The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 4, August 2024
https://doi.org/10.33022/ijcs.v13i4.4040

Implementation of Naïve Bayes Classifier And Support Vector Machine for Stunting Classification

Nilam Wahdiaz Azani[1], M. Afdal[2], Rice Novita[3], Mustakim[4]
12050320451@students.uin-suska.ac.id[1], m.afdal@uin-suska.ac.id[2], rice.novita@uin-suska.ac.id[3], mustakim@uin-suska.ac.id[4]
[1,2,3,4]Information System Departement, Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau

Article Information

Received : 18 May 2024
Revised : 3 Jun 2024
Accepted : 25 Jul 2024

Keywords

Classificaction, NBC, Stunting, SVM

Abstract

Stunting is a condition when a child's physical growth and development are stunted or delayed due to a lack of adequate nutritional intake over a long period of time, especially during the early years of life. Indonesia still has a stunting prevalence rate above the WHO standard, which is at 21.6%. 2020 UN statistics recorded more than 149 million (22%) toddlers worldwide were stunted, of which 6.3 million were early childhood or stunted toddlers were Indonesian toddlers. This study aims to create a classification model using Data Mining Algorithms NBC and SVM to analyze and describe the class of a total of 2018 toddler nutritional status data in Lima Puluh Kota Regency. The results of this study are expected to be an evaluation of whether the stunting prevention program implemented has been successful, and can be the basis for creating the next program.

A. Introduction

According to the World Health Organization (WHO), stunting is a condition when a child's physical growth and development are inhibited or delayed due to a lack of adequate nutritional intake over a long period of time, especially in the early growth period of life, namely in the first 1,000 days of life. Stunting is the nutritional status of toddlers based on the PB / U or TB / U index where in the anthropometric standard for assessing the nutritional status of children, the measurement results are at the threshold (Z-Score) <-2 SD to -3 SD (short / stunted) and <-3 SD (very short / severely stunted) [1]. Based on data obtained from the paudpedia.kemdikbud.go.id website page, UN 2020 statistics recorded more than 149 million (22%) toddlers worldwide experiencing stunting, of which 6.3 million are early childhood or stunting toddlers are Indonesian toddlers.

Indonesia still has a stunting prevalence rate above the WHO standard, which is at 21.6%. Cases of stunting in toddlers are almost evenly experienced in Indonesia, one of which is Lima Puluh Kota Regency with a stunting prevalence of 24.3% in 2022 (SSGI Ministry of Health Indonesia 2022). Of the 22 health centers in Lima Puluh Kota, there are 4 health centers that have experienced an increase in the percentage of stunting toddlers, namely the Koto Baru Health Center, Dangung - dangung Health Center, Mungka Health Center, and Sialang Health Center [2].

Stunting is a national issue, even now the government through the instruction of the President of the Republic of Indonesia through Presidential Regulation Number 72 of 2021 continues to strive so that in the future the problem of stunting in Indonesia can be resolved[3]. The government targets that by 2024 the stunting prevalence rate can drop to 14%[4]. The government focuses on 11 specific intervention programs to reduce stunting directed at 2 phases of growth. Namely, the phase of pregnant women or before giving birth and the postnatal phase, which is mainly in infants aged 0-24 months.

Data mining is a way of extracting patterns in the use of very large data[5][6]. Data mining can search and analyze a collection of data (database) so that an interesting pattern is found with the aim of extracting accurate and potential information and knowledge, and can be understood and useful for decision making[7]. Common techniques in data mining include description, prediction, estimation, classification, clustering and association[8]. Prediction and classification techniques are usually used to analyze data that can describe data classes or to predict future data.

