

Abstract

Diabetes mellitus remains a major global health challenge, requiring accurate and early prediction to reduce the risk of severe complications and improve patient outcomes. Although machine learning techniques have been widely applied to diabetes prediction, existing studies often address key challenges such as class imbalance, feature representation, and model optimization in isolation, leading to suboptimal and non-reproducible predictive performance. This study addresses this gap by proposing a comprehensive and integrated machine learning pipeline for diabetes prediction using the Pima Indians Diabetes Dataset. The proposed framework systematically combines data preprocessing, feature engineering, dimensionality reduction, class imbalance handling, and ensemble learning within a unified pipeline. Missing values are handled using median imputation, while outliers are treated through interquartile range (IQR)-based winsorization. Three domain-informed features Glucose_per_BMI, Age_BMI, and HighPreg are introduced to capture nonlinear relationships, followed by feature selection and Principal Component Analysis (PCA). To address class imbalance, SMOTE-Tomek resampling is applied, improving minority class representation. Multiple models, including Logistic Regression, Support Vector Machines, Random Forest, XGBoost, LightGBM, and a stacked ensemble, are evaluated using cross-validation and a hold-out test set. Experimental results demonstrate that the proposed pipeline significantly improves predictive performance, with accuracy increasing from 74.03% to 82.11%, F1-score from 53.33% to 71.11%, and balanced accuracy from 0.6893 to 0.7780. Notably, recall improved substantially across models, enhancing sensitivity to diabetic cases. Additionally, recall improved substantially, with Logistic Regression increasing from 0.6267 to 0.8000 and SVM from 0.5467 to 0.7467. LightGBM and the stacked ensemble achieved the best overall performance. These findings highlight the effectiveness of integrating preprocessing, feature engineering, and ensemble learning into a unified framework for robust and reliable diabetes prediction.