Classification of Toddler Nutritional Status Based on Antrophometric Index and Feature Discrimination using Support Vector Machine Hyperparameter Tuning

Much. Afif Masykur Mughni
Department of Informatics Engineering
Faculty of Computer Science
Universitas Narotama
Surabaya, Indonesia
masykurafif@gmail.com

Tresna Maulana Fahrudin
Department of Data Science
Faculty of Computer Science
Universitas Pembangunan Nasional
"Veteran" Jawa Timur
Surabaya, Indonesia
tresna.maulana.ds@upnjatim.ac.id

Made Kamisutara
Department of Informatics Engineer
Faculty of Computer Science
Universitas Narotama
Surabaya, Indonesia
made.kamisutara@narotama.ac.id

***Abstract*— The study of food and its relationship to health is known as nutritional status. Nutritional status is a criterion for determining a toddler's overall health. Body weight to age (BW / A), body height to age (BH / A), and body weight to body height (BW / BH) are three indexes used to determine a toddler's nutritional condition. Nutrition is still a crucial component in the development and growth of toddlers in Indonesia. In Indonesia, public health services like as the Public Health Center (Puskesmas) and the Integrated Healthcare Center (Posyandu) collaborate to manage toddler growth and development. We proposed a research to classify the nutritional status of toddler based on anthropometric index, to help control the growth and development of toddler. SVM Hyperparameter Tuning was used to turn the nutritional status of toddlers into a classification model. SVM is a machine learning algorithm that uses a hypothesis space in the form of linear functions in a high-dimensional feature space to classify data. To get a model that can solve machine learning issues optimally, the hyperparameter has to be adjusted. As a preprocessing stage, we used Fisher's Discriminant Ratio to choose important features, such as body weight (BB) and body height (BH). The classification model using SVM on training and testing data with a 70:30 ratio achieved 84% accuracy, while SVM Hyperparameter Tuning with parameters of Cost = 100 parameters, Gamma = 0.01, Kernel = RBF achieved 97% accuracy. They represented a significant accuracy difference of 13%.**

Keywords—Toddler; Nutritional Status; Fisher's Discriminant Ratio; Hyperparameter Tuning; Support Vector Machine

I. INTRODUCTION

Nutrition is a science that studies food and is related to health. The science of nutrition which discuss the properties of nutrients contained in food, nutritional status is a benchmark for assessing the health development of toddlers. Nutritional status helps to fulfill the nutritional status of toddler[1]. The issue of nutrition is still one of the main factors affecting the growth and development of toddler in Indonesia. The nutritional status of toddler is assessed according to three index, such as body weight to age (BW / A), body height to age (BH / A), and body weight to body height (BW / BH)[2].

Along with the development of technology, the assessment of nutritional status that initially used conventional methods can be carried out automatically. Data mining, commonly called Knowledge Discovery in Database (KDD), is often used for future decision making that is obtained from past data and information. For example, for classification, clustering, prediction, association, estimation, description [3]. Before applying classification, feature selection is performed, which is a technique for determining the most important features from a full collection of features while keeping as much of their class discriminating information as feasible [4].

Support Vector Machine (SVM) is one of the best methods of solving classification problems [5]. Support Vector Machine is machine learning that classification uses a hypothetical space in the form of linear functions in a high-dimensional feature space. In the concept of SVM, how to find the best separator (hyperlink) function among the unlimited number of functions[6]. SVM also has several hyperparameters such as the value of *c* or *gamma* used. In

general, the values selected for the hyperparameters affect the predictive performance of the model [7]. Hyperparameter searches are generally done manually, via a practically rule [8]. Hyperparameter is a parameter that value is used to control the learning process. Grid search is generally used as an approach to hyperparameter tuning that will methodically construct and evaluate a model for any combination of algorithm parameters defined in a grid [9].

