ABSTRAK

Sidi, Widya Dharma, (2026), Analisis Komparasi Algoritma C4.5, Random Forest, dan XGBoost Menggunakan Pendekatan Stratified Cross Validation dan SMOTE untuk Klasifikasi Kelulusan Tepat Waktu pada Imbalanced Small Dataset. Tesis Ilmu Komputer, Program Pascasarjana, Universitas Pendidikan Ganesha.

Kata Kunci: *Educational Data Mining, Imbalanced Small Dataset*, Klasifikasi Kelulusan, *Ensemble Learning*, SMOTE, *Feature Importance*.

Penerapan *Educational Data Mining* (EDM) untuk sistem peringatan dini akademik sering kali dihadapkan pada tantangan ketersediaan data yang terbatas (*small dataset*) dan distribusi kelas yang tidak seimbang (*imbalanced class*). Di samping itu, mayoritas model prediktif saat ini sangat bergantung pada fitur statis yang kurang merepresentasikan dinamika aktual mahasiswa selama masa studi. Penelitian ini berfokus pada data historis akademik di Universitas PGRI Mahadewa Indonesia (UPMI Bali) angkatan 2020 dan 2021, guna merancang model prediksi kelulusan tepat waktu yang stabil saat dihadapkan dengan bias ketimpangan data. Penelitian ini bertujuan untuk menginvestigasi ketangguhan (*robustness*) algoritma berbasis pohon, yaitu C4.5 (sebagai *baseline*), *Random Forest* (*bagging*), dan XGBoost (*boosting*). Untuk meningkatkan sensitivitas sistem, diusulkan ekstraksi fitur dinamis berupa tren Indeks Prestasi Semester (IPS) dan frekuensi tunggakan finansial. Penanganan ketimpangan kelas dilakukan melalui injeksi *Synthetic Minority Over-sampling Technique* (SMOTE) secara eksklusif pada subset data latih. Evaluasi komprehensif dilakukan menggunakan *Stratified 10-Fold Cross-Validation* dan divalidasi silang menggunakan dua *benchmark dataset* publik (Haberman dan Breast Cancer) guna menghindari bias lokal institusi. Hasil penelitian membuktikan bahwa meskipun Uji Friedman menunjukkan tidak ada perbedaan kinerja yang signifikan secara statistik antar-algoritma, XGBoost terpilih sebagai model usulan terbaik secara praktis. XGBoost mencapai *F1-Score* tertinggi (0,8207), tingkat stabilitas terbaik (simpangan baku 0,0387), dan kebal terhadap *overfitting* (*generalization gap* 0,0162). Uji Wilcoxon mengonfirmasi bahwa penerapan SMOTE secara signifikan mampu meningkatkan sensitivitas deteksi (*Recall*) model. Lebih lanjut, ekstraksi *Feature Importance* mengungkap bahwa capaian IPS Semester 4 dan Frekuensi Tunggakan merupakan determinan utama penentu risiko keterlambatan, sedangkan identitas demografis terbukti tidak relevan. Luaran klasifikasi ini tidak hanya menghasilkan prediksi label, melainkan bertransformasi menjadi instrumen pendukung keputusan strategis yang objektif bagi pemangku kebijakan akademik.

ABSTRACT

Sidi, Widya Dharma, (2026), Comparative Analysis of C4.5, Random Forest, and XGBoost Algorithms Using Stratified Cross Validation and SMOTE Approaches for On-Time Graduation Classification on an Imbalanced Small Dataset. Thesis, Computer Science, Postgraduate Program, Ganesha University of Education.

Keywords: Educational Data Mining, Imbalanced Small Dataset, Graduation Classification, Ensemble Learning, SMOTE, Feature Importance.

The application of Educational Data Mining (EDM) for academic early warning systems is often confronted with the challenges of limited data availability (small datasets) and uneven class distribution (imbalanced classes). Furthermore, the majority of current predictive models rely heavily on static features that do not adequately represent the actual dynamics of students during their study period. This study focuses on historical academic data at Universitas PGRI Mahadewa Indonesia (UPMI Bali) for the 2020 and 2021 cohorts, aiming to design an on-time graduation prediction model that is robust against data imbalance bias. This study investigates the robustness of tree-based algorithms, namely C4.5 (as the baseline), Random Forest (bagging), and XGBoost (boosting). To enhance the system's sensitivity, this research proposes the extraction of dynamic features in the form of Semester Grade Point Average (SGPA) trends and the frequency of financial arrears. The class imbalance issue was addressed by injecting the Synthetic Minority Over-sampling Technique (SMOTE) exclusively into the training data subset. A comprehensive evaluation was conducted using Stratified 10-Fold Cross-Validation and cross-validated using two public benchmark datasets (Haberman and Breast Cancer) to prevent local institutional bias. The results proved that although the Friedman test showed no statistically significant performance difference among the algorithms, XGBoost was selected as the best proposed model practically. XGBoost achieved the highest F1-Score (0.8207), the best stability level (standard deviation of 0.0387), and was immune to overfitting (generalization gap of 0.0162). The Wilcoxon test confirmed that the application of SMOTE significantly improved the model's detection sensitivity (Recall). Furthermore, Feature Importance extraction revealed that Semester 4 SGPA and Arrear Frequency were the primary determinants of study delay risk, while demographic identities proved irrelevant. This classification output does not merely generate prediction labels but transforms into an objective strategic decision-support instrument for academic policymakers.