambang and T. Khotiah, “Visualisasi Data Dari Dataset COVID-19 Menggunakan Pemrograman Python,” vol. 4902, no. x, 2021.
[32] Y. R. Silitonga, J. Arjuna, U. No, and K. Jeruk, “Sistem Pendeteksi Berita Hoax Di Media Sosial Dengan Teknik Data Mining Scikit Learn Pengumpulan Data Pada proses awal yang diperlukan adalah melakukan mining data pada media sosial Facebook dan Twitter untuk dijadikan data training dan data testing . Data training berisi berita yang,” vol. 4, pp. 173–179, 2019.
[33] P. B. Sianturi, Pengembangan Aplikasi Penerjemah Bahasa Lampung Berbasis Web Menggunakan Framework Streamlit. 2023.
[34] A. Lathifah, “Cross-Industry Standard Process For Data Mining (CRISP-DM) Untuk Menemukan Pola Asosiasi Pada Data Tracer Study Lulusan Perguruan Tinggi Keagamaan Islam (Studi Kasus : Uin Syarif Hidayatullah Jakarta) Cross-Industry Standard Process For Data Mining", (2023.)