INTISARI

PERBANDINGAN ANALISIS KLASIFIKASI SMOTE RANDOM FOREST DAN SMOTE K-NEAREST NEIGHBORS PADA DATA TIDAK SEIMBANG

Oleh

JUS PRASETYA

19/448805/PPA/05888

Pada studi *machine learning*, analisis klasifikasi bertujuan untuk meminimalkan kesalahan klasifikasi dan juga memaksimalkan hasil akurasi prediksi. Klasifikasi adalah pembelajaran penting untuk pengenalan pola (karakteristik pada data). Klasifikasi kelas tidak seimbang adalah proses untuk mengestimasi sampel baru anggota kelas dari kumpulan data yang ada di mana ukuran kelas bervariasi secara signifikan. Pada SMOTE data kelas minoritas dipelajari dan diekstrapolasi sehingga dapat menghasilkan sampel sintesis baru. *Random forests* adalah suatu metode klasifikasi yang terdiri dari gabungan pohon klasifikasi yang saling independen. Prediksi klasifikasi diperoleh melalui proses voting (jumlah terbanyak) dari pohon-pohon klasifikasi yang terbentuk. *k-Nearest Neighbors* yang merupakan metode klasifikasi yang melabelkan sampel baru berdasarkan k-tetangga terdekat dari sampel baru tersebut. *Synthetic Minority Oversampling Technique* (SMOTE) membangkitkan data sintesis pada kelas minoritas yakni kelas 1 (kanker serviks) menjadi 585 responden pengamatan (sampel) sehingga total responden pengamatan menjadi 1208 sampel. *Smote random forest* menghasilkan akurasi sebesar 96,28%, sensitifitas 99,17%, spesifisitas 93,44%, presisi 93,70%, AUC 96,30% dan akurasi dengan 5-*fold cross validation* sebesar 95,65%. *Smote k-nearest neighbors* menghasilkan akurasi sebesar 87,60%, sensitifitas 77,50%, spesifisitas 97,54%, presisi 96,88%, AUC 82,27% dan akurasi dengan 5-*fold cross validation* sebesar 87,78%. *Smote random forest* menghasilkan model klasifikasi yang sempurna, klasifikasi *smote k-nearest neighbors* menghasilkan model klasifikasi yang baik sedangkan klasifikasi *random forest* dan *k-nearest neighbors* pada *imbalanced data* menghasilkan model klasifikasi yang gagal.

Kata Kunci : **Machine Learning, Klasifikasi, SMOTE, Random Forest, k-Nearest Neighbors**

ABSTRACT

COMPARISON OF SMOTE RANDOM FOREST AND SMOTE K-NEAREST NEIGHBORS CLASSIFICATION ANALYSIS ON IMBALANCED DATA

By

JUS PRASETYA

19/448805/PPA/05888

In machine learning study, classification analysis aims to minimize misclassification and also maximize the results of prediction accuracy. Classification is an important learning for pattern recognition (characteristics in data). Imbalanced class classification is the process of estimating a new sample of class members from an existing data set where the class size varies significantly. SMOTE minority class data is studied and extrapolated so that it can produce new synthetic samples. Random forest is a classification method consisting of a combination of mutually independent classification trees. Classification prediction is obtained through a voting process (the highest number) of the classification trees formed. k-Nearest Neighbors which is a classification method that labels the new sample based on the k-nearest neighbors of the new sample. The Synthetic Minority Oversampling Technique (SMOTE) generates synthesis data in the minority class, namely class 1 (cervical cancer) to 585 observation respondents (samples) so that the total observation respondents are 1208 samples. Smote random forest resulted an accuracy of 96.28%, sensitivity 99.17%, specificity 93.44%, precision 93.70%, AUC 96.30% and accuracy with 5-fold cross validation of 95.65%. Smote k-nearest neighbors resulted an accuracy of 87.60%, sensitivity 77.50%, specificity 97.54%, precision 96.88%, AUC 82.27% and accuracy with 5-fold cross validation of 87.78%. Smote random forest produces a perfect classification model, smote k-nearest neighbors produces a good classification model, while the random forest and k-nearest neighbors classification on imbalanced data results a failed classification model.

Keywords: **Machine Learning, Classification, SMOTE, Random Forest, k-Nearest Neighbors**