Research conducted by Putri and friends, named Classification System Of Toddler Nutrition Status using Naïve Bayes Classifier Based on Z- Score Value and Anthropometry Index, utilized data from 225 toddlers. The testing system employed 55 data points for training and 175 data points for testing, achieving 100% accuracy. The outcome of the study was the development of a system capable of classifying nutritional status based on a combination of three anthropometric indices using the Naive Bayes Classifier[9]. Then Purnomo, applied the Naïve Bayes Classification Method used in classifying the stunting status of toddlers in Madiun. The test results indicate that the Naive Bayes method is effective for classifying stunting status in toddlers. The implemented Naive Bayes algorithm achieved an average performance of 58% accuracy, 68% precision, and

58% recall, as determined by the confusion matrix test using 30% of the data for testing and 70% for training [10].

Rahmi have done research on Classification Of Stunting In Children Under Five Years In Padang City Using Support Vector Machine, concluded that the test results using the Support Vector Machine (SVM) method had a high accuracy value. The optimal parameters for predicting the classification of stunting in children were determined to be a cost value of 10 and a gamma value of 5, using the Radial Basis Function (RBF) kernel. This classification achieved an accuracy of approximately 100%, indicating that all observations were correctly predicted [11]. In Byna and Anisa's research, the accuracy value of the SVM algorithm for predicting stunting incidence is 81.62% [12]. In a study entitled Performance Test Of Naive Bayes And Svm Methods On Classification Of Malnutrition Status In Children written by Anamisa and friends, shows that the SVM method with polynomial kernel has the highest accuracy value of 89.76%, more accurate when compared to the Naïve Bayes method, with an accuracy rate of 86.31% [13].

Based on previous research, the authors used the NBC and SVM algorithms in classifying stunting cases in Lima Puluh Kota Regency. It is hoped that this research will provide information about stunting cases in Lima Puluh Kota Regency with a machine learning approach so that it can be an illustration for the government in determining intervention steps to reduce and prevent stunting.

B. Research Method

This research begins by collecting data on the nutritional status of toddlers in each health center in Lima Puluh Kota Regency. After the data is collected, data preprocessing is carried out through several processes, such as cleaning and transformation. After the data is ready to be processed, then the NBC and SVM algorithms are modeled. After all the processes are complete, then a comparison between models is made to determine the performance of each algorithm and then an evaluation is carried out to determine the performance of the model.

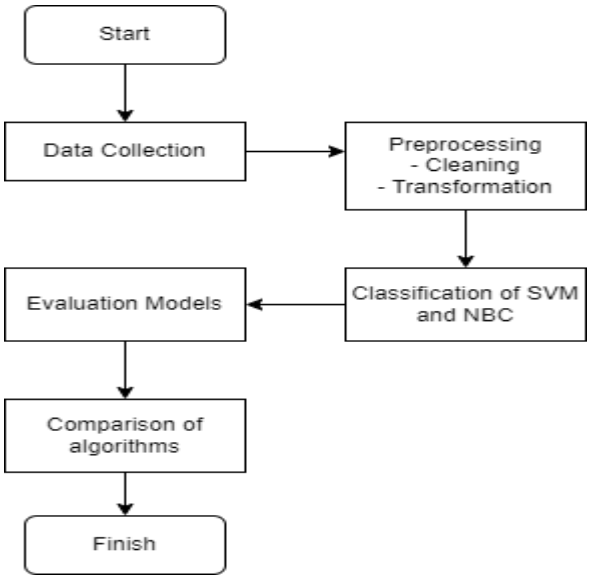

Figure1. Research Methodology

1. Data Collection
   The data collection stage is carried out to start the research. the data in this study uses data on the nutritional status of toddlers in Lima Puluh Kota Regency in 2023. For the initial data there are 2018 rows and 16 attributes.
2. Preprocessing
   In the preprocessing stage, data cleaning and transformation are performed. Cleaning is done to remove unused data by deleting incorrect data so that there is no more empty, irrelevant, and duplicated data. Transformation is done to normalize the data so that the data is more suitable for processing.
3. Classification
   The next step is modeling the NBC and SVM algorithms to classify the data

C. Result and Discussion

1. Initial Data
   The initial data obtained is 2018 rows with 16 attributes consisting of Gender, Date of Birth, Birth Weight, Birth Height, Health Center, Age at Measurement, Measurement Date, weight (w), height (h), weight per age (W/A), Z-Score weight for age (WAZ), height per age (H/A), Z-Score height per age (HAZ), W/H, Z-Score W/H (WHZ), and Stunting Status. This initial data will then be processed first by selecting attributes that are in accordance with stunting anthopometry according to the World Health Organization. After that, data preprocessing is carried out.

