

Predictive Classification of Diabetes Mellitus in Indian Adults Using Machine Learning: Analysis of National Family Health Survey-5 Data

Dr. Prem Kumar Chandrakar

Mahant Laxminarayan Das College, Raipur, Chhattisgarh, India

Mridula Patel

M.Sc. Computer Science

Pandit Sundarlal Sharma (Open) University, Bilaspur, Chhattisgarh, India

Vivek Prakash Sahu

Mahant Laxminarayan Das College, Raipur, Chhattisgarh, India

premchandrakar@gmail.com, mridulapatel5@gmail.com, viveknetcom@gmail.com

Abstract: *Background: Diabetes mellitus represents a critical public health emergency in India, with projections indicating 134 million affected adults by 2045 (Pradeepa et al., 2020). The integration of machine learning with nationally representative survey data offers promising approaches for population-level risk stratification in resource-constrained settings.*

Methods: Utilizing the National Family Health Survey-5 (NFHS-5) dataset comprising 724,115 women and 101,839 men, this study implemented six machine learning algorithms. Comprehensive preprocessing addressed complex survey design and missing data through multiple imputation techniques (Van Buuren, 2018).

Results: Ensemble methods demonstrated superior performance, with Random Forest achieving AUC-ROC 0.891 (95% CI: 0.884-0.898) and XGBoost 0.874 (95% CI: 0.866-0.882). The models identified age (22.3%), BMI (18.7%), and waist-to-hip ratio (15.2%) as primary predictors, consistent with known pathophysiological mechanisms while revealing novel socioeconomic determinants.

Conclusion: Machine learning algorithms effectively predict diabetes risk using nationally representative data, potentially enabling cost-effective screening strategies. Implementation research is needed to translate these findings into public health practice..

Keywords: Diabetes Prediction, Machine Learning, Ensemble Methods, NFHS-5, India, Public Health Informatics

