# Introduction to MySQL and Data Analysis

# By Adnan Rasool

## 1. What is MySQL?

MySQL is an open-source relational database management system (RDBMS) that uses Structured Query Language (SQL) to store, manage, and retrieve data efficiently. It supports operations such as data insertion, querying, updating, and deletion in a structured format, ideal for applications requiring complex data relationships.

### Key Features

- Easy-to-use interface for managing relational data.
- Cross-platform support.
- Scalability and high performance.
- Extensive community and enterprise support.

### Applications

- Web development.
- Data analytics.
- Enterprise applications.
- IoT (Internet of Things).

## 2. Dataset Overview

This project uses a dataset of patients containing health-related metrics to study diabetes prevalence. The dataset includes fields like:

| Field | Description |
|---|---|
| `gender` | Patient's gender |
| `age` | Patient's age |
| `hypertension` | Whether the patient has hypertension |
| `heart_disease` | History of heart disease |
| `smoking_history` | Patient's smoking history |
| `bmi` | Body Mass Index |
| `HbA1c_level` | Glycated hemoglobin level |
| `blood_glucose_level` | Blood glucose concentration |
| `diabetes` | Whether the patient has diabetes (1/0) |

### Sample Dataset

| gender | age | hypertension | heart_disease | smoking_history | bmi | HbA1c_level | blood_glucose_level | diabetes |
|---|---|---|---|---|---|---|---|---|
| Female | 80.0 | 0 | 1 | never | 25.19 | 6.6 | 140 | 0 |
| Female | 54.0 | 0 | 0 | No Info | 27.32 | 6.6 | 80 | 0 |
| Male | 28.0 | 0 | 0 | never | 27.32 | 5.7 | 158 | 0 |
| Female | 36.0 | 0 | 0 | current | 23.45 | 5.0 | 155 | 0 |

---

## 3. Code Overview

### 3.1 `create_database.py`

#### Purpose

This script creates a MySQL database, imports the dataset, and generates a table to store the data.

#### Key Steps

1. **Connect to MySQL Server**

```
conn = mysql.connector.connect(
    host='localhost',
    user='root',
    password='root@123'
)
```

2. **Create Database and Table**

```
cursor.execute('CREATE DATABASE IF NOT EXISTS diabetes_db')
cursor.execute('''CREATE TABLE IF NOT EXISTS diabetes_data (...)
```

3. **Import Dataset**

```
with open(csv_filename, 'r', encoding='utf-8') as file:
    csv_reader = csv.DictReader(file)
    for row in csv_reader:
        cursor.execute('''INSERT INTO diabetes_data (...) VALUES (...)''', row.values())
```

4. **Sample Output**

   When the script is executed with a valid dataset:

```
Successfully inserted 10 records into the database.
```

---

### 3.2 `sql_queries.py`

#### Purpose

This script executes predefined SQL queries to analyze the diabetes dataset.

#### Key Queries

1. **Total Number of Patients**

```
SELECT COUNT(*) as total_patients FROM diabetes_data;
```

2. **Diabetes Prevalence by Age Group**

```
SELECT CASE
         WHEN age < 20 THEN '0-19'
         WHEN age < 40 THEN '20-39'
         WHEN age < 60 THEN '40-59'
         ELSE '60+'
       END as age_group,
       COUNT(*) as total_count,
       SUM(diabetes) as diabetes_count,
       ROUND(SUM(diabetes) / COUNT(*) * 100, 2) as diabetes_percentage
FROM diabetes_data
GROUP BY age_group
ORDER BY age_group;
```

3. **Smoking History Distribution**

```
SELECT smoking_history, COUNT(*)
FROM diabetes_data
GROUP BY smoking_history
ORDER BY COUNT(*) DESC;
```

4. **Average Metrics by Diabetes Status**

```
SELECT
    diabetes,
    ROUND(AVG(bmi), 2) as avg_bmi,
    ROUND(AVG(HbA1c_level), 2) as avg_HbA1c,
    ROUND(AVG(blood_glucose_level), 2) as avg_glucose
FROM diabetes_data
GROUP BY diabetes;
```

5. **Gender Distribution**

```
SELECT gender, COUNT(*) as count
FROM diabetes_data
GROUP BY gender;
```

6. **Patients by Age Group**

```
SELECT
    CASE
        WHEN age < 40 THEN 'Under 40'
        WHEN age < 60 THEN '40-59'
        ELSE '60 and above'
    END as age_group,
    COUNT(*) as count
FROM diabetes_data
GROUP BY age_group
ORDER BY age_group;
```

#### Sample Outputs

- **Diabetes by Age Group:**

```
Age group: 60+ - 3 people, 1 with diabetes (33.33%)
```

- **Average BMI for Diabetics vs Non-Diabetics:**

```
With diabetes: Avg BMI = 28.5
Without diabetes: Avg BMI = 24.7
```

---

## 4. Getting Started

### Prerequisites

- Install MySQL Server.
- Python 3.x with `mysql-connector-python` library.

### Running the Scripts

1. Create and populate the database:

```
python create_database.py
```

2. Execute SQL queries for analysis:

```
python sql_queries.py
```