

ABSTRAK

Penelitian ini bertujuan untuk membandingkan kinerja algoritma *Extreme Gradient Boosting* (XGBoost) dan *Logistic Regression* dalam klasifikasi status stunting pada balita di Kecamatan Cilongok, Kabupaten Banyumas. Data yang digunakan berupa 7.089 catatan antropometri balita dari dua Puskesmas, dengan prevalensi stunting sebesar 18,51%. Setelah proses pembersihan data, diperoleh 6.984 record dengan proporsi stunting 18,67% yang digunakan dalam pemodelan. Tahap pra-pemrosesan meliputi parsing usia ke satuan bulan, imputasi median untuk nilai hilang, normalisasi dengan *Min-Max Scaling*, serta penanganan ketidakseimbangan kelas menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE). Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, F1-Score, dan ROC-AUC dengan skema 5-fold cross-validation. Hasil menunjukkan bahwa XGBoost memiliki performa lebih tinggi dibanding Logistic Regression, dengan rata-rata F1-Score sebesar $0,8765 \pm 0,0186$. Analisis interpretabilitas model dilakukan menggunakan metode SHAP (*Shapley Additive Explanations*) untuk mengidentifikasi kontribusi variabel terhadap prediksi. Hasil analisis SHAP konsisten menempatkan tinggi badan, usia, dan berat badan sebagai variabel paling berpengaruh terhadap klasifikasi stunting. Penelitian ini menyimpulkan bahwa XGBoost lebih unggul dalam akurasi dan konsistensi dibanding Logistic Regression, sementara metode SHAP memberikan transparansi dalam interpretasi model. Temuan ini diharapkan dapat mendukung tenaga kesehatan dalam memahami faktor dominan stunting serta meningkatkan efektivitas pemantauan gizi balita.

Kata kunci: stunting, SHAP, xgboost, regresi logistik, klasifikasi

ABSTRACT

This study aimed to compare the performance of Extreme Gradient Boosting (XGBoost) and Logistic Regression algorithms in classifying stunting status among children under five in Cilongok District, Banyumas Regency. The dataset consisted of 7,089 anthropometric records collected from two community health centers, with an initial stunting prevalence of 18.51%. After the data cleaning process, 6,984 records were retained for modeling, resulting in a stunting proportion of 18.67%. The preprocessing stage included age parsing into months, median imputation for missing values, Min-Max Scaling normalization, and class imbalance handling using the Synthetic Minority Over-sampling Technique (SMOTE). Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics under a 5-fold cross-validation scheme. The results demonstrated that XGBoost outperformed Logistic Regression, achieving an average F1-score of 0.8765 ± 0.0186 . Model interpretability was assessed using SHAP (Shapley Additive Explanations) to identify the contribution of each variable to the prediction outcomes. The SHAP analysis consistently identified height, age, and weight as the most influential variables in stunting classification. This study concludes that XGBoost provides superior accuracy and consistency compared to Logistic Regression, while SHAP enhances model transparency and interpretability. These findings are expected to support healthcare professionals in understanding the dominant factors associated with stunting and improving the effectiveness of nutritional monitoring among children under five.

Keywords: stunted, SHAP, xgboost, logistic regression, classification