

ABSTRAK

Stroke merupakan gangguan kesehatan serius yang terjadi karena pembuluh darah di otak pecah atau terhambatnya aliran darah dan nutrisi ke jaringan otak. Penelitian ini bertujuan untuk membandingkan performa tiga algoritma pengajaran mesin, XGBoost, Random Forest, dan Logistic Regression, digunakan untuk mengkategorikan penyakit stroke berdasarkan berbagai faktor risiko. Dataset yang digunakan diambil dari Kaggle dan terdiri dari 5.110 data pasien dengan 12 atribut, antara lain Age, gender, hypertension, heart disease, ever married, work type, residence type, average glucose level, body mass index, dan smoking status. Menghadapi ketidakseimbangan data, di mana hanya terdapat 249 kasus stroke dari total 5.110 sampel, penelitian ini mengimplementasikan Teknik Over-sampling Minoritas Sintetis (SMOTE) untuk meningkatkan kualitas dan keseimbangan data dalam pelatihan model. Hasil pengujian menunjukkan bahwa algoritma XGBoost memiliki kinerja terbaik dengan akurasi 95% dan nilai presisi 0,96 untuk kelas non-stroke dan 0,95 untuk kelas stroke. kelas stroke, nilai recall masing-masing 0,95 dan 0,96, serta skor F1 sebesar 0,95 pada kedua kelas. Algoritma Random Forest menduduki peringkat kedua dengan akurasi 94%, sedangkan Logistic Regression menunjukkan performa terendah dengan akurasi 82%. Penelitian ini menunjukkan bahwa algoritma ensemble seperti XGBoost dan Random Forest unggul dalam klasifikasi penyakit stroke dibandingkan metode konvensional, serta menyoroti pentingnya faktor usia, kadar glukosa darah, dan riwayat penyakit jantung sebagai prediktor signifikan dalam diagnosis stroke.

Kata Kunci : Stroke; machine learning; xgboost; random forest; logistic regression