Tabel 1. Initial Data

No	Sex	*Health Center*	*Age*	*...*	*Weight (kg)*	*Height(cm)*
1	F	Mungo	4y 7m 13d	...	13,9	96
2	F	Koto Baru	2y 8m 9d	...	10	84
3	M	Baruah Gunung	4y 11m 16d	...	13	97,4
4	M	Mungo	3y 2m 15d	...	12	-
5	F	Pakan Rabaa	4y 10m 28d	...	13	-
...	...	...	...	...	...	...
2016	F	Baruah Gunung	0y 0m 0d	...	2.5	47
2017	M	Suliki	0y 0m 0d	...	2000	39
2018	M	Taram	0y 0m 0d	...	3,3	48

The table above is raw data obtained from Health District, Lima Puluh Kota Regency, which is a recap of the nutritional status of toddlers from 2023 and early 2024. In this dataset there are still many cells that are abnormally empty. besides that, some columns are also irrelevant to the calculation of stunting so normalization needs to be done.

2. Preprocessing Data
   Data preprocessing needs to be done to ensure that the data to be processed is ready for analysis. The stages of data preprocessing are as follows:

Cleaning

Cleaning is a stage to eliminate empty data, delete irrelevant attributes, and duplicate so that all data can be used.

Tabel 2. Data Cleansing

No	Sex	*Health Center*	*Age*	...	*Weight (kg)*	*Height(cm)*
1	F	Mungo	55	...	13,9	96
2	F	Koto Baru	32	...	10	84
3	M	Baruah Gunung	59	...	13	97,4
4	M	Mungo	38	...	12	87,5
5	F	Pakan Rabaa	58	...	13	90,2
...	...	...	...	...	...	...
2016	F	Baruah Gunung	0	...	2.5	47
2017	M	Suliki	0	...	2	39
2018	M	Taram	0	...	3.3	48

The table above is a dataset that has passed the cleaning stage. It can be seen that empty cells have been filled, and abnormal data has the same value. Date of Birth, Birth Weight, Birth Height, Measurement Date, weight per age (W/A), Z-Score weight for age (WAZ), height per age (H/A), Z-Score height per age (HAZ) have been removed.

Transformation

After the cleaning stage and there is no empty data, the next step is the data transformation stage. In the transformation stage, the nominal data type will be converted to numeric form.

Tabel 3. Data Transformation

No	Sex	*Health Center*	*Age*	...	*Weight (kg)*	*Height(cm)*
1	1	10	55	...	13,9	96
2	1	5	32	...	10	84
3	0	0	59	...	13	97,4
4	0	10	38	...	12	87,5
5	1	12	58	...	13	90,2
...	...	...	...	...	...	...
2016	1	1	0	...	2.5	47
2017	0	19	0	...	2	39
2018	0	21	0	...	3.3	48

At this stage the nominal data is converted into numeric. Such as gender, where 0 is male and 1 is female

3. Splitting Data

Before building the model, the data is separated into training data for model learning, and test data for evaluating the model. In this study, splitting data using the k-fold cross validation, with k=10.

4. Naive Bayes Classifier

Naïve Bayes Classifier was first proposed by Thomas Bayes, an English scientist. Naive Bayes is a classification technique that uses probabilistic and statistical approaches to estimate future probabilities by utilizing previous experience; hence, it is referred to as Bayes' Theorem. This theorem deals with the relationship between features that are considered unrelated. Naive Bayes classifiers make the assumption that the presence or absence of a particular attribute in a category is completely unrelated to the attributes of other categories [14].

