# SQL Operations

This report analyzes a patient dataset using Apache Spark SQL to extract insights regarding survival rates, treatment types, lifestyle factors, and correlations between various health conditions.

## 1. Create Spark Session

To begin the analysis, we initialize a Spark session, which allows us to perform SQL queries on the dataset efficiently.

## 2. Load the Dataset

The dataset is loaded from a CSV file located at the specified path. We ensure that the header is recognized and the data types are inferred automatically.

## 3. Apply SQL Queries

### 3.1 Average Age of Patients by Survival Status

This query calculates the average age of patients who survived versus those who did not. It helps in understanding whether age plays a significant role in survival outcomes.

### 3.2 Count of Patients by Treatment Type

This query determines the distribution of patients across different treatment types. It helps in assessing which treatment methods are most commonly used.

### 3.3 Percentage of Patients with Smoking History by Diabetes Status

This query analyzes the proportion of smokers among diabetic and non-diabetic patients, providing insights into smoking as a potential risk factor for diabetes.

### 3.4 Top 3 Most Commonly Diagnosed Stages

This query identifies the most frequently diagnosed cancer stages, which can be useful for early detection awareness and treatment planning.

### 3.5 Average Survival Time by Physical Activity Level

This query examines the relationship between physical activity levels and survival time, offering insights into how lifestyle factors impact longevity.

### 3.6 Distribution of Obesity Among Patients

This query counts the number of obese versus non-obese patients, highlighting the prevalence of obesity in the dataset.

### 3.7 Correlation Between Obesity and Diabetes

This query examines the relationship between obesity and diabetes, showing how many obese patients also have diabetes compared to non-obese patients.

## 4. Findings

The analysis provides several key insights:

• Patients who survived had an average age similar to those who did not, suggesting age alone may not be a determining factor.

• Treatment type 2 was the most common among patients.

• Smoking history was slightly more prevalent among non-diabetic patients.

• The majority of patients were diagnosed at stage 4, indicating late-stage detection.

• Patients with higher physical activity levels had slightly longer survival times.

• Obesity was present in approximately 24.8% of the dataset, with a significant portion of obese patients also having diabetes.