Project Report
Predicting Diabetes Risk Using Machine Learning

Objective

The primary objective of this project is to develop a predictive model using machine-learning techniques to assess the likelihood of individuals developing diabetes. By leveraging patient data, the model aims to provide healthcare professionals with valuable insights, allowing for early interventions and personalized health recommendations. This initiative contributes to improving patient outcomes and addressing the growing prevalence of diabetes globally.

Source of Data

The dataset utilized in this project is derived from publicly available health databases. It includes various features related to patient demographics and health metrics, which are critical for predicting diabetes risk. Key attributes in the dataset include:

Age: The age of the patient.

BMI (Body Mass Index): A measure of body fat based on height and weight.

Blood Pressure: The patient's blood pressure reading.

Smoking Status: A binary indicator of whether the patient is a smoker.

Exercise Frequency: The frequency of exercise in a week.

Diabetes: The target variable indicating whether the patient has diabetes (1) or not (0).

Focus on Sustainable Development Goals (SDG)

This project aligns with the United Nations Sustainable Development Goal (SDG) 3: Good Health and Wellbeing. By developing predictive analytics for diabetes risk, we aim to:

Reduce Mortality: Enable early detection and management of diabetes, ultimately reducing associated health complications and mortality rates.

Promote Health: Provide data driven insights to promote healthier lifestyles through personalized recommendations.

Strengthen Health Systems: Support healthcare systems by integrating technology that enhances decision-making and patient management.

Methodology

The methodology for this project is structured into several key steps:

Data Pre-processing

Data Loading

The dataset is loaded into a pandas Data-Frame.

Feature Selection

Relevant features are selected for modelling:

Any missing values are addressed (if present) to ensure data integrity.

Feature Scaling

Standardization of features using `StandardScaler` is performed to normalize the range of independent variables:

Model Development

Data Splitting

The dataset is divided into training and testing sets using an 8020 split.

Model Selection

Logistic Regression is chosen for its effectiveness in binary classification problems:

Model Training

The model is trained using the training dataset.

Prediction and Evaluation

Predictions are made on the test dataset, and model performance is evaluated using metrics such as classification report and ROC AUC score.

ROC Curve Visualization

The ROC curve is plotted to visualize the tradeoff between sensitivity and specificity.

User Interaction

An interactive function is implemented to allow users to input their health metrics and receive personalized predictions and recommendations regarding diabetes risk. The user inputs are processed and scaled before being passed to the model for prediction.

Conclusion

This project demonstrates the application of machine learning techniques to predict diabetes risk based on health metrics. By focusing on data-driven insights, we can empower healthcare providers to make informed decisions, ultimately improving patient outcomes. Future work may involve expanding the dataset, incorporating more features, and exploring advanced modelling techniques to enhance prediction accuracy.