ABSTRACT

Heart disease is a crucial global health issue, requiring accurate detection and prediction for optimal treatment. However, existing methods often fail to handle the complexity of medical data. To address this key issue, a robust and easy-to-understand machine learning model was developed. This model uses the UCI Heart Disease dataset, which has undergone a data preparation process, including handling missing data with KNN Imputer. This study introduces a hybrid XGBoost-Random Forest (XGB-RF) model that operates in two modes: stacking, where XGBoost predictions serve as additional input to Random Forest, and XGB-RF Feature Engineering, where XGBoost generates new features that then serve as input to Random Forest. Understanding the prediction results is carried out by SHapley Additive exPlanations (SHAP), thus identifying the causes of heart disease in each individual. With the model achieving an accuracy of 83.69% and ROC-AUC of 86.27%, demonstrating a significant improvement in the ability to predict heart disease risk. In conclusion, this XGB-RF model is effective in predicting heart disease risk and provides clear explanations through SHAP. This model has great potential to help doctors in faster and more accurate patient risk assessment in the future, and can even be developed for early monitoring through technology integration.

Keywords: *Heart Disease, XGBoost, Random Forest, Stacking, Feature Engineering*