INTISARI

Penelitian ini bertujuan untuk memprediksi penjualan buku dengan memanfaatkan model regresi *Random Forest* yang dikombinasikan dengan teknik augmentasi data menggunakan *Conditional Generative Adversarial Network* (cGAN). Dataset yang digunakan bersumber dari Zahira Media Publisher, mencakup data penjualan buku dari berbagai platform marketplace seperti Shopee, Tokopedia, Google Play Books, dan situs resmi penerbit untuk periode 2021–2024. Fokus penelitian adalah mengatasi permasalahan ketidakseimbangan data, khususnya terkait genre buku yang kurang terwakili dalam dataset. Proses penelitian diawali dengan eksplorasi data dan preprocessing, kemudian dilakukan pelatihan model *Random Forest* pada data yang belum seimbang. Hasilnya menunjukkan bahwa performa model masih kurang optimal, dengan nilai R^2 hanya sebesar 0,22 dan MAPE sebesar 163,56%, yang mengindikasikan adanya bias terhadap genre mayoritas. Untuk mengatasi hal tersebut, dilakukan augmentasi data menggunakan cGAN yang berhasil menyeimbangkan distribusi genre dalam dataset. Setelah dilakukan augmentasi, model *Random Forest* dilatih ulang dan menunjukkan peningkatan kinerja yang signifikan, dengan nilai R^2 mencapai 0,96 dan MAPE menurun menjadi 9,93%. Analisis feature importance menunjukkan bahwa faktor paling berpengaruh dalam prediksi penjualan buku adalah rating pengguna, diikuti oleh platform, harga, dan genre. Temuan ini membuktikan bahwa augmentasi data menggunakan cGAN dapat meningkatkan akurasi model regresi dalam konteks data yang tidak seimbang. Hasil penelitian diharapkan dapat memberikan kontribusi bagi penerbit dan pemasar dalam merumuskan strategi berbasis data untuk optimalisasi penjualan buku.

Kata kunci: prediksi penjualan buku, *Random Forest*, Conditional GAN, augmentasi data, machine learning

.

ABSTRACT

This research aims to predict book sales by utilizing the Random Forest regression model combined with data augmentation through Conditional Generative Adversarial Network (cGAN). The dataset used in this study is sourced from Zahira Media Publisher, covering book sales data from various marketplace platforms including Shopee, Tokopedia, Google Play Books, and the publisher's official website for the period 2021–2024. The research focuses on addressing issues arising from data imbalance, particularly related to underrepresented book genres in the dataset. The research process begins with data exploration and preprocessing, followed by training the initial Random Forest model on unbalanced data. The results show that the model's performance was suboptimal, with an R^2 score of only 0.22 and a high MAPE value of 163.56%, indicating bias toward majority genres. To address this, cGAN was applied for data augmentation, successfully balancing the genre distribution. After augmentation, the Random Forest model was retrained, showing a significant improvement with an R^2 score of 0.96 and a reduced MAPE of 9.93%. Feature importance analysis revealed that the most influential factor in book sales prediction is user rating, followed by platform, price, and genre. These findings demonstrate that data augmentation using cGAN can effectively improve prediction accuracy in regression models with imbalanced data. The results are expected to provide valuable insights for publishers and marketers in formulating data-driven strategies for sales optimization.

Keywords: book sales prediction, Random Forest, Conditional GAN, data augmentation, machine learning