INTISARI

OPTIMALISASI KLASIFIKASI RISIKO STUNTING BERDASARKAN ASPEK SOSIAL EKONOMI DAN LINGKUNGAN DENGAN METODE SAMPLING DAN COST-SENSITIVE LEARNING UNTUK MENGATASI DATA TIDAK SEIMBANG

Marina Indah Prasasti

23/530768/PPA/06744

Ketidakseimbangan kelas merupakan permasalahan umum yang sering ditemukan pada dataset *stunting*, di mana jumlah data pada kelas minoritas (*Stunted* dan *Severely Stunted*) jauh lebih sedikit dibandingkan dengan kelas mayoritas (*Normal*). Ketidakseimbangan ini menyebabkan model klasifikasi *Machine Learning* (ML) cenderung bias terhadap kelas mayoritas, sehingga kinerja prediksi pada kelas minoritas menjadi rendah.

Penelitian ini bertujuan untuk mengoptimalkan klasifikasi risiko *stunting* dengan mengatasi ketidakseimbangan kelas melalui dua pendekatan, yaitu pendekatan sampling dan *cost-sensitive learning*. Pendekatan sampling yang digunakan meliputi Borderline SMOTE, IHT, dan SMOTE-ENN, yang diterapkan menggunakan model *stacking ensemble* dengan Gradient Boosting dan XGBoost sebagai *base learners*, serta Random Forest sebagai *meta learner*. Sementara itu, pendekatan *cost-sensitive learning* menggunakan metode AdaCost. Data yang digunakan dalam penelitian ini diperoleh dari Survei Kesehatan Indonesia tahun 2023 yang berjumlah 306.281 data balita.

Hasil pengujian menunjukkan bahwa Model tanpa penanganan ketidakseimbangan memiliki recall hanya 0.33 dan F1-score 0.29. Sedangkan Borderline SMOTE dengan *sampling_strategy* 0.75 menghasilkan recall sebesar 0.35 dan F1-score sebesar 0.34, menjadi hasil terbaik dalam mendeteksi kelas minoritas. Pendekatan IHT memberikan F1-score yang stabil pada kisaran 0.32–0.33. Sebaliknya, SMOTE-ENN menunjukkan performa terendah, dengan F1-score hanya 0.25 pada rasio 0.5 dan 0.23 pada *sampling_strategy* 0.75. Sementara itu, pendekatan *cost-sensitive learning* menggunakan AdaCost belum mampu memberikan perbaikan terhadap bias kelas mayoritas.

Kata Kunci: *Stunting*, Ketidakseimbangan Kelas, Resampling, *Cost-Sensitive Learning*, *Stacking Ensemble*

ABSTRACT

OPTIMIZING STUNTING RISK CLASSIFICATION USING SOCIOECONOMIC AND ENVIRONMENTAL FACTORS WITH SAMPLING AND COST-SENSITIVE LEARNING METHODS IN IMBALANCED DATASET

Marina Indah Prasasti
23/530768/PPA/06744

Class imbalance is a common issue frequently encountered in stunting datasets, where the number of records in the minority classes (Stunted and Severely Stunted) is significantly lower than in the majority class (Normal). This imbalance leads Machine Learning (ML) classification models to be biased toward the majority class, resulting in poor predictive performance for the minority classes.

This study aims to optimize stunting risk classification by addressing class imbalance through two approaches: sampling and cost-sensitive learning. The sampling methods used include Borderline SMOTE, IHT, and SMOTE-ENN, which are applied using a stacking ensemble model with Gradient Boosting and XGBoost as base learners and Random Forest as the meta learner. Meanwhile, the cost-sensitive learning approach employs the AdaCost method. The dataset used in this study was obtained from the 2023 Indonesian Health Survei or Survei Kesehatan Indonesia, comprising 306.281 child records.

The results show that Borderline SMOTE with a 0.75 sampling ratio on the stacking ensemble model achieved the best performance in detecting minority classes, with a recall of 0.35 and an F1-score of 0.34. The model without any imbalance handling produced a recall of only 0.33 and an F1-score of 0.29. The IHT approach yielded relatively stable F1-scores, ranging from 0.32 to 0.33. In contrast, SMOTE-ENN showed the lowest performance, with F1-scores of just 0.25 at a 0.5 sampling ratio and 0.23 at 0.75. Meanwhile, the cost-sensitive learning approach using AdaCost did not show any significant improvement in mitigating majority class bias.

Keywords: Stunting, Class Imbalance, Resampling, Cost-Sensitive Learning, Stacking Ensemble