World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:19, No:04, 2025

Machine Learning Prediction of Diabetes Prevalence in the U.S. Using Demographic, Physical, and Lifestyle Indicators: A Study Based on NHANES 2009-2018

Authors : Oluwafunmibi Omotayo Fasanya, Augustine Kena Adjei

Abstract : To develop a machine learning model to predict diabetes (DM) prevalence in the U.S. population using demographic characteristics, physical indicators, and lifestyle habits, and to analyze how these factors contribute to the likelihood of diabetes. We analyzed data from 23,546 participants aged 20 and older, who were non-pregnant, from the 2009-2018 National Health and Nutrition Examination Survey (NHANES). The dataset included key demographic (age, sex, ethnicity), physical (BMI, leg length, total cholesterol [TCHOL], fasting plasma glucose), and lifestyle indicators (smoking habits). A weighted sample was used to account for NHANES survey design features such as stratification and clustering. A classification machine learning model was trained to predict diabetes status. The target variable was binary (diabetes or non-diabetes) based on fasting plasma glucose measurements. The following models were evaluated: Logistic Regression (baseline), Random Forest Classifier, Gradient Boosting Machine (GBM), Support Vector Machine (SVM). Model performance was assessed using accuracy, F1-score, AUC-ROC, and precision-recall metrics. Feature importance was analyzed using SHAP values to interpret the contributions of variables such as age, BMI, ethnicity, and smoking status. The Gradient Boosting Machine (GBM) model outperformed other classifiers with an AUC-ROC score of 0.85. Feature importance analysis revealed the following key predictors: Age: The most significant predictor, with diabetes prevalence increasing with age, peaking around the 60s for males and 70s for females. BMI: Higher BMI was strongly associated with a higher risk of diabetes. Ethnicity: Black participants had the highest predicted prevalence of diabetes (14.6%), followed by Mexican-Americans (13.5%) and Whites (10.6%). TCHOL: Diabetics had lower total cholesterol levels, particularly among White participants (mean decline of 23.6 mg/dL). Smoking: Smoking showed a slight increase in diabetes risk among Whites (0.2%) but had a limited effect in other ethnic groups. Using machine learning models, we identified key demographic, physical, and lifestyle predictors of diabetes in the U.S. population. The results confirm that diabetes prevalence varies significantly across age, BMI, and ethnic groups, with lifestyle factors such as smoking contributing differently by ethnicity. These findings provide a basis for more targeted public health interventions and resource allocation for diabetes management.

Keywords : diabetes, NHANES, random forest, gradient boosting machine, support vector machine
Conference Title : ICDM 2025 : International Conference on Diabetes and Metabolism
Conference Location : Washington, Australia
Conference Dates : April 24-25, 2025