# Predicting Risk of Type 2 Diabetes: Stay Healthy, Stay Ahead

## Problem Statement

Health issues like heart disease, diabetes, and obesity are on the rise, and understanding how lifestyle choices impact these conditions is more important than ever. Early prediction of health risks can help people make better choices and lead healthier lives.

### 1. Data Cleaning: Missing Values and Normalization of Life-related Variables

- Lifestyle variables such as exercise frequency and diet should be normalized into one common system of measurement (e.g., exercise in hours per week).

### 2. EDA: Exploring Relationships Between Lifestyle Factors and Health Outcomes

- EDA involves seeking correlations between lifestyle factors and health outcomes such as type 2 diabetes.
- Visualizations and correlation tests help uncover associations between poor diet or lack of exercise with increased risk.
- EDA focuses on identifying the major lifestyle factors for model development.

### 3. Feature Engineering: Aggregate Health Indices, Lifestyle Scores, Demographic Pro

- Feature engineering combines lifestyle factors into composite scores like a 'Health Score' based on diet, exercise, and smoking status.
- A demographic profile, including age and family history, can be added to enhance diabetes risk.

### 4. Modeling: Logistic Regression, k-Nearest Neighbors (kNN), Tree-Based Models

- Logistic regression and kNN classification use lifestyle factors to predict the likelihood of type 2 diabetes.
- Logistic regression provides probability, while kNN classifies individuals based on similarity with others.
- Tree-based models, such as random forests, identify the most influential factors on diabetes risk.

### 5. Evaluation: Use of Precision, Recall, and ROC-AUC as Performance Metrics

- Evaluation metrics such as precision, recall, and ROC-AUC are used to assess model performance.
- Precision and recall help measure the accuracy of positive predictions, while ROC-AUC demonstrates the model's ability to distinguish between at-risk and not-at-risk individuals.

### 6. Insights: Identifying Major Lifestyle Factors and Suggesting Preventive Measures

- Major risks to diabetes include low physical activity levels and poor dietary habits.
- After modeling, recommendations can be made for promoting exercise and healthier eating.
- Public health interventions can focus on these factors to reduce the risk of type 2 diabetes and heart disease.
- 80% of cardiovascular disease is preventable with a healthy diet, physical activity, in taking less tobacco and alcohol

**Race strategy (metrology): How are you going to use Python and data to solve your problem? What tools and techniques will you use?**

Python and data can be used to improve individual health by collecting and analyzing data from various sources (e.g., fitness trackers, medical records). Predictive models assess conditions like diabetes or heart disease, and personalized recommendations can be made based on an individual's data. Privacy and security are ensured using encryption, and continuous monitoring systems can provide real-time feedback and reminders. The goal is to deliver actionable insights, personalized health plans, and early warnings for health risks. We can also use data visualization tools like bar charts to describe correlation between ages and diseases, regression, ROC-AUC and kNN used like diagrams of the graphs.

**Dataset:**

- 2011-2015 Dataset: Behavioral Risk Factor Surveillance System (BRFSS) Dataset (2011-2015)
- 2023 Dataset: 2023 Dataset
- The Original BRFSS Website: The Website