Stunting Classification Analysis for Toddlers in Bojongsoang: A Data-Driven Approach

Muhammad Ghiyaats Daffa[1], **Putu Harry Gunawan**[2]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]Divisi Digital Service PT Telekomunikasi Indonesia
[1]ghiyats@students.telkomuniversity.ac.id, [2]phgunawan@telkomuniversity.ac.id

Abstract

Stunting merupakan salah satu masalah kesehatan yang menjadi prioritas masalah kesehatan anak di Indonesia. Pencegahan stunting pada balita sangat diperlukan untuk menghindari dampak jangka panjang bagi balita dan masyarakat. Pencegahan stunting dapat dilakukan dengan memantau pertumbuhan balita. Oleh karena itu, dibutuhkan sebuah sistem yang dapat memprediksi kondisi stunting pada balita. Machine learning menawarkan banyak metode yang dapat digunakan untuk membangun sistem prediksi kondisi stunting pada balita. Penelitian ini menganalisis beberapa model machine learning yang berpotensi cocok untuk memprediksi kelas stunting, yaitu K-Nearest Neighbor (KNN), Random Forest (RF), dan Ensemble Learning yang disebut Boosted KNN (BK). Dataset yang digunakan dalam penelitian ini memiliki masalah ketidakseimbangan, dimana data stunting hanya sebesar 1% dari total dataset. Oleh karena itu, dilakukan oversampling pada dataset dengan cara membangkitkan dataset secara acak berdasarkan distribusi data yang tergolong dalam kelas minoritas. Hasil dari penjabaran oversampling ini terbukti memuaskan. Menerapkan data yang tidak seimbang memberikan rata-rata akurasi sebesar 98% untuk semua metode yang digunakan; namun, rata-rata makro skor F-1 terbukti tidak optimal untuk masing-masing metode, dengan 51,95% untuk KNN, 52,45% untuk RF, dan 53,55% untuk BK. Setelah data diseimbangkan dengan melakukan oversampling, rata-rata makro skor F-1 untuk semua metode meningkat secara substansial. Hasil yang baru adalah 93,55% untuk KNN, 97,70% untuk RF, dan 98,00% untuk BK, menggarisbawahi peran penting dalam mengatasi ketidakseimbangan data dalam meningkatkan akurasi prediksi.

Abstract

Stunting is one of the health problem priorities for children in Indonesia. Prevention of stunting in toddlers is needed to avoid the long-term effects for both the toddlers and the public. Stunting prevention can be done by monitoring the growth of toddlers. Therefore, a system that can predict stunting conditions in toddlers is needed. Machine learning offers many methods that can be used to build a system to predict stunting conditions in toddlers. This research analyzes some machine learning models that are potentially suitable to predict stunting classes, which are K-Nearest Neighbor (KNN), Random Forest (RF), and Ensemble Learning called Boosted KNN (BK). The dataset has an imbalance issue in this research, with the stunting data at only 1% of the total dataset. Therefore, oversampling of the dataset is done by generating a random dataset based on the distribution of the data that are classified as the minority class. The results of elaborating on this oversampling are shown to be satisfying. Applying imbalanced data gives an average of 98% accuracy for all methods used; however, the F-1 score macro average is shown not optimal for each of the methods, with 51.95% for KNN, 52.45% for RF, and 53.55% for BK. After the data is balanced by oversampling, the F-1 score macro average for all methods substantially increases. The new results were 93.55% for KNN, 97.70% for RF, and 98.00% for BK, underscoring the critical role of addressing data imbalance in improving predictive accuracy.

Keywords: stunting, machine learning, k-nearest neighbor, random forest, ensemble learning

1. Introduction

Stunting ranks among the foremost health concerns for children in Indonesia, denoting a condition where children under the age of five fail to thrive due to chronic malnutrition within the initial 1000 days of life, resulting in stunted growth [2, 17, 11]. Addressing this issue is crucial to circumvent enduring consequences for toddlers and the wider public, preventing long-term health decline [12]. Notably, President Joko Widodo has set a target to reduce stunting prevalence from 21% in 2022 to 14% in 2024, as outlined on the https://sehatnegeriku.kemkes.go.id website. Effective stunting prevention involves meticulously monitoring toddler growth, underscoring the need for an implementable system to predict stunting conditions.

To predict stunting conditions in toddlers, a suitable algorithm to perform classification is needed. Based on the research in [15], the K-Nearest Neighbor performs well in predicting stunting conditions. In 2020, the stunting problem was discussed in [15], which used K-Nearest Neighbor (KNN) to predict toddler stunting conditions. The KNN method with K-Fold cross-validation got 95.26% accuracy as the highest result after several iterations. On the other hand, the research in [7] used the Naive Bayes method to predict the same stunting dataset and got 64.36% accuracy as the highest result.

In their effort to address stunting prevention, researchers [3] introduced the Sagita application in their recent research. Sagita is a valuable tool for parents, enabling them to monitor their toddlers' height and weight growth. Beyond mere observation, Sagita also sends timely notifications to parents, alerting them to the necessity of professional medical monitoring to ensure optimal nutrition for their child. The post-test results of the research indicate that Sagita effectively enhances parental understanding of stunting, encourages the adoption of healthier food alternatives, and disseminates other crucial information. Despite its success in information dissemination, Sagita currently possesses limited features [3]. A promising avenue for enhancement involves incorporating a machine learning-based prediction system into Sagita, as suggested in [3].

This study seeks to develop a robust machine-learning model dedicated to detecting stunting conditions in toddlers. Leveraging a dataset sourced from the Bojongsoang Community Health Center, consisting of over 6000 records of toddlers' physical measurements, the research focuses on training a model capable of providing early warnings. The comparison between three machine learning models—K-Nearest Neighbor, Random Forest, and the novel Ensemble Learning, Boosted KNN (a fusion of Random Forest principles with K-Nearest Neighbor)—is meticulously outlined to identify the most effective model. This research aims to contribute to stunting prevention in Indonesia by deploying machine learning models to provide timely alerts, potentially mitigating the impact of this critical health issue.

2. Methods

2.1 Research Design

This research starts by reviewing related literature to this research to get the background of the problem and acquire some method ideas to solve the problem. The research continues by collecting toddler data from the Bojongsoang Community Health Center. The collected data then needs to be understood first. After understanding each feature, some features are selected to be processed later. The features that have been selected are visualized so they can be easier to explain. Before the method chosen is implemented in the data, the data needs to be cleaned and converted to compatible data types to be processed. The data then needs to be checked to see whether it is balanced. The data will be balanced first if it turns out that the data is an imbalanced dataset. After preprocessing, K-Nearest Neighbor (KNN), Random Forest (RF), and Boosted KNN (BK) will try to classify the data. The performance of each method will be compared to get a conclusion. In order to get a better explanation, Fig. 1 will show the flowchart used for the research.

2.2 K-Nearest Neighbor

K-Nearest Neighbor (KNN) algorithm is one of the most common algorithms used to do classification or regression on data[5]. The idea of the KNN algorithm is to find similarity in each data, selecting k objects set that are the most similar, and labeling the new data based on the selected k objects set[13]. The similarity between data is determined by calculating the distance between data using Euclidean distance[13]. Euclidean distance from $l = (l_1, \cdots, l_n)$ to $m = (m_1, \cdots, m_n)$ is given as,

$$\text{euc}(l,m) = \sqrt{\sum_{i=1}^{n} (l_i - m_i)^2} \tag{1}$$

where n is the number of columns or features in the data and the smaller the distance, the more similar the data are[13]. The KNN algorithm classifies data based on the distance between each unlabeled data and all labeled data in the dataset. The classification is based on k-nearest neighbors (smallest distances), where k is the number of neighbors involved in the voting process. The class label for the test data is determined based on the majority votes[1]. The KNN algorithm can be seen in Algorithm 1: