

INTISARI

PERBANDINGAN METODE *SMOTE-k-NN* DAN METODE *SMOTE-ENN* DALAM PENANGANAN KLASIFIKASI DATA TIDAK SEIMBANG

oleh

Dennis Dharmawira Angjaya

16/398654/PA/17615

Perkembangan pesat dalam data menuntut perkembangan yang sama pesatnya dalam proses pengolahan data agar data dapat diolah secara lebih cepat dan lebih tepat. Data yang jumlahnya banyak dan beragam terkadang perlu untuk dilakukan klasifikasi sehingga terlihat lebih tertata. Dalam klasifikasi data, seringkali ditemukan kasus data yang tidak seimbang. Melakukan klasifikasi pada data tidak seimbang seringkali menimbulkan masalah berupa kurangnya kemampuan model untuk memprediksi data dari kelas minoritas. Metode *SMOTE* sering digunakan sebagai metode *oversampling* untuk menyeimbangkan data dengan membentuk instance sintetis untuk kelas minoritas. Metode *ENN* merupakan metode *undersampling* yang menghapus data jika label kelasnya tidak sama dengan mayoritas label kelas dari tetangga terdekatnya. *k-NN* merupakan metode klasifikasi yang melabelkan sampel baru berdasarkan *k*-tetangga terdekat sampel baru tersebut. *Random forest* merupakan metode klasifikasi yang terdiri dari gabungan pohon klasifikasi yang saling independen, dimana prediksi klasifikasi diperoleh melalui proses voting jumlah terbanyak dari pohon-pohon klasifikasi yang terbentuk. Pada skripsi ini dilakukan analisis perbandingan *SMOTE-k-NN* dan *SMOTE-ENN* dalam penanganan data *Wine Quality* yang tidak seimbang. Analisis klasifikasi dengan metode *k-NN* diterapkan pada *SMOTE-k-NN*, sementara itu metode *random forest* diterapkan pada *SMOTE-ENN*. Dari analisis yang dilakukan diperoleh kesimpulan bahwa metode *SMOTE-k-NN* menghasilkan performa klasifikasi yang lebih baik dibandingkan metode *SMOTE-ENN*.

Kata kunci: klasifikasi, data tidak seimbang, *SMOTE*, *SMOTE-k-NN*, *SMOTE-ENN*

ABSTRACT

COMPARISON OF SMOTE – k -NN METHOD AND SMOTE-ENN METHOD IN HANDLING IMBALANCED DATA CLASSIFICATION

by

Dennis Dharmawira Angjaya

16/398654/PA/17615

The rapid development of data demands equally rapid development of data processing so that data can be processed quicker and more accurate. Large and various data is sometimes needed to be classified so that it looks more organized. In data classification, cases of imbalanced data are often found. Attempting classification to imbalanced data often causes problem such as lack of model ability to predict data from minority class. SMOTE method is often used as an oversampling method to balance data by creating synthetic instances for minority class. ENN method is an undersampling method that deletes data if the class label is different from the majority of class label from its nearest neighbors. k -NN is a classification method that labels new samples based on its k -nearest neighbors. Random forest is a classification method consisting of a combination of mutually exclusive decision trees, where the classification prediction is obtained through a voting process of majority from formed decision trees. In this undergraduate thesis, comparison analysis of SMOTE- k -NN and SMOTE-ENN in handling imbalance data of Wine Quality is carried out. Classification analysis with k -NN method is applied to SMOTE- k -NN, while random forest method is applied to SMOTE-ENN. From the conducted analysis, it's concluded that SMOTE- k -ENN method produced better classification performance compared to SMOTE-ENN method.

Keywords: classification, imbalanced data, SMOTE, SMOTE- k -NN, SMOTE-ENN