Summary :

The Dataset

[www.kaggle.com](http://www.kaggle.com)

OUR DATA SOURCE - KAGGLE

OUR OBJECTIVE

**In this project, the objective is to predict a probability of being diabetic or non-diabetic based on various features like blood glucose levels, HBA1C levels, BMI etc.   
The project uses a diabetes prediction dataset from Kaggle.   
The code is divided in 9 parts that starts from data gathering to model deployment to testing the prediction hypothesis with test data via a frontend web page.  
This model can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans.**

**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, BMI, Heart disease, Smoking history, Hyper-Tension, HBA1C level, Blood Clucose Level**

OUR DATA SOURCE - KAGGLE

[**https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset**](https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset)

**Data Exploration**

**Using tools such as Spark, and Python, we explored the data to find additional details :  
- Count of the number of rows  
- Study of the schema  
- Count of total number of diabetic and non-diabetic values.   
 (*Note: 1 - Presence of diabetes, 0 - Absence of diabetes)*- Count of total number of gender types  
- Check for empty values in the gender column  
- Get summary statistics to describe the data frames count, mean, stddev, min and max**

**DATA CLEANING**

**Cleaned and Prepared data   
- Check for null values  
- Get rid of the "Other" value in the *gender* column to limit it to Male and Female values  
- Remove the "No Info" value from the *smoking\_history* column as that will not be useful data for our ML model  
- Assign 0 and 1 values in *gender* column where 0 means "Female" and 1 means "Male"  
- Assign numeric values to the *smoking\_history* column where "never" = 0, "ever" = 1, "not current" = 2, "current" = 3, "former" = 4**