The Bayes theorem process is generally shown in the following equation [15] :

$$P(C|X) = (P(X|C)P(C))/(P(X)) \quad (1)$$

Description:

X	: Unknown data class
C	: Hypothesis Data
P(C|X)	: Probability of hypothesis H based on condition X
P(C)	: Probability of hypothesis H
P(X|C)	: Probability of X based on condition in hypothesis H
P(X)	: Probability of X

Through experiments using Kfold Cross Validation, the highest accuracy for training model NBC are at K-10 with an accuracy value of 81.09% as shown in Figure 1.

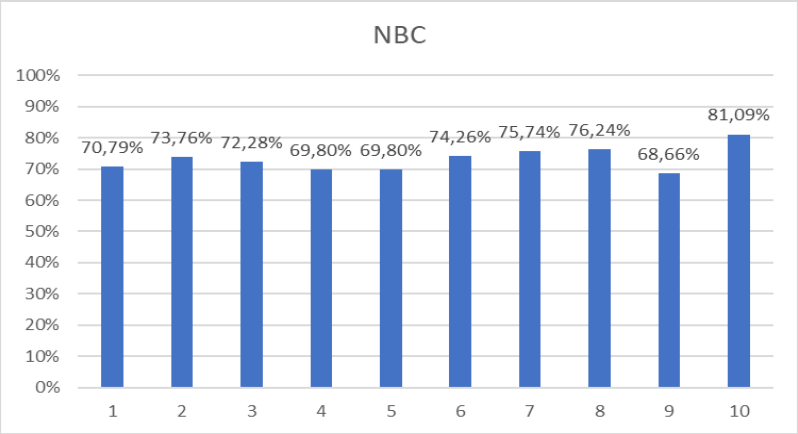

Figure 1. Result of K-fold Cross Validation NBC

After training the dataset, the next step is evaluation to measure the performance of the NBC algorithm model. The model evaluation results can be seen in table 4.

Table 4. Confusion Matrix NBC

Stunting	**True Yes**	***True No***
Prediction Yes	147	11

Prediction No	27	16

The confusion matrix above shows that the NBC model that has been made successfully classifies 147 stunting data and 16 non-stunting data. From the classification results, the accuracy value is 81%, precision 84%, and recall 93%.

5. Support Vector Machine

Support Vector Machine is a learning system that uses linear function theory on features that have been trained by utilizing optimal theory[16]. SVM is a supervised learning algorithm with the basic principle of using kernel technology, first introduced by Vapnik, et al. in 1955[17]. The SVM algorithm has the ability to learn well with only a very small number of free parameters, resilient to several types of outliers, and computationally efficient compared to several other methods[18].

Through experiments using Kfold Cross Validation, the highest accuracy for training model on SVM are at K-10 with an accuracy value 90,05% as shown in Figure 2.

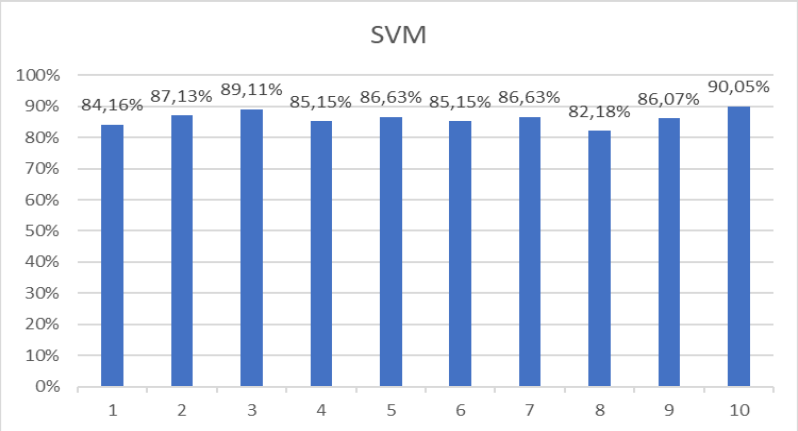

Figure 2. Result of K-fold Cross Validation SVM

The next step is evaluation to measure the performance of the SVM algorithm model. The model evaluation results can be seen in table below.

