A Comprehensive Data-Driven Machine Learning Framework for Diabetes Prediction and Diagnosis

Suejit Pechprasarn*, Sasipatcha Hanmanop, Tatpol Jongsiri, Kittitat Waiprasit

College of Biomedical Engineering, Rangsit University, Pathum Thani 12000, Thailand
*Corresponding author; E-mail: suejit.p@rsu.ac.th

Abstract

Diabetes remains one of the most prevalent and rapidly growing global health challenges, necessitating more accurate and accessible diagnostic tools. This study presents an extensive evaluation of 24 machine learning models designed to predict and diagnose diabetes through the integration of eight critical predictors: sex, gender, history of heart disease, hypertension status, smoking habits, body mass index (BMI), Hemoglobin A1c (HbA1c) levels, and blood glucose concentrations. The dataset underwent rigorous preprocessing—comprising data curation and cleaning—to ensure unbiased and robust training. Additionally, the study investigated three distinct data partitioning strategies (70/30, 80/20, and 90/10 splits) to ascertain the optimal balance between training and evaluation sets. The predictive task involved binary classification, designating subjects as non-diabetic (0) or diabetic (1). Among the tested models, the Ensemble Boosted Trees algorithm demonstrated superior performance, achieving an accuracy of 91.45%, precision of 91.29%, recall of 91.65%, and an F1-score of 91.37% with the 70/30 split, thereby underscoring its efficacy. Furthermore, a dimensionality reduction approach, relying solely on age and BMI, yielded a respectable accuracy of 74.27%. This simplified model illustrates a cost-effective screening alternative, particularly valuable in resource-constrained settings where comprehensive blood testing is impractical. These findings highlight the potential of advanced, data-driven methodologies in enhancing early diabetes detection and diagnosis.

Keywords: *artificial intelligence, diabetes, dimensionality reduction, feature selection method, machine learning*