ABSTRACT

Cardiovascular disease (CVD) is the leading cause of death worldwide and often develops without clear early symptoms. Early risk prediction is crucial to reduce mortality rates and support timely clinical intervention. This study aims to develop a heart disease risk prediction model using the Framingham Heart Study dataset by combining Adaptive Synthetic Sampling (ADASYN) and the Extreme Gradient Boosting (XGBoost) algorithm. ADASYN is employed to address class imbalance, while XGBoost is selected for its robustness in handling tabular data and its ability to produce accurate and efficient predictive models. The study explores six experimental scenarios with variations in the application of ADASYN, hyperparameter tuning, and feature selection. Model performance is evaluated using accuracy, recall, precision, F1-score, and area under the curve (AUC), with particular focus on performance in the minority class (at-risk patients). The best results were achieved in the scenario that utilized ADASYN, tuning, and feature selection, yielding a recall of 25%, an F1-score of 26%, and an AUC of 64.3%. This represents a substantial improvement over the baseline model, which only achieved a recall of 7% and an F1-score of 10%. Although the metrics remain relatively low in absolute terms, the improvement in sensitivity toward identifying high-risk patients is clinically meaningful. Therefore, the integration of ADASYN and XGBoost in this study demonstrates promising potential to support the development of data-driven early detection systems for cardiovascular risk. This approach offers a more balanced, sensitive, and practical solution for clinical decision support.

***Keywords**: heart disease prediction, ADASYN, XGBoost, imbalanced data, feature selection, framingham heart study, machine learning.*