INTISARI

ANALISIS *RANDOM OVER SAMPLING* MENGGUNAKAN KLASIFIKASI *RANDOM FOREST* DAN *C5.0* PADA DATA TIDAK SEIMBANG

Oleh

MARLEN WATTIMENA
20/466528/PPA/06094

Analisis klasifikasi merupakan salah satu studi machine learning yang bertujuan untuk menemukan suatu fungsi keputusan yang secara akurat memprediksi kelas dari data *testing* yang berasal dari fungsi distribusi yang sama dengan data untuk *training*. Klasifikasi adalah proses mengelompokkan objek kedalam suatu kelas atau kategori yang telah ditentukan. Ketidakseimbangan data adalah salah satu masalah dalam klasifikasi yang merupakan kondisi data tidak berimbang antara kelas data satu dengan kelas data yang lain. Ketidakseimbangan data dapat diselesaikan dengan menggunakan metode *random over sampling* (ROS) dengan cara menambahkan data secara random kedalam kelas minoritas sehingga data menjadi seimbang. Random forest 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. Decision tree 5.0 merupakan metode klasifikasi berbentuk pohon keputusan dengan hasil akhir berupa aturan (rule). *Random over sampling* (ROS) menambahkan data secara random pada kelas minoritas yakni kelas 1 (stroke) menjadi 4700 responden pengamatan (sampel) sehingga total responden pengamatan menjadi 9400 sampel. *Random over sampling* pada klasifikasi random forest menghasilkan akurasi sebesar 99,47%, sensitifitas 100%, spesifisitas 98,96%, presisi 98,92%, AUC 99,48% dan akurasi dengan 10-fold cross validation sebesar 79,4%. *Random over sampling* pada klasifikasi decision tree algoritma C5.0 menghasilkan akurasi sebesar 97,45%, sensitifitas 100%, spesifisitas 95,29%, presisi 94,72%, AUC 97,6%. *Random over sampling* pada klasifikasi random forest menghasilkan model klasifikasi yang sangat baik, random over sampling pada klasifikasi decision tree algoritma C5.0 menghasilkan model klasifikasi yang sangat baik sedangkan klasifikasi random forest dan klasifikasi decision tree algoritma C5.0 pada imbalanced data menghasilkan model klasifikasi yang gagal.

Kata Kunci: **Machine Learning, Klasifikasi, Random Over Sampling (ROS), Random Forest, Decision Tree C5.0**

ABSTRACT

RANDOM OVER SAMPLING ANALYSES USING RANDOM FOREST CLASSIFICATION AND C5.0 WITH IMBALANCED DATA

By

MARLEN WATTIMENA
20/466528/PPA/06094

Classification analysis is a machine learning study that aims to find a decision function that accurately predicts the class of testing data that comes from the same distribution function as the data for training. Classification is the process of grouping objects into a predetermined class or category. Data imbalance is one of the problems in classification which is an unbalanced data condition between one data class and another data class. The data imbalance can be solved by using the random over sampling (ROS) method by adding random data into the minority class so that the data becomes balanced. Random forest is a classification method consisting of a combination of mutually independent classification trees. The classification prediction is obtained through a voting process (the highest number) of the classification trees formed. Decision tree 5.0 is a classification method in the form of a decision tree with the final result in the form of a rule. Random over sampling (ROS) added data randomly in the minority class, namely class 1 (stroke) to 4700 observation respondents (samples) so that the total observation respondents became 9400 samples. Random over sampling in the random forest classification resulted in an accuracy of 99.47%, sensitivity 100%, specificity 98.96%, precision 98.92%, AUC 99.48% and accuracy with 10-fold cross validation of 79.4%. Random over sampling in the decision tree classification algorithm C5.0 resulted in an accuracy of 97.45%, sensitivity 100%, specificity 95.29%, precision 94.72%, AUC 97.6%. Random over sampling on the random forest classification resulted in a very good classification model, random over sampling on the decision tree classification algorithm C5.0 produced a very good classification model, while the random forest classification and decision tree classification algorithm C5.0 on imbalanced data resulted in a failed classification model.

Keywords: **Machine Learning, Classification, Random Over Sampling (ROS), Random Forest, Decision Tree C5.0**