II. FUNDAMENTAL THEORY

A. Data Normalization

Normalization is the scaling steps to change the attribute values to make it more balanced. Z-Score is one of the best normalization methods, Z-score formula is shown by equation (1). The importance of normalizing data to build an accurate prediction model [10].

$$\text{new data} = \frac{\text{data}-\text{mean}}{\text{standard deviation}} \quad (1)$$

B. Feature Selection

Feature selection is a technique for selecting features. The selection step is critical because if one feature is chosen with low discriminating power, the following classifier model will perform poorly [4]. The goal is to be able to choose the best features, eliminate noise data and irrelevant data.

Fisher's Discriminant Ratio as a feature selection method based on attribute ranking is commonly used to measure the discrimination of a feature that is separated into two classes based on the class it belongs to. The formula for Fisher's Discriminant Ratio (FDR) is shown in equation (2).

$$\text{FDR} = \frac{(m_1-m_2)^2}{(s_1+s_2)} \quad (2)$$

Where $m_1$dan m_2 are the average values of the two classes, respectively, while $s_1$and s_2 are the variants of the two classes in the features to be measured.

C. SVM Hyperparameter Tuning

Support Vector Machine is a machine learning classification using a hypothetical space in the form of a linear function in a high-dimensional feature space [6]. SVM also has several hyperparameters such as the value of *c* or *gamma* used. In general, the values selected for the hyperparameters affect the predictive performance of the model [7]. The SVM formula is shown in the equation (3).

$$\min \frac{1}{2}|w|^2 = \frac{1}{2}(w_1^2 + w_2^2) \quad (3)$$

with the provision of $y_i(x_i w + b) - 1 \geq 0$

where $(x_i w + b) \geq 1$ for class 1, and $(x_i w + b) - 1 \geq 0$ for class 2.

Hyperparameter is a parameter that value is used to control the learning process. Hyperparameters must be set so that the model can solve machine learning problems optimally [11]. Hyperparameter Tuning in SVM usually using parameters such as Cost (C) and Gamma (γ).

D. Confusion Matrix

Confusion Matrix is a method for measuring classification performance. There is an information in the confusion matrix that can compare the predict label with the actual label. Confusion matrix is used to find accuracy, precision, and recall of prediction models that have been formed [12]. Figure 1 shows the square table between actual labels and predictive labels grouped into 4 categories such as True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

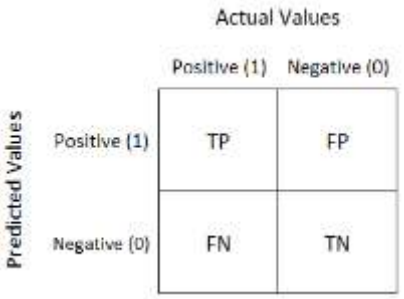

Figure 1. TP, TN, FP dan FN

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \times 100\% \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{\text{FP}+\text{TP}} \times 100\% \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{FN}+\text{TP}} \times 100\% \quad (6)$$

The term accuracy refers to a model's ability to accurately classify data. The ratio of correct positive and negative predictions of the whole data is called accuracy. To put it another way, accuracy refers to the degree to which the predicted value is near to the actual value. Equation (4) shows the accuracy formula.

Precision refers to the degree of accuracy between the desired data and the model's forecast outputs. The ratio of true positive predictions to total positive predicted results is called precision. All the positive classes that have been predicted correctly, how many data are really positive. The precision formula is shown in the equation (5).

The success of the model in retrieving information is referred to as recall. The ratio of true positive predictions to total true positive data is known as recall. In the equation (6), the recall formula is shown.

III. RESEARCH METHODOLOGY

Figure 2 shows the proposed system design of the research which consist of 5 steps.

