ABSTRAK

Ketidakseimbangan data merupakan salah satu permasalahan utama dalam pemodelan klasifikasi, khususnya dalam prediksi gagal bayar pinjaman, di mana jumlah peminjam yang tidak gagal bayar jauh lebih dominan dibandingkan yang mengalami gagal bayar. Kondisi ini menyebabkan model cenderung bias terhadap kelas mayoritas sehingga mengabaikan kelas minoritas yang justru penting. Penelitian ini bertujuan untuk mengevaluasi kinerja algoritma *Light Gradient Boosting Machine* (LightGBM) dalam mengklasifikasikan risiko gagal bayar pinjaman serta menganalisis efektivitas teknik penyeimbangan data, yaitu *Synthetic Minority Over-sampling Technique* (SMOTE) dan *Random Under-Sampling* (RUS).

Data yang digunakan diperoleh dari *platform* Kaggle, terdiri atas 255.348 entri dengan 17 atribut yang dipilih melalui proses seleksi fitur. Evaluasi model dilakukan menggunakan metode *K-Fold Cross Validation* dengan nilai K = 3, 5, dan 10, serta pengaturan *hyperparameter* melalui *Grid Search* dan *Random Search*. Hasil pengujian menunjukkan bahwa penerapan SMOTE memberikan hasil akurasi tertinggi, yaitu sebesar 92,88% pada konfigurasi 17 atribut dengan K = 10. Akurasi pada data tanpa balancing tercatat sebesar 88,64%, sementara metode RUS hanya mencapai 71,52%. Seluruh atribut yang digunakan terbukti memberikan kontribusi terhadap pembentukan prediksi kelas default.

Dari hasil tersebut dapat disimpulkan bahwa kombinasi algoritma LightGBM, pemilihan atribut yang optimal, serta penerapan teknik balancing menggunakan SMOTE merupakan pendekatan yang efektif dalam meningkatkan performa klasifikasi risiko gagal bayar. Penelitian ini diharapkan dapat menjadi acuan dalam pengembangan sistem pendukung keputusan di sektor keuangan berbasis data yang tidak seimbang.

Kata kunci: LightGBM, gagal bayar pinjaman, data tidak seimbang, SMOTE, RUS, klasifikasi.

ABSTRACT

Imbalanced data is a major challenge in classification modeling, particularly in loan default prediction, where the number of non-default borrowers significantly exceeds those who default. This condition often leads to a model that is biased toward the majority class while neglecting the minority class, which is critical in risk assessment. This study aims to evaluate the performance of the Light Gradient Boosting Machine (LightGBM) algorithm in classifying loan default risk and to analyze the effectiveness of data balancing techniques, namely Synthetic Minority Over-sampling Technique (SMOTE) and Random Under-Sampling (RUS).

The dataset used was obtained from the Kaggle platform, consisting of 255,348 entries with 17 selected attributes through feature selection. Model evaluation was conducted using the K-Fold Cross Validation method with K values of 3, 5, and 10, and hyperparameter tuning through Grid Search and Random Search. The results show that SMOTE achieved the highest accuracy of 92.88% on the configuration with 17 attributes and K = 10. The accuracy without balancing was 88.64%, while the RUS method only reached 71.52%. All selected attributes were found to contribute meaningfully to the classification of the default class.

These findings indicate that the combination of the LightGBM algorithm, optimal feature selection, and data balancing using SMOTE is an effective approach to improve the performance of loan default classification. This research is expected to serve as a reference for developing decision support systems in the financial sector that deal with imbalanced data.

***Keywords**: LightGBM, loan default, imbalanced data, SMOTE, RUS, classification.*