DEVELOPING A MACHINE LEARNING- BASED OF A DIABETES MELLITUS DETECTION APPS WITH UNBALANCED DATA HANDLING USING SMOTE AND OPTUNA HYPERPARAMETER OPTIMIZATION

YUWANIS FAZLINA AGUSTIA
Program Studi Informatika Medis, Fakultas Sains & Teknologi
Universitas Teknologi Yogyakarta
Email: yuwanisfazlinaagustia@gmail.com

ABSTRACT

The study developed a machine learning-based Diabetes Mellitus prediction system to support case identification using laboratory examination data. This study used Prolanis Diabetes Mellitus patient data, comprising 484 records: 313 Diabetes Mellitus patients and 171 non-Diabetes Mellitus patients, resulting in a class imbalance. The research stages included data preprocessing, including handling missing data, data normalization, and addressing data imbalance using SMOTE, ADASYN, and Tomek Links. Classification models were developed and compared using the Random Forest, XGBoost, and LightGBM algorithms, followed by optimization through hyperparameter tuning. Test results demonstrated that the Random Forest model combined with SMOTE and hyperparameter tuning achieved the best performance, with an accuracy of 71.89% on the training data and 72.60% on the test data; precision of 70.81% and 72.13%, respectively; recall of 96.35% and 93.62%; F1-score of 81.62% and 81.48%; and an AUC/ROC of 79.21 on the training data and 76.74 on the test data. This model also showed a relatively small difference in metric values between the training and test data, indicating good generalization and minimal overfitting. Therefore, the Random Forest model combined with SMOTE and hyperparameter tuning was implemented in a web application as a machine learning-based Diabetes Mellitus prediction system.

Keywords: Diabetes Mellitus, Machine Learning, Random Forest, SMOTE, Hyperparameter Tuning.