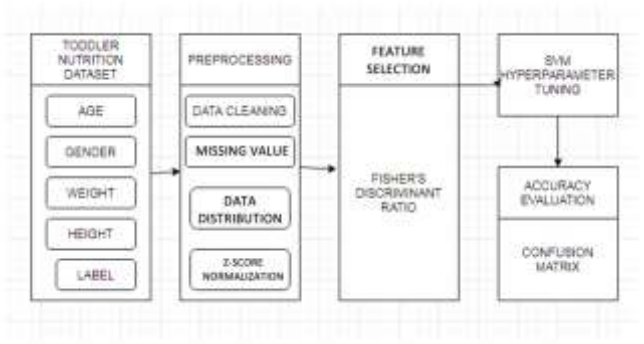

Figure 2. The Proposed System Design of Research

The first stage of this research is the collection of a dataset obtained from the anthropometric standards of toddler published by the Ministry of Health [13] with a total of 869 samples of 5 attributes. Table I shows the toddler nutritional dataset.

TABLE I. Toddler Anthropometric Dataset

No.	Atribute	Description
1.	Age	0-60 months
2.	Gender	1=Male 2=Female
3.	Body Weight	in kilogram
4.	Body Height	in centimeter
5.	Nutritional Status Label	-3, -2, -1, 0, +1, +2, +3
Label: The more it looks to the left (-): Skinny Toddler The more it looks to the right (+): Fat Toddler		

The next stage is preprocessing which includes data cleaning and data normalization. Data cleaning is removing duplicate data, checking for inconsistent data, and checking for typos. Data cleaning also affect the performance of data mining techniques because it reduces the number and complexity[14]. In addition, normalization to ensure the data distribution is normal distribution.

Feature selection using Fisher's Discriminant Ratio (FDR) to find out which of the attributes of age, sex, body weight, and body height are the most discriminator features in the dataset. After selected the attributes that have a high level of discrimination, then building a classification model that can predict the membership of a class from the individual testing data. The class label is classified as under nutrition, good nutrition, risk of overnutrition, and overnutrition that have been previously assumed. The classification method used Support Vector Machine (SVM) with optimization of hyperparameter tuning. Hyperparameters are proposed to the model can solve machine learning problems optimally. Grid search is performed to minimize over and under fitting and to obtain parameter combinations that produce the best accuracy in classification modeling [15].

The last stage is the evaluation model of the classification performances using confusion matrix to get the accuracy of model that has been trained in predicting class labels. The model will be tested by testing data to know the performance related to the actual labels matches the predictive labels.

IV. Experimental Result

A. *Preprocessing*

The preprocessing data includes data cleaning and normalization. After the dataset was cleaned, exploratory data analysis can be implemented to visualize the condition of dataset, for the example, data distribution, histogram, and percentage using pie diagram.

Data distribution of age attribute can be seen in Figure 3, which the age attribute of toddler in the range 0 to 60 months.

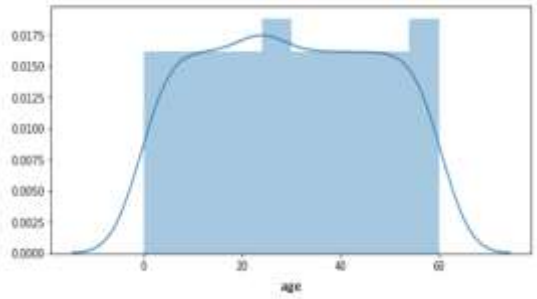

Figure 3. Data Distribution of Age Attribute

Furthermore, the percentage of gender attributes categorized at the discrete value 1 = male by 50%, while 2 = female as much as 50%. Figure 4 shows a pie chart of the percentage of nutritional status data based on gender attribute.

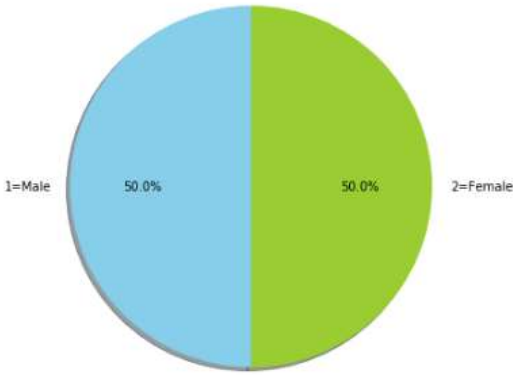

Figure 4. Percentage of Toddler Nutritional Status based on Gender Attribute

Data distribution of body height and weight attributes can be seen in Figure 5 and 6 which both are included in the normal distribution.

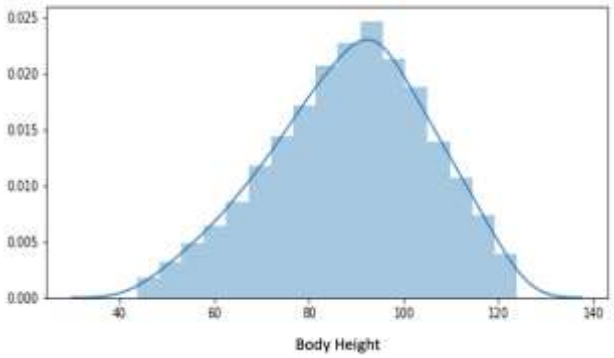

Figure 5. Data Distribution of Height Attribute

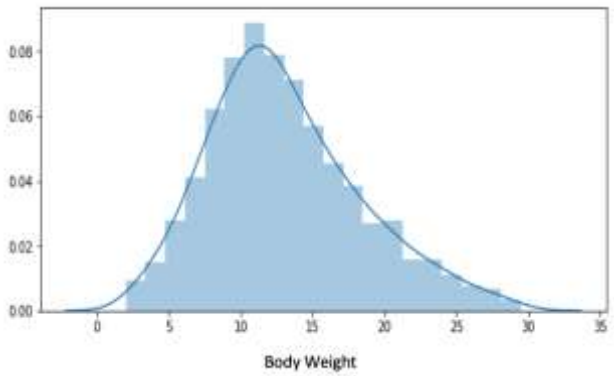

Figure 6. Data Distribution of Weight Attribute

The percentage of the nutritional status label is shown in Figure 7 which shows the label -3 = less nutrition, -2 = good nutrition, -1 = good nutrition, 0 = good nutrition, 1 = good nutrition, 2 = risk of over nutrition, 3 = over nutrition based on the weight index according to body length or height (BW / BL or BW / BH).

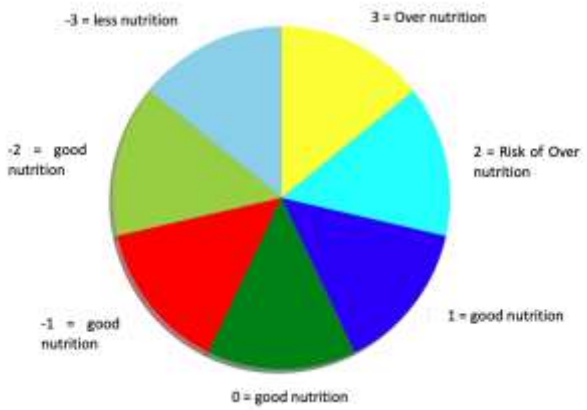

Figure 7. Percentage of Toddler Nutritional Status based on Class Label

B. Normalization using Z-score

The Z-score normalization result by calculating the mean and standard deviation of each attribute. Figure 8 shows the normalized of toddler nutritional dataset, which the dataset has been quite balanced or in the same range of each attribute.