Table 5. Confusion Matrix SVM

Stunting	**True Yes**	***True No***
Prediction Yes	160	6
Prediction No	14	21

As we can see that the confusion matrix above shows the SVM model has been made successfully classifies 160 stunting data and 21 non-stunting data. From the classification results, the accuracy value is 90%, precision 92%, and recall 96%.

D. Conclusion

From the results of testing the NBC and SVM algorithms, the accuracy of each algorithm is 81% and 90%, which means that the SVM algorithm is better at predicting stunting classification.

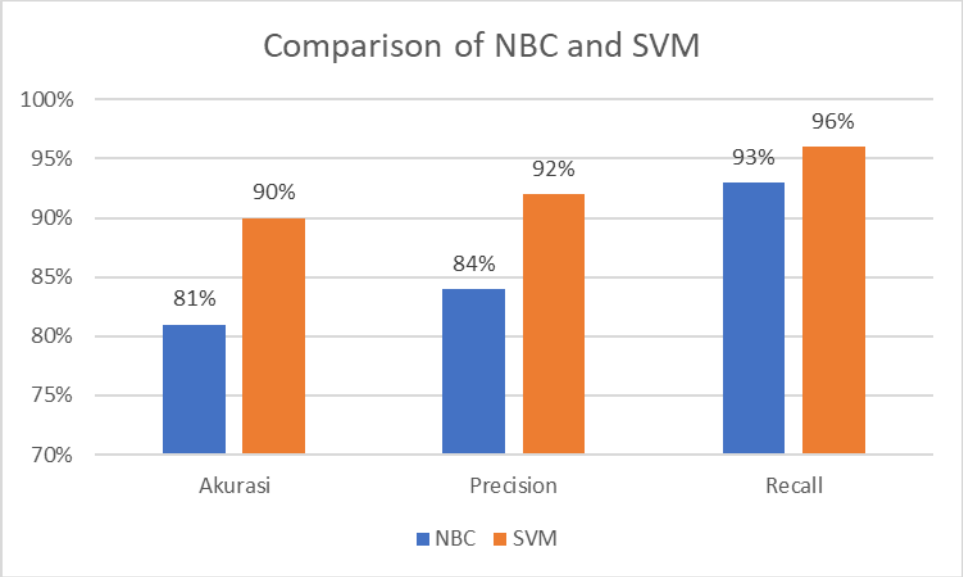

Figure 3. NBC and SVM algorithm modeling results

According to the study, it is concluded that data mining techniques with SVM and NBC algorithms can be applied to classify stunting cases. The SVM algorithm is better at classifying large datasets, compared to the NBC algorithm. when making predictions with new data, both algorithms can still classify well.

E. Acknowledgment

Thanks to the health department of Lima Puluh Kota Regency for giving permission and access to process the data needed in the study.

F. References

[1] J. L. Leroy and E. A. Frongillo, "Perspective: What Does Stunting Really Mean? A Critical Review of the Evidence," *Adv. Nutr.*, vol. 10, no. 2, pp. 196–204, 2019, doi: 10.1093/advances/nmy101.

[2] Dinas Kesehatan Lima Puluh Kota, "Laporan Data Stunting Di Kabupaten Lima Puluh Kota Tahun 2019 dan 2020," pp. 1–5, 2020.

[3] H. S. Salmon and D. K. Moninjta, "Jurnal governance," vol. 1, no. 2, pp. 1–14, 2022.

[4] N. Annisa, A. Nurdin, A. Tihardimanto, U. Rimayanti, and A. Ahmad, "Implementation of Stunting Prevention Program in Indonesia: Literature Review," *Citizen-Based Mar. Debris Collect. Train. Study case Pangandaran*, vol. 7, no. 4, pp. 56–61, 2024.

