INTISARI

Penerapan hyperparameter tuning pada algoritma XGBoost untuk prediksi penyakit diabetes mellitus dilakukan untuk meningkatkan akurasi prediksi diabetes. Penelitian ini memanfaatkan dataset terbuka yang terdiri dari 70.000 baris data dan 22 variabel, dengan fokus pada sembilan variabel utama yakni GenHlth, HighBP, BMI, HighChol, Age, DiffWalk, Income, HeartDiseaseorAttack dan PhysHlth. Metode yang digunakan mencakup pembersihan data, seleski fitur penanganan outlier menggunakan metode winsorization, pembagian data latih dan data uji dengan stratified sampling, serta pengembangan model menggunakan algoritma XGBoost yang dioptimalkan dengan teknik Randomized Search CV. Hasil penelitian menunjukkan bahwa model baseline XGBoost tanpa tuning menghasilkan akurasi sebesar 73%, sementara model dengan hyperparameter tuning mencapai akurasi 74%. Analisis Confusion Matrix dan Classification Report menunjukkan peningkatan pada nilai precision, recall, dan f1-score, terutama pada kelas positif atau terdeteksi diabetes. Sebagai langkah lanjutan, model ini dideploy dalam aplikasi web sederhana menggunakan Flask untuk menguji fungsionalitas model. Meskipun demikian, aplikasi ini bersifat prototipe beta dan belum dirancang untuk penggunaan medis karena model masih memerlukan validasi lebih lanjut. Penelitian ini menunjukkan bahwa optimasi hyperparameter pada XGBoost memberikan hasil yang lebih baik dibandingkan model baseline. Model yang dihasilkan memiliki potensi untuk digunakan dalam prediksi awal diabetes, namun memerlukan pengujian tambahan sebelum diimplementasikan secara luas.

Kata kunci: prediksi diabetes, hyperparameter tuning, XGBoost, confusion matrix

ABSTRACT

Hyperparameter tuning is applied to the XGBoost algorithm to predict diabetes mellitus and increase the accuracy of diabetes predictions. This research utilizes an open dataset consisting of 70,000 rows of data and 22 variables, with a focus on nine main variables, namely GenHlth, HighBP, BMI, HighChol, Age, DiffWalk, Income, HeartDiseaseorAttack and PhysHlth. The methods used include data cleaning, selecting outlier handling features using the winsorization method, dividing training data and test data using stratified sampling, and developing a model using the XGBoost algorithm optimized with the Randomized Search CV technique. The research results show that the XGBoost baseline model without tuning produces an accuracy of 73%, while the model with hyperparameter tuning achieves an accuracy of 74%. Confusion Matrix and Classification Report analysis show increased precision, recall, and f1-score values, especially in the positive or detected diabetes class. As a further step, the model was deployed in a simple web application using Flask to test the model's functionality. However, this application is a beta prototype and has not been designed for medical use as the model still requires further validation. This research shows that hyperparameter optimization in XGBoost provides better results than the baseline model. The resulting model has the potential to be used in early diabetes prediction but requires additional testing before widespread implementation.

Keywords: diabetes prediction, hyperparameter tuning, XGBoost, confusion matrix