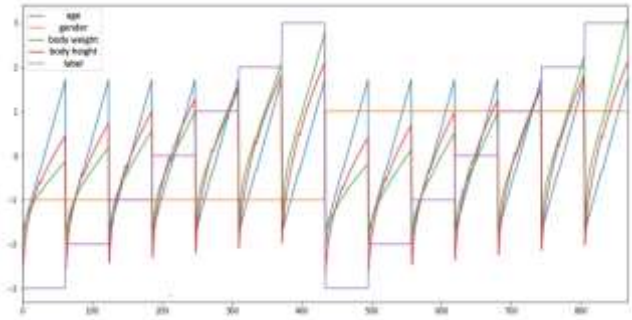

Figure 8. Z-score Normalization of Toddler Nutritional Status Dataset

C. Feature Selection using Fisher's Discriminant Ratio (FDR)

To get attributes that have a high discrimination score, FDR is a alternative of feature selection method based on attribute ranking. Figure 9 shows the attribute ranking of several important attributes and have the high discrimination score evaluated by FDR, such as the body weight (BB) has a score of 0.8 and the body height (BH) has a score of 0.3, the both attributes are the most important as a predictor variable to determine the toddler nutritional status.

Figure 9. Attribute Ranking of Nutritional Status of Toddler using FDR

D. Classification Model of Toddler Nutrition Status using SVM Hyperparameter Tuning

There were several scenarios in the research, such as comparing the accuracy performance result of SVM model between using the full-set attribute and the selected attribute along with their hyperparameter tuning. The accuracy of the toddler nutritional status classification model using SVM involving all attributes is shown in Table II.

TABLE II. Accuracy Performance of SVM Model using Full-set Attribute

No.	Ratio of training and testing data	Accuracy
1.	90:10	85%
2.	80:20	84%
3.	70:30	84%
4.	60:40	83%
5.	50:50	84%

TABLE III. Accuracy Performance of SVM Model using Selected Attribute

No.	Ratio of training and testing data	Accuracy
1.	90:10	64%
2.	80:20	66%
3.	70:30	67%
4.	60:40	61%
5.	50:50	58%

While the accuracy of the classification model using SVM involving only 2 attributes, such as Body Weight (BW) and Body Height (BH) based on the selected attributes using FDR is shown in Table III. There is a difference in accuracy in Tables II and III, which SVM does not given good performance when a number of attributes in the dataset must be selected by the FDR as shown in Table III with an accuracy range of 58-67%. SVM given good performance using a full-set attribute with an accuracy range of 83-85% as shown in Table II. Therefore, full-set attributes are still used at a later stage to build a model using SVM Hyperparameter Tuning.

After searching for parameters using the Grid Search with the ratio of training data and testing data, such as 50:50, 60:40, 70:30, 80:20, and 90:10, the best values were found for the parameters Cost = 100, Gamma = 0.01, and Kernel. = RBF. The ratio of training data and testing data of 60:40, the best values were found from the parameters Cost = 0.1, Gamma = 1, and Kernel = Polynominal. Then the classification modeling was using SVM Hyperparameter Tuning by involving all attributes which reached the significant accuracy above 95% which can be seen in Table IV.

TABEL IV. Accuracy Performance of SVM Hyperparameter Tuning Model

No.	Ratio of training and testing data	Accuracy
1.	90:10	96%
2.	80:20	97%
3.	70:30	97%
4.	60:40	96%
5.	50:50	95%

E. Confusion Matrix and Evaluation

In Table IV, it can be seen that the ratio of the best training and testing data is 70:30, which the accuracy reached 97%. Table V shows the results of the confusion matrix that evaluates the accuracy of each class label based on the precision, recall and F-measure using SVM Hyperparameter Tuning model, with a ratio of training data and testing data of 70:30. SVM Hyperparameter Tuning given the best performance in classifying the toddler nutritional status dataset.

TABEL V. The Performance of SVM Hyperparameter Tuning Model

Label	Precision	Recall	F1-Score
-3	100%	100%	100%
-2	97%	100%	98%
-1	100%	96%	98%
0	92%	100%	96%
1	100%	94%	97%
2	100%	94%	97%

V. CONCLUSION

We concluded that after the feature selection using Fisher's Discriminant Ratio (FDR) was applied, the features that have a high discrimination score were body weight (BW) and body height (BH). However, feature selection using FDR does not promise high accuracy, because the attribute of dataset condition is leaving 2 out of 4 attributes. Therefore, all attributes are retained in building the classification model. The experiment result showed SVM given quite good accuracy with the ratio of training and testing is 70:30, which the accuracy reached 84%. While SVM Hyperparameter Tuning with the parameters Cost = 100, Gamma = 0.01, and Kernel = RBF given the best accuracy reached 97%. It is a significant difference in accuracy of 13% between models using standard SVM compared with using Hyperparameter Tuning. The experiment proved that SVM with Hyperparameter Tuning improved the accuracy performance.

