The **Diabetes prediction dataset** is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. This dataset can be used to build machine learning models to predict diabetes in patients based on their medical history and demographic information. This can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans. Additionally, the dataset can be used by researchers to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes.

**Research proposal introduction**

In this project we will be predicting that whether the patient has diabetes or not.

**Research question**

* What is the relationship between various medical and demographic factors in relation to the likelihood of developing diabetes?

**Machine learning and methodology to be used**

Firstly in our project we will conduct data visualization using histogram and box plots. Correlation matrix will be used to visualize the correlation between variables. Simple linear model and multiple linear model will be made. A logistic regression algorithm will be use to make the predictions and check for accuracy which can be random forest model.

**Data Description**

Gender :Gender refers to the biological sex of the individual, which can have an impact on their susceptibility to diabetes.

Age: Age is an important factor as diabetes is more commonly diagnosed in older adults. Age ranges from 0-80 in our dataset.

Hypertension: hypertension is a medical condition in which the blood pressure in the arteries is persistently elevated. It has values a 0 or 1 where 0 indicates they don’t have hypertension and for 1 it means they have hypertension.

 Heart disease: heart disease is another medical condition that is associated with an increased risk of developing diabetes. It has values a 0 or 1 where 0 indicates they don’t have heart disease and for 1 it means they have heart disease.

Smoking history is also considered a risk factor for diabetes and can exacerbate the complications associated with diabetes. In our dataset we have 5 categories i.e not current,former,No Info,current,never and ever.

 BMI(Body Mass Index) is a measure of body fat based on weight and height. Higher BMI values are linked to a higher risk of diabetes. The range of BMI in the dataset is from 10.16 to 71.55. BMI less than 18.5 is underweight, 18.5-24.9 is normal, 25-29.9 is overweight, and 30 or more is obese.

HbA1c \_level: HbA1c(Hemoglobin A1c) level is a measure of a person's average blood sugar level over the past 2-3 months. Higher levels indicate a greater risk of developing diabetes. Mostly more than 6.5% of HbA1c Level indicates diabetes

Blood glucose level refers to the amount of glucose in the bloodstream at a given time. High blood glucose levels are a key indicator of diabetes.

Diabetes is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes.

**Data source**

https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset