

ABSTRAK

Stroke merupakan salah satu penyakit mematikan yang menjadi penyebab utama kematian dan kecacatan di seluruh dunia. Penelitian ini bertujuan untuk membandingkan tingkat akurasi algoritma *Naive Bayes* dan *K-Nearest Neighbor* (KNN) dalam mengklasifikasikan data penyakit stroke. Dataset yang digunakan diambil dari situs Kaggle dan telah mengalami proses *preprocessing* serta penyeimbangan data menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE) guna mengatasi masalah *imbalanced data*. Penelitian ini juga menerapkan teknik validasi silang (*K-Fold Cross Validation*) untuk menguji performa model.

Hasil penelitian menunjukkan bahwa algoritma *K-Nearest Neighbor* dengan kombinasi k tetangga terdekat (k) dan *k-fold crossvalidation* menunjukkan bahwa kombinasi $k = 2$ dan *k-fold* = 3 menghasilkan performa terbaik dengan mencapai akurasi rata-rata tertinggi sebesar 95.59% dengan evaluasi model sebesar 87%.

Penelitian ini diharapkan dapat memberikan wawasan baru tentang penggunaan algoritma pembelajaran mesin dalam proses mencari akurasi dari dataset stroke, dengan tujuan meningkatkan efisiensi dan akurasi dalam menentukan dataset stroke.

Kata kunci: Stroke, Naive Bayes, K-Nearest Neighbor, SMOTE, K-Fold Cross Validation, klasifikasi.

ABSTRACT

Stroke is one of the deadliest diseases and a leading cause of death and disability worldwide. This study aims to compare the accuracy levels of the Naive Bayes and K-Nearest Neighbor (KNN) algorithms in classifying stroke disease data. The dataset used was obtained from the Kaggle website and has undergone preprocessing and data balancing using the Synthetic Minority Over-sampling Technique (SMOTE) to address the issue of imbalanced data. This study also applies K-Fold Cross Validation to evaluate the performance of the models.

The results show that the K-Nearest Neighbor algorithm, with a combination of the number of nearest neighbors (k) and k -fold cross-validation, achieved the best performance when $k = 2$ and $k\text{-fold} = 3$, yielding the highest average accuracy of 95.59% and a model evaluation score of 87%.

This study is expected to provide new insights into the use of machine learning algorithms for accuracy assessment of stroke datasets, aiming to enhance efficiency and accuracy in stroke data classification.

Keywords: Stroke, Naive Bayes, K-Nearest Neighbor, SMOTE, K-Fold Cross Validation, classification.