GitHub Read me:

Project title: Diabetes risk prediction using Machine learning

Objective / Problem statement: Diabetes is recognized as a chronic condition in which insufficient insulin is produced by the body, leading to elevated blood sugar levels. Various factors, including genetics, lifestyle, age, and socioeconomic status, influence the development of diabetes. While type 1 diabetes cannot be prevented, type 2 diabetes and prediabetes can often be delayed or prevented through measures such as weight loss, regular exercise, and a balanced diet. The main objective of this project is to build a machine learning model to accurately predict an individual's risk of developing diabetes based on demographics, lifestyle, key health indicators, and socioeconomic factors.

Dataset source(s): For this project, I will be using the CDC Diabetes Health Indicators Dataset (Balanced), which I obtained from Kaggle. The specific file I will use is:

diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv

The dataset has more than 70,000 rows (balanced: 50% diabetic and 50% non-diabetic). The target variable is “Diabetes\_binary” (0 = No, 1 = Yes). There are 3 numerical variables: BMI, MentHlth, PhysHlth and 18 categorical variables: Age, HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, DiffWalk, Sex, Education, Income.

Predictors: Various health indicators (such as BMI, physical activity, high blood pressure, and cholesterol) along with demographic variables (including age, sex, education, and income).