

KOMPARASI KINERJA ALGORITMA RANDOM FOREST DAN XGBOOST UNTUK KLASIFIKASI KASUS DIABETES

Oleh

Resa Astari Br Tarigan, NIM 2215091018

Jurusan Teknik Informatika

E-Mail: Resa.astari@student.undiksha.ac.id

ABSTRAK

Diabetes merupakan salah satu penyakit kronis yang prevalensinya terus meningkat dan berpotensi menimbulkan berbagai komplikasi serius apabila tidak terdeteksi secara dini. Tingginya jumlah penderita yang belum terdiagnosis menunjukkan pentingnya pemanfaatan teknologi untuk mendukung proses deteksi dan klasifikasi penyakit secara lebih cepat dan akurat. Salah satu pendekatan yang banyak digunakan adalah machine learning, yang mampu mengidentifikasi pola dari data kesehatan dan menghasilkan prediksi yang mendukung pengambilan keputusan di bidang medis. Penelitian ini bertujuan untuk membandingkan kinerja algoritma XGBoost dan Random Forest dalam klasifikasi Diabetes berdasarkan metrik accuracy, precision, recall, dan F1-score, serta menentukan algoritma dengan performa terbaik dalam mengklasifikasikan data pasien Diabetes. Dataset yang digunakan adalah Diabetes Dataset yang diperoleh dari Mendeley Data dengan jumlah 1.000 data dan 12 atribut. Penelitian ini menerapkan pendekatan Knowledge Discovery in Databases (KDD) yang meliputi tahapan pemilihan data, exploratory data analysis, preprocessing, data cleaning, encoding, feature filtering, pembagian data menjadi data latih dan data uji dengan rasio 80:20, serta pemodelan menggunakan algoritma Random Forest dan XGBoost. Untuk meningkatkan keandalan evaluasi, digunakan metode 5-Fold Cross Validation pada data latih, sedangkan data uji digunakan untuk mengukur performa akhir model terhadap data yang belum pernah diproses sebelumnya. Evaluasi dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa algoritma XGBoost memiliki performa yang lebih baik dibandingkan Random Forest dengan nilai accuracy sebesar 98,78%, precision sebesar 99,52%, recall sebesar 96,49%, dan F1-score sebesar 97,91%. Berdasarkan hasil tersebut, dapat disimpulkan bahwa XGBoost merupakan algoritma yang lebih efektif dalam klasifikasi kasus Diabetes pada dataset yang digunakan dalam penelitian ini.

Kata Kunci: Diabetes, Machine Learning, Klasifikasi, Random Forest, XGBoost.

PERFORMANCE COMPARISON OF RANDOM FOREST AND XGBOOST ALGORITHMS FOR DIABETES CLASSIFICATION

By

Resa Astari Br Tarigan, NIM 2215091018

Informatics Engineering Department

E-mail: Resa.astari@student.undiksha.ac.id

ABSTRACT

Diabetes is a chronic disease whose prevalence continues to increase and has the potential to cause various serious complications if not detected at an early stage. The high number of undiagnosed cases highlights the importance of utilizing technology to support disease detection and classification processes more quickly and accurately. One of the most widely used approaches is machine learning, which is capable of identifying patterns from healthcare data and generating predictions to support decision-making in the medical field. This study aims to compare the performance of the XGBoost and Random Forest algorithms in diabetes classification based on accuracy, precision, recall, and F1-score metrics, as well as to determine the algorithm with the best performance in classifying diabetes patient data. The dataset used in this study was the Diabetes Dataset obtained from Mendeley Data, consisting of 1,000 records and 12 attributes. This study adopted the Knowledge Discovery in Databases (KDD) approach, which includes data selection, exploratory data analysis, preprocessing, data cleaning, encoding, feature filtering, data splitting into training and testing sets with an 80:20 ratio, and model development using the Random Forest and XGBoost algorithms. To improve the reliability of the evaluation process, 5-Fold Cross Validation was applied to the training data, while the testing data were used to measure the final performance of the models on previously unseen data. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. The results showed that the XGBoost algorithm outperformed Random Forest, achieving an accuracy of 98.78%, precision of 99.52%, recall of 96.49%, and F1-score of 97.91%. Based on these findings, it can be concluded that XGBoost is a more effective algorithm for diabetes classification on the dataset used in this study.

Keywords: Diabetes, Machine Learning, Classification, Random Forest, XGBoost.