# ABSTRAK

*Imbalanced Data* dalam klasifikasi adalah salah satu topik yang penting dalam *data mining* dan *Machine Learning*. Pada kasus nyata di kehidupan, dataset mengenai nasabah yang berkaitan dengan penentuan kelayakan pemberian kredit adalah salah satu dari sekian jenis dataset yang identik dengan ketimpangan tinggi antar kelas di dalamnya. Dalam klasifikasi tidak seimbang biner terdapat salah satu kelas yang memiliki lebih banyak *instance* sehingga dimaknai sebagai kelas mayoritas dan kelas lain dengan lebih sedikit *instance* yang dimaknai sebagai kelas minoritas. Model yang terbuat dari data tidak seimbang beresiko besar menyebabkan prediksi kelas minoritas yang rendah dan *overfitting* karena informasi dari kelas mayoritas lebih mendominasi daripada kelas minoritas, hal tersebut berdampak pada diragukannya kualitas data dan keputusan dalam sistem klasifikasi.

Menggunakan metode SMOTE yang diusulkan sebelum masuk tahap klasifikasi dengan Random Forest. Dataset *imbalanced* akan melalui proses ekstrapolasi dengan SMOTE sehingga dihasilkan data sintesis pada data minoritas (*bad*) sebanyak 400 data sintesis dari yang sebelumnya hanya berjumlah 200 data, persebaran data menjadi seimbang dengan masing-masing kelas *good* dan *bad* berjumlah 600 data. Selanjutnya pada tahap klasifikasi dengan Random Forest, data yang telah seimbang akan melalui proses *resampling bootstrap* dengan pengembalian sehingga dataset terbagi ke dalam beberapa subset data. Beberapa pohon (CART) akan terbentuk sesuai dengan jumlah subset, setiap pohon akan dianalisis sehingga menghasilkan prediksi masing-masing. Hasil klasifikasi akhir ditentukan berdasarkan voting atau pengambilan suara terbanyak dari hasil prediksi setiap pohon yang terbentuk. Model yang telah dibangun akan diuji kehandalannya dengan *confusion matrix*. Dari tabel *confusion matrix* akan diketahui nilai *accuracy, precision, recall, specificity, f1-score,* dan AUC-ROC.

Hasil pengujian terbaik didapatkan pada model SMOTE-Random Forest dengan nilai akurasi 84%, presisi 86%, sensitivitas 81%, spesifisitas 87%, *F1 Score* 83%, dan AUC 84%. Sedangkan pada metode Random Forest tanpa SMOTE, didapatkan nilai akurasi 74%, presisi 81%, sensitivitas 87%, spesifisitas 34%, *F1 Score* 84%, dan AUC 60%. Terdapat kenaikan sebesar 10% pada akurasi, 5% pada presisi, 53% pada spesifisitas, dan 24% pada AUC. Dengan hasil tersebut, SMOTE-Random Forest dikategorikan sebagai model klasifikasi yang baik.



**Kata kunci :** Klasifikasi, *Credit scoring*, *Imbalanced data*, SMOTE, Random Forest.

## ABSTRACT

Imbalanced data in classification is one of the most important topics in data mining and machine learning. In real life cases, datasets about customers related to determining credit worthiness are one of the many types of datasets that are identical with high disparities between classes in them. In binary unbalanced classification there is one class that has more instances so it is interpreted as the majority class and another class with fewer instances is interpreted as the minority class. Models made of unbalanced data have a big risk of causing low and overfitting minority class predictions because information from the majority class dominates over the minority class, this has an impact on doubting the quality of the data and decisions in the classification system.

Using the proposed SMOTE method before entering the classification stage with Random Forest. Imbalanced datasets will go through an extrapolation process with SMOTE so that 400 synthesis data are produced on minority (bad) data from previously only 200 data, the data distribution becomes balanced with each good and bad class totaling 600 data. Furthermore, at the classification stage with Random Forest, the balanced data will go through a bootstrap resampling process with returns so that the dataset is divided into several data subsets. Several trees (CART) will be formed according to the number of subsets, each tree will be analyzed to produce its own predictions. The results of the final classification are determined based on voting or voting the most votes from the predicted results of each tree that is formed. The model that has been built will be tested for reliability with a confusion matrix. From the confusion matrix table it will be known the value of accuracy, precision, recall, specificity, f1-score, and AUC-ROC.

The best test results were obtained in the SMOTE-Random Forest model with an accuracy of 84%, precision of 86%, sensitivity of 81%, specificity of 87%, F1 Score of 83%, and AUC of 84%. Meanwhile, the Random Forest method without SMOTE obtained an accuracy of 74%, precision of 81%, sensitivity of 87%, specificity of 34%, F1 score of 84%, and AUC of 60%. There was a 10% increase in accuracy, 5% in precision, 53% in specificity, and 24% in AUC. With these results, SMOTE-Random Forest is categorized as a good classification model.



**Keywords :** *Classification, Credit scoring, Imbalanced data*, SMOTE, Random Forest.