

KLASIFIKASI RISIKO STROKE MENGGUNAKAN LR DENGAN PENDEKATAN SMOTE-TOMEK LINKS DAN PENERAPAN XAI SHAP

Ekonovian Nababan, 202231072
Dibimbing Oleh Yozika Arvio, ST., M.Kom

ABSTRAK

Stroke merupakan penyebab utama kecacatan dan kematian di Indonesia dengan prevalensi 8,3 per mil pada penduduk usia 15 tahun ke atas dan pembiayaan kesehatan nasional mencapai Rp5,2 triliun pada tahun 2023. Penerapan *machine learning* untuk klasifikasi risiko stroke menghadapi dua tantangan utama, yaitu ketidakseimbangan data yang ekstrem dengan rasio 1:19 antara kelas stroke dan tidak stroke, serta kurangnya transparansi model dalam menjelaskan hasil prediksinya. Penelitian ini bertujuan untuk mengevaluasi kinerja model *Logistic Regression* dalam mengklasifikasikan risiko stroke menggunakan dataset hasil penyeimbangan data dengan metode SMOTE dibandingkan SMOTE-Tomek Links berdasarkan metrik recall, serta menginterpretasikan hasil klasifikasi menggunakan metode SHapley Additive exPlanations (SHAP). Dataset yang digunakan adalah *Brain Stroke Dataset* dari Kaggle yang terdiri dari 4.981 data dan 11 fitur. Penanganan ketidakseimbangan data dilakukan dengan membandingkan SMOTE dan SMOTE-Tomek Links yang diterapkan secara khusus pada data latih, diikuti pelatihan model *Logistic Regression* dan interpretasi model menggunakan SHAP. Hasil penelitian menunjukkan bahwa model *Logistic Regression* yang dilatih menggunakan dataset hasil SMOTE-Tomek Links mencapai *recall* kelas stroke sebesar 0,80 pada data uji, meningkat signifikan dari 0,00 pada skenario tanpa penanganan, dengan SMOTE-Tomek Links menghasilkan data latih yang lebih bersih melalui penghapusan 20 sampel ambigu pada batas keputusan antar kelas dibandingkan SMOTE murni. Interpretasi SHAP mengidentifikasi usia sebagai fitur paling dominan, diikuti status tidak pernah merokok, yang keduanya konsisten dengan pedoman klinis mengenai faktor risiko stroke.

Kata Kunci: stroke, klasifikasi risiko, *logistic regression*, SMOTE-Tomek Links, SHAP

STROKE RISK CLASSIFICATION USING LR WITH SMOTE-TOMEK LINKS APPROACH AND XAI SHAP APPLICATION

Ekonovian Nababan, 202231072
Supervised by Yozika Arvio, ST., M.Kom

ABSTRACT

Stroke is a leading cause of disability and mortality in Indonesia, with a prevalence of 8.3 per thousand among the population aged 15 years and above and a national healthcare expenditure of IDR 5.2 trillion in 2023. The application of machine learning for stroke risk classification faces two primary challenges: extreme class imbalance with a 1:19 ratio between stroke and non-stroke cases, and insufficient model transparency in explaining prediction outcomes. This study aims to evaluate the performance of a Logistic Regression model in classifying stroke risk using datasets balanced by SMOTE compared to SMOTE-Tomek Links based on recall, and to interpret classification results using the SHapley Additive exPlanations (SHAP) method. The dataset used is the Brain Stroke Dataset from Kaggle comprising 4,981 records and 11 features. Imbalanced data handling was performed by comparing SMOTE and SMOTE-Tomek Links applied specifically to the training data, followed by Logistic Regression model training and model interpretation using SHAP. Results indicate that the Logistic Regression model trained on the SMOTE-Tomek Links dataset achieved a stroke class recall of 0.80 on the test data, a significant improvement from 0.00 in the baseline scenario without imbalance handling, with SMOTE-Tomek Links producing cleaner training data by removing 20 ambiguous samples at the decision boundary compared to SMOTE. SHAP interpretation identified age as the most dominant feature, followed by never-smoked status, both consistent with clinical guidelines on stroke risk factors.

Keywords: stroke, risk classification, logistic regression, SMOTE-Tomek Links, SHAP