

## ABSTRACT

Diabetes mellitus continues to rise globally, with 537 million people affected in 2021 and an estimated 643 million by 2030. The Pima Indians Diabetes Dataset (PIDDD) is often used in diabetes risk prediction studies due to its representation of high-risk populations, making it an important reference in the development of machine learning-based models. The application of machine learning in the healthcare sector is growing, particularly with the XGBoost algorithm for clinical prediction. However, the complexity of XGBoost poses challenges in interpretability, so the SHAP method is used to explain the contribution of variables. The XGBoost model achieved an accuracy of 83% on the test data. SHAP analysis identified Glucose, BMI, and Age as the most significant predictors for diabetes. The integration of XGBoost and SHAP resulted in a diabetes prediction model that is not only accurate but also interpretable. These findings highlight the potential of machine learning in facilitating early detection and improving clinical decision-making, particularly in identifying key risk factors for patients at risk of diabetes.

**Keywords:** Diabetes Mellitus, XGBoost, SHAP, Diabetes Risk Prediction