VI. REFERENCE

[1] F. R. Ambarwati, Gizi dan Kesehatan Reproduksi. Cakrawala Ilmu, 2015.

[2] Kementerian Kesehatan RI, “Buku saku pemantauan status gizi,” Buku saku pemantauan status gizi tahun 2017, pp. 7–11, 2018.

[3] C. S. Nur Aeni Widiastuti, Stefanus Santosa, “Algoritma Klasifikasi Data Mining Naive Bayes Berbasis Particle Swarm Optimization Untuk Deteksi Penyakit Jantung,” 2014.

[4] T. H. Lin, H. Te Li, and K. C. Tsai, “Implementing the Fisher’s Discriminant Ratio in a k-Means

Clustering Algorithm for Feature Selection and Data Set Trimming," J. Chem. Inf. Comput. Sci., vol. 44, no. 1, pp. 76–87, 2004, doi: 10.1021/ci030295a.
[5] A. Pratama, R. C. Wihandika, and D. E. Ratnawati, "Implementasi Algoritme Support Vector Machine (SVM) untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. March, pp. 1704–1708, 2018.
[6] D. Widiastuti, J. S. Informasi, and U. Gunadarma, "Analisa Perbandingan Algoritma Svm , Naive Bayes , Dan Decision Tree Dalam Mengklasifikasikan Serangan (Attacks)," pp. 1–8, 2007.
[7] R. G. Mantovani, A. L. D. Rossi, J. Vanschoren, B. Bischl, and A. C. P. L. F. De Carvalho, "Effectiveness of Random Search in SVM hyper-parameter tuning," Proc. Int. Jt. Conf. Neural Networks, vol. 2015-Septe, 2015, doi: 10.1109/IJCNN.2015.7280664.
[8] M. Claesen and B. De Moor, "Hyperparameter Search in Machine Learning," pp. 10–14, 2015, [Online]. Available: http://arxiv.org/abs/1502.02127.
[9] M. I. Chacon-murguia, L. E. Guerra-fernandez, and H. Erives, "Hyper-Parameter Tuning for Support Vector Machines by Estimation of Distribution Algorithms," vol. 667, no. December, pp. 175–185, 2017, doi: 10.1007/978-3-319-47054-2.
[10] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," Appl. Soft Comput. J., no. xxxx, p. 105524, 2019, doi: 10.1016/j.asoc.2019.105524.
[11] Z. A. Sunkad and Soujanya, "Feature Selection and Hyperparameter Optimization of SVM for Human Activity Recognition," Proc. - 2016 3rd Int. Conf. Soft Comput. Mach. Intell. ISCMI 2016, pp. 104–109, 2017, doi: 10.1109/ISCMI.2016.30.
[12] Eko Prasetyo, DATA MINING KONSEP DAN APLIKASI MENGGUNAKAN MATLAB. CV.ANDI OFFSET(Penerbit Andi), 2012.
[13] Kemenkes, "Peraturan Menteri Kesehatan RI tentang Standar Antropometri Penilaian Status Gizi Anak," Standar Antropometri Penilaian Status Gizi Anak. p. 78, 2020.
[14] Y. Asriningtias et al., "Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa," vol. 8, no. 1, pp. 837–848, 2014, doi: 10.12928/jifo.v8i1.a2082.
[15] F. Budiman, "SVM-RBF parameters testing optimization using cross validation and grid search to improve multiclass classification," Sci. Vis., vol. 11, no. 1, pp. 80–90, 2019, doi: 10.26583/sv.11.1.07.