### **REPORT**

### **1. Problem Statement:**

Diabetes is a significant public health issue, impacting millions globally and leading to serious health complications if not managed effectively. The aim of this analysis is to understand trends in diabetes prevalence, identify key contributing factors, and provide actionable insights for effective prevention and management. The investigation will include analyzing variables such as age, lifestyle habits, comorbidities, and glucose levels to uncover patterns and inform healthcare strategies.

### **2. Data Requirements:**

#### **Patient Demographics**

* Age
* Pregnancy history (for women)

#### **Health Metrics**

* Plasma glucose levels
* Blood pressure (diastolic)
* Skinfold thickness (as a proxy for body fat)
* Insulin levels
* BMI (Body Mass Index)
* Diabetes pedigree function (genetic risk score)

#### **Diabetes Status**

* Outcome (1: diabetic, 0: non-diabetic)

#### **Aggregated Insights**

* Average glucose levels by age group
* Percentage of individuals at high risk for diabetes
* Correlation between BMI and diabetes prevalence

### **3. Data Collection Process**

#### **Medical Records**

* Gather patient data from electronic health records (EHR) systems, including demographic and clinical information.

#### **Health Surveys**

* Extract data on lifestyle habits and dietary patterns from patient self-reported surveys or public health databases.

#### **Diagnostic Labs**

* Collect lab results for blood glucose, HbA1c, and cholesterol levels from diagnostic centers or hospital reports.

#### **Integrated Health Monitoring Systems**

* Use wearable device data (e.g., activity trackers, glucose monitors) to obtain real-time physical activity and glucose trends.

#### **Data Integration**

* Consolidate all datasets into a single repository using a unique Patient ID as the primary key.
* Ensure the data is cleaned, anonymized, and free of inconsistencies to comply with patient privacy regulations (e.g., HIPAA, GDPR).

#### **Validation and Enrichment**

* Validate the data against population health benchmarks and ensure consistency across data sources.
* Enrich datasets with contextual features such as socioeconomic status or geographical distribution for more comprehensive analysis.