IMPLEMENTASI KOMBINASI METODE *RESAMPLING* PADA KLASIFIKASI PENYAKIT STROKE DENGAN ALGORITMA *K-NEAREST NEIGHBOR* DAN SELEKSI FITUR *INFORMATION GAIN*

MUHAMMAD FATHURRAHMAN

ABSTRAK

Salah satu masalah utama dalam dunia medis adalah stroke. Stroke menjadi penyebab kematian kedua di dunia. Berdasarkan hasil Riset Kesehatan Dasar (Riskesdar) tahun 2018 prevalensi stroke di Indonesia sebanyak 713,783 orang yang menderita stroke setiap tahunnya. Namun, dalam mendiagnosa stroke diperlukan waktu yang cukup panjang. Mengingat tiap menit ada sel yang mati karena penyumbatan aliran pada otak. Perlu dilakukan diagnosis sedini mungkin untuk mengurangi resiko dari pasien. *Data mining* dapat digunakan sebagai prediksi penyakit. Dalam pembuatan model *data mining, imbalanced data* merupakan masalah karena dapat berdampak buruk pada hasil klasifikasi dimana model *machine learning* akan lebih memperhatikan kelas mayoritas dan mengabaikan kelas minoritas. Pada penelitian telah dilakukan prediksi penyakit stroke menggunakan algoritma K-Nearest Neighbor dengan mengkombinasi teknik *resampling* seperti SMOTE, *Tomek Links* dan ENN. Serta penelitian dilakukan untuk mengetahui pengaruh seleksi fitur information gain terhadap model. Melalui proses *10 fold cross validation* diketahui model *machine learning* K-NN dengan SMOTE dan *Tomek Links* mampu memprediksi stroke dengan akurasi 83,5%, *f1-score* 12,5%, dan recall 24,7%. Kemudian untuk K-NN dengan SMOTE dan ENN diperoleh akurasi 78%, *f1-score* 16,8%, dan *recall* 45%. Ketika dilakukan seleksi fitur *information gain* terdapat peningkatan performa pada kedua metode tersebut. SMOTE dan *Tomek Links* menghasilkan akurasi 79,9%, *f1-score* 18,3%, dan recall 46,6% serta kombinasi SMOTE dan ENN diperoleh akurasi 76%, *f1-score* 20%, dan *recall* 59%. Setelah dilakukan pengujian diketahui bahwa teknik *resampling* dapat meningkatkan performa model pada kasus data yang tidak seimbang dari nilai *recall* dan *f1-score* sebesar 54% dan 7%.

Kata kunci : *Synthetic Minority Over-sampling*, *K-Nearest-Neighbor*, Stroke, *Tomek Links*, *Edited Nearest Neighbor*, *Information gain*.

IMPLEMENTATION OF A COMBINATION OF RESAMPLING METHODS IN STROKE CLASSIFICATION USING K-NEAREST NEIGHBOR ALGORITHM AND INFORMATION GAIN FEATURE SELECTION

MUHAMMAD FATHURRAHMAN

ABSTRACT

One of the main problems in the medical world is stroke. Stroke is the second cause of death in the world. Based on the results of Basic Health Research (Riskesdar) in 2018, the prevalence of stroke in Indonesia is 713,783 people who suffer from stroke every year. However, diagnosing a stroke takes quite a long time. Considering that every minute there are cells that die due to blockage of flow in the brain. Data mining can be used as a prediction of disease. In making data mining models, data imbalance is a problem because it can have a negative impact on the classification results where the machine learning model will pay more attention to the majority class and ignore the minority class. In this study, stroke prediction was carried out using the K-Nearest Neighbor algorithm by combining resampling techniques such as SMOTE, Tomek Links, and ENN. As well as research conducted to determine the effect of the search feature information obtained on the model. Through a 10 fold cross validation process, it is known that the K-NN machine learning model with SMOTE and Tomek Links is able to predict stroke with an accuracy of 83.5%, an f1-score of 12.5%, and a recall of 24.7%. Then K-NN with SMOTE and ENN obtained 78% accuracy, f1 score 16.8%, and recall 45%. When the selection of information gain features is carried out, there is an increase in performance in both methods. SMOTE and Tomek Links produce 79.9% accuracy, 18,3% f1-score, and 46,6% recall and the combination of SMOTE and ENN obtains 76% accuracy, 20% f1-score, and 59% recall. After the experiments, it is known that the resampling technique can improve the performance of the model in the case of imbalanced data from the recall and f1-score values by 54% and 7%.

Keyword*: Synthetic Minority Over-sampling, K-Nearest-Neighbor, Stroke, Tomek Links, Edited Nearest Neighbor, Information gain.*