

INTISARI

Ketidakseimbangan kelas (class imbalance) merupakan permasalahan mendasar dalam machine learning yang terjadi ketika distribusi kelas dalam dataset tidak proporsional. Kondisi ini menyebabkan model klasifikasi cenderung berpihak pada kelas mayoritas dan mengabaikan kelas minoritas, padahal kelas minoritas justru seringkali lebih penting untuk dideteksi. Permasalahan ini ditemukan secara luas pada berbagai domain kritis, seperti deteksi penipuan keuangan, diagnosis penyakit, dan keamanan siber. Salah satu pendekatan yang paling banyak digunakan untuk mengatasi ketidakseimbangan kelas adalah teknik oversampling berbasis synthetic minority, yang bekerja pada tahap preprocessing sebelum pelatihan model. Meskipun demikian, belum ada konsensus yang jelas mengenai teknik mana yang paling efektif, terutama ketika dikombinasikan dengan algoritma boosting modern yang kini semakin dominan di industri.

Penelitian ini melakukan analisis komparatif secara empiris terhadap empat teknik oversampling utama, yaitu SMOTE, ADASYN, Borderline-SMOTE, dan SMOTE-ENN, pada tiga dataset benchmark UCI Machine Learning Repository yang memiliki karakteristik berbeda: Bank Marketing (IR 7,55:1), Glass Identification (IR 15,46:1), dan Musk Version I (IR 1,3:1). Eksperimen dilakukan menggunakan dua algoritma boosting modern, yaitu XGBoost dan LightGBM, dan dievaluasi melalui metrik F1-Score, G-Mean, dan ROC-AUC menggunakan skema 5-fold stratified cross-validation.

Hasil penelitian menunjukkan bahwa tidak terdapat satu teknik oversampling yang secara universal unggul pada seluruh dataset. SMOTE-ENN memberikan peningkatan performa paling signifikan pada dataset Bank Marketing (F1-Score 0,6025; G-Mean 0,8161), sedangkan ADASYN secara konsisten unggul pada dataset Glass Identification (F1-Score 0,8333) dan Musk Version I (F1-Score 0,9211) sekaligus menjadi teknik dengan kontribusi paling stabil lintas seluruh kombinasi dataset dan classifier. Teknik oversampling berbasis jarak seperti SMOTE-ENN dan Borderline-SMOTE terbukti rentan terhadap curse of dimensionality pada dataset berdimensi tinggi, sementara kombinasi ADASYN dengan LightGBM menghasilkan performa terbaik secara keseluruhan, melampaui capaian penelitian pembandingan (Sutarman dkk., 2024). Penelitian ini turut menghasilkan panduan empiris pemilihan teknik oversampling berdasarkan karakteristik dataset, serta catatan metodologis mengenai pentingnya pelaporan nilai cross-validation pada evaluasi dataset berukuran kecil. Hasil ini memberikan rekomendasi berbasis bukti empiris mengenai teknik oversampling yang paling sesuai untuk karakteristik dataset dengan tingkat ketidakseimbangan yang berbeda, sekaligus menjadi referensi ilmiah bagi peneliti dan praktisi dalam memilih strategi preprocessing yang tepat ketika menggunakan XGBoost dan LightGBM pada data nyata yang tidak seimbang.

Kata kunci: Oversampling, Class Imbalance, SMOTE, ADASYN, Borderline-SMOTE, SMOTE-ENN, XGBoost, LightGBM, Klasifikasi

ABSTRACT

Class imbalance is a fundamental problem in machine learning that occurs when the class distribution within a dataset is disproportionate. This condition causes classification models to favor the majority class while neglecting the minority class, even though the minority class is often more critical to detect. This problem is widely encountered across various high-stakes domains, including financial fraud detection, medical diagnosis, and cybersecurity. One of the most widely adopted approaches to address class imbalance is the synthetic minority oversampling technique, which operates at the preprocessing stage prior to model training. Nevertheless, no clear consensus exists regarding which technique performs best, particularly when combined with modern boosting algorithms that are increasingly dominant in industry.

This study conducted an empirical comparative analysis of four major oversampling techniques SMOTE, ADASYN, Borderline-SMOTE, and SMOTE-ENN on three UCI Machine Learning Repository benchmark datasets with distinct characteristics: Bank Marketing (IR 7.55:1), Glass Identification (IR 15.46:1), and Musk Version I (IR 1.3:1). Experiments were conducted using two modern boosting algorithms, namely XGBoost and LightGBM, and evaluated through F1-Score, G-Mean, and ROC-AUC metrics under a 5-fold stratified cross-validation scheme.

The results showed that no single oversampling technique universally outperformed the others across all datasets. SMOTE-ENN produced the most significant performance gain on the Bank Marketing dataset (F1-Score 0.6025; G-Mean 0.8161), while ADASYN consistently outperformed the other techniques on the Glass Identification (F1-Score 0.8333) and Musk Version I (F1-Score 0.9211) datasets, and demonstrated the most stable contribution across all dataset-classifier combinations. Distance-based oversampling techniques such as SMOTE-ENN and Borderline-SMOTE proved vulnerable to the curse of dimensionality on high-dimensional datasets, while the ADASYN and LightGBM combination achieved the best overall performance, surpassing a comparable prior study (Sutarman et al., 2024). This study also produced empirical guidance for selecting oversampling techniques based on dataset characteristics, along with methodological notes on the importance of reporting cross-validation scores when evaluating small datasets. These findings provide evidence-based recommendations on the most suitable oversampling technique for datasets with varying levels of class imbalance, as well as a practical reference for researchers and practitioners in selecting the appropriate preprocessing strategy when employing XGBoost and LightGBM on real-world imbalanced data.

Keywords: Oversampling, Class Imbalance, SMOTE, ADASYN, Borderline-SMOTE, SMOTE-ENN, XGBoost, LightGBM, Classification