**Diabetes Datathon**

**Change is everywhere**

**Table of Contents**

1. **Problem Statement**
2. **Part A: Setting up GitHub environment**
3. **Part B: Data Extraction, Cleaning, and Initial Exploratory Data Analysis (EDA)**
4. **Part C: Loading Data into a SQL Database**
5. **Part D: Performing Data Analysis with SQL**
6. **Part E: Build Visualizations in Power BI**

### **Problem Statement:**

Using Python, SQL & Power BI: perform exploratory data analysis, and visualisation, of Diabetes data. You have between 11th Nov – 22nd Nov to solutionize.

**Task:**

Diabetes is a complex, chronic endocrine disorder, affecting millions of British people per year. November is Diabetes Awareness month, and your client - a drug manufacturer, have just secured funding to break ground in the critical disease segment: Diabetes. So, in preparation to launch the novel product, and in honour of Diabetes awareness month (November), and they want you to analyse the impact of the disease on the human body.

They have hired you, a team of Data Enthusiasts, to spend two weeks performing data analysis and visualisation, to inform their approach.

There are some deliverables to win the best solution, and an overall winning team will also be announced. Feel free to innovate along the way. The more the merrier!

**Deliverables:**• A detailed report outlining the data cleaning and preprocessing steps taken, with justifications for each decision.• An initial EDA report that includes key insights and visualizations (e.g., histograms, scatter plots, heatmaps) generated using Python libraries like Matplotlib or Seaborn.   
*•* A clean, well-prepared dataset stored as a new CSV file or as Data frame (Pandas, Pyspark).• A brief report explaining schema design with implemented SQLs.• A visualisation report that translates the data analysis into actionable insights for stakeholders.

### **Part A:** **Setting up GitHub environment**

**In the first part of the challenge, participants will set up Git Hub, in preparation to commit their work to GitHub.**

#### **GitHub Task:**

1. Create a project folder.
2. Initialize the Git repository.
3. Configure the repo for remote collaboration.
4. Create a main branch.

### **Part B: Data Extraction, Cleaning, and Initial Exploratory Data Analysis (EDA)** **Next, participants are required to begin by extracting, cleaning and performing exploratory data analysis.**

1. Extract the Diabetes dataset from a provided CSV file using any preferred language (for e.g. Python, R, Pyspark, Java).
2. Clean and preprocess the data:
   1. Handle missing values, outliers, and inconsistencies.
   2. Encode categorical variables and normalize/scale numerical features where appropriate.
3. Following the cleaning process, participants should perform an initial EDA to uncover key patterns and insights in the data, such as distributions and correlations.
4. Save your changes to Github/Bit Bucket.

**Deliverables:** *•* A clean, well-prepared dataset stored as a new CSV file or as DataFrame (Pandas, Pyspark).• A detailed report outlining the data cleaning and preprocessing steps taken, with justifications for each decision.• An initial EDA report that includes key insights and visualizations (e.g., histograms, scatter plots, heatmaps) generated using Python libraries like Matplotlib or Seaborn.

OPTIONAL:

Use ML models to derive better insights from the dataset. Suggest Logistic Regression, Decision Trees, Random Forest, k-NN, XG Boost, Neural Network etc.

### **Part C: Loading Data into a SQL Database** **After completing the initial EDA, participants will load the cleaned and processed dataset into a SQL database. This step requires them to:**

1. Create a database using a local SQL server or a cloud-based service (e.g., MySQL, PostgreSQL, SQLite for simplicity).
2. Pull data from Github, and insert into SQL tables, ensuring that the schema is well-structured (e.g., appropriate data types, primary keys).
3. Push changes to GitHub.

#### **Deliverables:** • A SQL script or Python code (using libraries like sqlite3, SQLAlchemy, or psycopg2) that creates the database schema and populates it with the cleaned data.

#### • A brief report explaining the database schema design, including table structures and relationships.

# **Part D: Performing Data Analysis with SQL** **In the penultimate part of the challenge, participants will use SQL to query the data stored in the database and retrieve specific subsets for further analysis. They will then connect Power BI/ Tableau (or other viz tool) to the SQL database to create visualizations and dashboards that model the insights gained from the data.** **SQL Tasks:** 1. Retrieve Summarized Data: Write SQL queries to extract aggregated data, such as average glucose levels by age group, or the distribution of Diabetes cases across different demographics. You will receive more points by using more than the queries provided below. Add some suggestions here (maybe w3 school, and/or a list of common clauses).

• Query Example:  
  
SELECT age\_group, AVG(glucose) AS avg\_glucose, COUNT(\*) AS cases  
FROM diabetes\_data  
GROUP BY age\_group;  
  
  
2. Filter Data for Specific Analyses: Extract subsets of data for more detailed analysis, such as patients with high blood pressure or cholesterol.

• Query Example:  
  
SELECT \* FROM diabetes\_data WHERE glucose > 200;  
  
3. Create Views for Power BI: Use SQL to create views that simplify complex queries or combine multiple tables, making it easier to work with the data in Power BI.

• Query Example:  
  
CREATE VIEW high\_risk\_patients AS  
SELECT patient\_id, age, cholesterol, resting\_glucose  
FROM diabetes\_data   
WHERE glucose > 240 OR resting\_blood\_pressure > 140;  
  
OPTIONAL:

* Generate additional queries.

**Deliverables:**  
  
• A set of SQL queries that effectively retrieve and summarize the data needed for analysis.

### **Part E: Build Visualizations in Power BI**

#### **Power BI Tasks:**

In the final part of the challenge, participants will visualise their findings.

