

## ABSTRAK

Status gizi balita merupakan indikator penting dalam pemantauan kesehatan anak, namun permasalahan ketidakseimbangan data (*imbalanced data*) pada dataset kesehatan sering menjadi kendala dalam membangun model klasifikasi yang optimal. Penelitian ini bertujuan untuk menganalisis dan membandingkan kinerja algoritma *Naïve Bayes*, *Support Vector Machine* (SVM), dan *Random Forest* dalam mengklasifikasikan status gizi balita, serta menentukan algoritma dengan performa terbaik dan mengidentifikasi fitur yang paling berpengaruh dalam proses klasifikasi. Data yang digunakan merupakan data antropometri balita dari Puskesmas Purwokerto Selatan periode Januari hingga Desember 2025 dengan total 37.308 data yang setelah preprocessing menjadi 33.313 data. Variabel yang digunakan meliputi berat badan, tinggi badan, lingkar lengan atas (LiLA), usia, dan jenis kelamin sebagai fitur input, serta status gizi BB/TB sebagai variabel target yang terdiri dari 6 kelas. Metode yang diterapkan meliputi preprocessing data, pembagian dataset dengan rasio 80:20, penerapan *Synthetic Minority Over-sampling Technique* (SMOTE) pada data latih untuk mengatasi ketidakseimbangan kelas, dan validasi menggunakan *5-Fold Cross validation*. Hasil penelitian menunjukkan bahwa *Random Forest* dengan SMOTE merupakan model terbaik dengan nilai *accuracy* 95,75%, *precision* 95,88%, *recall* 95,75%, dan *F1-score* 95,79%. Penerapan SMOTE meningkatkan performa *Random Forest* dan *Support Vector Machine* (SVM), namun menurunkan performa *Naïve Bayes* akibat ketidaksesuaian asumsi distribusi normal dengan karakteristik data sintetis. Hasil *feature importance* menunjukkan berat badan (39,8%), tinggi badan (27,1%), LiLA (18,1%), usia (13,3%), dan jenis kelamin (1,6%) sebagai fitur paling berpengaruh dalam klasifikasi status gizi balita.

Kata kunci: *Random Forest*, klasifikasi Status Gizi Balita, SMOTE, *imbalance data*, antropometri balita.

## ABSTRACT

*The nutritional status of toddlers is an important indicator in monitoring children's health, however, the problem of imbalanced data in health datasets often becomes an obstacle in building optimal classification models. This study aims to analyze and compare the performance of Naïve Bayes, Support Vector Machine (SVM), and Random Forest algorithms in classifying toddler nutritional status, as well as to determine the best performing algorithm and identify the most influential features in the classification process. The data used is anthropometric data of toddlers from Puskesmas Purwokerto Selatan for the period January to December 2025 with a total of 37,308 data which after preprocessing became 33,313 data. The variables used include weight, height, mid-upper arm circumference (MUAC), age, and gender as input features, and BB/TB nutritional status as the target variable consisting of 6 classes. The methods applied include data preprocessing, dataset splitting with an 80:20 ratio, application of Synthetic Minority Over-sampling Technique (SMOTE) on training data to overcome class imbalance, and validation using 5-Fold Cross validation. The results show that Random Forest with SMOTE is the best model with accuracy 95.75%, precision 95.88%, recall 95.75%, and F1-score 95.79%. The application of SMOTE improved the performance of Random Forest and SVM, but decreased the performance of Naïve Bayes due to the incompatibility of the normal distribution assumption with the characteristics of synthetic data. Feature importance results show that weight (39.8%), height (27.1%), MUAC (18.1%), age (13.3%), and gender (1.6%) are the most influential features in toddler nutritional status classification.*

*Keywords: Random Forest, toddler nutritional status classification, SMOTE, Imbalance data, toddler anthropometry*