ABSTRACT

Diabetes is a chronic disease that often goes undetected in its early stages due to its common symptoms, posing a risk of serious complications. Early detection is key to preventing these outcomes. This research aims to develop a website-based early prediction system for diabetes using the Categorical Boosting (CatBoost) machine learning algorithm. This study also identifies the most influential symptoms and evaluates the performance of CatBoost against Random Forest, KNN, XGBoost, AdaBoost, LightGBM, and Logistic Regression algorithms. Using two datasets—primary data from healthcare workers in Kupang and Southeast Maluku, and a secondary dataset from the UCI repository—this research applied Chi-Square feature selection to determine the main predictors. The results indicate that Polydipsia, Polyuria, sudden weight loss, partial paresis, and Polyphagia are the five most dominant symptoms. The optimized CatBoost model achieved an accuracy of 96.77%, a precision of 97%, and an F1-score of 96.78%, outperforming the other comparison models. The developed website-based system was successfully validated using the Black Box Testing method, proving its functionality as an effective and accurate initial screening tool.

Keywords : *diabetes, early stage, catboost, machine learning, prediction*