[5] M. Afdal and R. P. Disastra, "Analisis Pola Kecelakaan Lalu Lintas Menggunakan Algoritma K-Means Dan Fp-Growth Studi Kasus: Polres Solok," *J. Ilm. Rekayasa dan Manaj. Sist. Inf.*, vol. 8, no. 1, p. 31, 2022, doi: 10.24014/rmsi.v8i1.15656.

[6] S. Ndagijimana, I. H. Kabano, E. Masabo, and J. M. Ntaganda, "Prediction of Stunting Among Under-5 Children in Rwanda Using Machine Learning Techniques.," *J. Prev. Med. Public Health*, vol. 56, no. 1, pp. 41–49, Jan. 2023, doi: 10.3961/jpmph.22.388.

[7] Mambang, F. D. Marleny, and M. Zulfadhilah, "Prediction of linear model on stunting prevalence with machine learning approach," *Bull. Electr. Eng. Informatics*, vol. 12, no. 1, pp. 483–492, 2023, doi: 10.11591/eei.v12i1.4028.

[8] S. M. Birjandi and S. H. Khasteh, "A survey on data mining techniques used in medicine," *J. diabetes Metab. Disord.*, vol. 20, no. 2, pp. 2055–2071, 2021.

[9] T. E. Putri, R. T. Subagio, Kusnadi, and P. Sobiki, "Classification System of Toddler Nutrition Status using Naïve Bayes Classifier Based on Z- Score Value and Anthropometry Index," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012005.

[10] A. Rozaq and A. J. Purnomo, "Classification of Stunting Status in Toddlers Using Naive Bayes Method in the City of Madiun Based on Website," *J. Techno Nusa Mandiri*, vol. 19, no. 2, pp. 69–76, 2022, doi: 10.33480/techno.v19i2.3337.

[11] I. Rahmi, M. Susanti, H. Yozza, and F. Wulandari, "Classification of Stunting in Children Under Five Years in Padang City Using Support Vector Machine," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 16, no. 3, pp. 771–778, 2022, doi: 10.30598/barekengvol16iss3pp771-778.

[12] A. Byna, "Comparative Analysis of Machine Learning Algorithms for classification about Stunting Genesis," 2020, doi: 10.4108/eai.23-11-2019.2298349.

[13] D. R. Anamisa, A. Jauhari, and F. A. Mufarroha, "Performance Test of Naive Bayes and Svm Methods on Classification of Malnutrition Status in Children," *Commun. Math. Biol. Neurosci.*, vol. 2024, pp. 1–18, 2024, doi: 10.28919/cmbn/8429.

[14] S. Stephen, "Implementasi Metode Algoritma Naive Bayes Untuk Memprediksi Customer Churn Pada Bisnis Usaha Mikro Kecil Dan Menengah," vol. 4, no. 1, pp. 86–96, 2022, [Online]. Available: http://repositori.buddhidharma.ac.id/1613/%0Ahttp://repositori.buddhidharma.ac.id/1613/1/COVER – BAB III.pdf

[15] D. P. Santoso and W. Wibowo, "Analisis Sentimen Ulasan Aplikasi Buzzbreak Menggunakan Metode Naïve Bayes Classifier pada Situs Google Play Store," *J. Sains dan Seni ITS*, vol. 11, no. 2, 2022, doi: 10.12962/j23373520.v11i2.72534.

[16] R. N. Yusra, O. S. Sitompul, and Sawaludin, "Kombinasi K-Nearest Neighbor (KNN) dan Relief-F Untuk Meningkatkan Akurasi Pada Klasifikasi Data," vol. 1, pp. 0–5, 2021.

[17] Y. Yan, Y. Wang, and Y. Lei, "Micro Learning Support Vector Machine for Pattern Classification : A High-Speed Algorithm," vol. 2022, 2022.

[18] I. Steinwart and A. Christmann, *Support Vector Machines (1st edition)*. Springer Publishing Company, 2008.