ABSTRAK

Deteksi dini penyakit stroke memiliki peranan penting dalam mencegah dan mengurangi jumlah kematian. Meskipun demikian, data medis yang komprehensif dan seimbang mengenai stroke masih terbatas. Data medis mengenai stroke umumnya berukuran besar, memiliki banyak fitur, dan distribusi kelas yang tidak seimbang (*imbalanced*). Hal ini menjadi masalah karena algoritma klasifikasi cenderung dominan pada kelas mayoritas dan mengabaikan kelas minoritas, serta dimensi data yang tinggi dapat menyebabkan redundansi fitur. Oleh karena itu, data yang tidak seimbang dan berdimensi tinggi dapat menyebabkan kinerja algoritma pembelajaran mesin seperti *Naïve Bayes*, *K-Nearest Neighbors*, dan *Decision Tree* menjadi kurang optimal.

Untungnya, dalam bidang pembelajaran mesin terdapat metode penggabungan (*ensemble*) dengan mengintegrasikan metode SMOTE untuk menyeimbangkan data melalui pendekatan *over-sampling* dan metode PCA untuk mengurangi dimensionalitas data dengan mereduksi jumlah fitur. Sedangkan data yang digunakan adalah data sekunder dari repositori kaggle, khususnya dataset medis stroke. Data yang didapatkan kemudian dibagi menjadi data latih sebesar 80% dan data uji sebesar 20%. Selanjutnya dilakukan 2 skenario pemodelan dengan tahapan *hyperparameter tuning* dan 10-*fold cross validation*.

Hasil penelitian menunjukan peningkatan kinerja klasifikasi yang ditandai dengan meningkatnya nilai AUC secara berturut-turut sebesar 0,12 untuk *Naïve Bayes*, 0,08 untuk KNN dan *Decision Tree*, serta peningkatan yang signifikan pada nilai G-Mean sebesar 0,22 untuk *Naïve Bayes*, 0.35 untuk KNN, dan 0,26 untuk *Decision Tree*. Meskipun *ensemble* SMOTE dan PCA memberikan peningkatan kinerja yang baik pada nilai AUC dan G-Mean, namun terjadi penurunan pada nilai akurasi sebesar 0,14 pada *Naïve Bayes*, 0,12 pada KNN dan 0,10 pada *Decision Tree*. Penurunan ini terjadi karena SMOTE mensintesis sampel baru pada kelas minoritas agar seimbang dengan kelas mayoritas. Hal ini tentu mengakibatkan peningkatan ukuran data dan membebani kinerja algoritma klasifikasi. Di sinilah peran PCA dibutuhkan untuk mengurangi dimensi fitur dari 10 fitur menjadi 8 fitur dengan tetap mempertahankan 95% kandungan informasi, sehingga menghasilkan data yang lebih kecil dan sederhana.

Kata Kunci : Stroke, *Imbalanced*, Dimensionalitas, SMOTE, PCA, *hyperparameter tuning*, 10-*fold cross validation*, *Naïve Bayes*, *K-Nearest Neighbors*, *Decision Tree*

ABSTRACT

Early detection of stroke plays a crucial role in preventing and reducing the number of deaths. However, comprehensive and balanced medical data regarding stroke are still limited. Medical data related to stroke are generally large in size, have numerous features, and exhibit imbalanced class distribution. This poses a problem as classification algorithms tend to be biased towards the majority class and overlook the minority class, and high-dimensional data can lead to feature redundancy. Consequently, imbalanced and high-dimensional data can result in suboptimal performance of machine learning algorithms such as Naïve Bayes, K-Nearest Neighbors, and Decision Tree.

Fortunately, in the field of machine learning, there are ensemble methods that integrate SMOTE for data balancing through oversampling and PCA for dimensionality reduction by reducing the number of features. The data used in this study is secondary data obtained from the Kaggle repository, specifically a medical stroke dataset. The data is then divided into 80% training data and 20% testing data. Two modeling scenarios were performed, including hyperparameter tuning and 10-fold cross-validation.

The results of the research demonstrate improved classification performance, indicated by an increasing AUC value of 0.12 for Naïve Bayes, 0.08 for KNN and Decision Tree, and a significant increase in G-Mean value of 0.22 for Naïve Bayes, 0.35 for KNN, and 0.26 for Decision Tree. Although ensemble SMOTE and PCA provide good performance in terms of AUC and G-Mean, there is a decrease in accuracy of 0.14 for Naïve Bayes, 0.12 for KNN, and 0.10 for Decision Tree. This decline occurs because SMOTE synthesizes new samples for the minority class to balance it with the majority class, which leads to an increase in data size and burdens the performance of the classification algorithm. This is where PCA plays a role in reducing the feature dimension from 10 features to 8 features while retaining 95% of the information content, resulting in smaller and simpler data.

Keywords **:** *Stroke, Imbalanced, Dimensionality, SMOTE,* PCA, *hyperparameter tuning,* 10-*fold cross validation, Naïve Bayes, K-Nearest Neighbors, Decision Tree*