1. Connect Power BI to the SQL Database: Establish a connection between Power BI and the SQL database where the data is stored.  
2. Create Visualizations: Use the data retrieved via SQL queries to build visualizations that highlight key insights, such as:  
 - Distribution of Diabetes cases by age group.  
 - Correlation between cholesterol levels and Diabetes risk.  
 - Geographic or demographic patterns in the data.  
3. Build a Dashboard: Compile these visualizations into a cohesive dashboard that provides a comprehensive overview of the factors associated with Diabetes.

#### **Deliverables:** • A Power BI dashboard that visualizes key insights, supported by explanations of how the visualizations help to understand the data. • A final presentation or report summarizing the findings and their potential implications for Diabetes prevention or management.

#### **Summary:**

#### This structured problem statement guides participants through the full data analysis process, from extraction and cleaning through to database management, SQL querying, and final visualization in Power BI. This approach not only tests participants’ technical skills in Python and SQL but also emphasizes the importance of integrating these tools into a coherent workflow for data-driven decision-making.

**About Dataset**

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, cholesterol\_level, diabetes\_pedigree\_function, family\_diabetes\_history, diet\_type, star\_sign, social\_media\_usage, physical\_activity\_level, and sleep\_duration. 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.

**gender**

Gender refers to the biological sex of the individual, which can have an impact on their susceptibility to diabetes. There are four categories in it male, female and other. Due to the limitations of standard medical records, individuals are categorized into two gender groups for the purposes of this exercise: male, female. Please note that this categorization is simplified and does not accurately reflect the full spectrum of gender identities

**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.

**cholesterol\_level**

Healthcare providers measure cholesterol levels as milligrams of cholesterol per deciliter of blood. The abbreviation is mg/dL. Total Cholesterol is the total amount of cholesterol that’s circulating in the patient's blood.

**diabetes\_pedigree\_function**

Refers to a specific calculation used in the context of diabetes research, particularly in the assessment of Type 2 diabetes risk based on family history.This function might represent a numeric score derived from a patient's family history of diabetes, which helps in predicting the likelihood of developing diabetes. It typically takes into account factors such as the number of relatives with diabetes and their degree of relatedness to the individual.

**diet\_type**

Refers to the classification of an individual's dietary habits or preferences. This feature can play a significant role in assessing the risk of developing diabetes, especially Type 2 diabetes, as diet is a key factor in metabolic health. Categories include:

Vegetarian: Primarily plant-based diet, may include dairy and eggs.

Vegan: Strictly plant-based diet, excludes all animal products.

Low-Carbohydrate: Diet focused on reducing carbohydrate intake, often high in proteins and fats. Mediterranean: Emphasizes fruits, vegetables, whole grains, healthy fats (like olive oil), and lean protein sources.

Standard American Diet (SAD): Typical Western diet, often high in processed foods, sugars, and fats. Free: Diet excluding gluten-containing foods, suitable for individuals with gluten sensitivities or celiac disease.

Pescatarian: Primarily plant-based diet that includes fish and seafood but excludes other meats.

Carnivore: Diet consisting entirely of animal products, excluding all plant-based foods.

**star\_sign**

Refers to the astrological sign assigned to an individual based on their birthdate.

**bmi**

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.

**weight**

This refers to an individual's body weight, typically measured in kilograms or pounds. This feature is crucial for assessing the risk of developing diabetes, particularly Type 2 diabetes, as body weight is a significant risk factor for metabolic health.

**family\_diabetes\_history**

Categorical variable based on if there is history or none.

**social media usage**

The feature "social\_media\_usage" refers to an individual's habits and behaviors related to social media platforms. This feature can provide insights into lifestyle, mental health, and social connections, all of which can influence diabetes risk and management. Categories include:

Never: No use of social media.

Rarely: Uses social media infrequently (e.g., less than once a week).

Occasionally: Uses social media a few times a month (more than rarely but not regularly).

Moderate: Uses social media several times a week (active engagement without being overwhelming).

Excessive: Uses social media daily or multiple times a day.

**physical\_activity\_level**

This refers to an individual's engagement in physical exercise and movement. This feature is critical for assessing overall health and predicting the risk of developing Type 2 diabetes, as regular physical activity has a well-established impact on metabolic health. Categories include:

Sedentary: Little to no physical activity (e.g., desk job, minimal exercise).

Lightly Active: Light physical activity (e.g., light walking, some household chores, or stretching). Moderately Active: Engages in moderate exercise (e.g., brisk walking, cycling, or recreational sports a few times a week).

Very Active: Regular vigorous exercise (e.g., running, weight training, or sports several times a week).

Extremely Active: High levels of physical activity (e.g., professional athletes, daily intense workouts).

**sleep\_duration**

How many hours on average the patient sleeps per night, between 4 - 10 hours.

**stress\_level**

This refers to the degree of psychological stress an individual experiences. This feature can be significant for understanding various health outcomes, including the risk of developing Type 2 diabetes, as chronic stress is known to affect metabolic processes and overall health. Categories include:

Low: Minimal stress; generally coping well with daily life.

Moderate: Some stress; occasional challenges but manageable.

Elevated: Increased stress; frequent challenges that may impact health.

High: Significant stress; ongoing challenges affecting daily functioning.

Extreme: Severe stress; overwhelming feelings that may require intervention.

**pregnancies**

Number of times the patient has been pregnant

**alcohol\_consumption**

This refers to the amount and frequency of alcohol intake by an individual. This feature is significant in assessing the risk of developing Type 2 diabetes, as alcohol consumption can have both positive and negative effects on metabolic health. Categories include:

None: No alcohol consumption.

Light: Occasional drinking (e.g., up to 1 drink per week).

Moderate: Regular drinking (e.g., up to 2-3 drinks per week).

Heavy: Frequent or excessive drinking (e.g., more than 3 drinks per week).

**diabetes**

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