# Dataset Exploration Report

Project 3: Healthcare Predictive Analytics

Milestone 1 - Data Collection, Exploration, and Preprocessing

## 1. General Overview

The dataset contains 100000 rows and 9 columns. It focuses on predicting whether a patient has diabetes based on several medical and personal attributes.

The columns include:

- gender (categorical): Gender of the patient.  
- age (numerical): Age of the patient.  
- hypertension (binary): Whether the patient has hypertension (0 or 1).  
- heart\_disease (binary): Whether the patient has heart disease (0 or 1).  
- smoking\_history (categorical): Smoking status.  
- bmi (numerical): Body Mass Index.  
- HbA1c\_level (numerical): Glycated hemoglobin level.  
- blood\_glucose\_level (numerical): Blood glucose level.  
- diabetes (binary): Whether the patient is diagnosed with diabetes (target).

## 2. Data Quality

✔ No missing values are currently present in the dataset. Missing values were previously treated using the ‘KNN’ imputation technique.

✔ Number of duplicate rows found: 4969

## 3. Feature Distribution Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Mean | Std Dev | Min | Max |
| age | 41.89 | 22.52 | 0.08 | 80.0 |
| bmi | 27.32 | 6.64 | 10.01 | 95.69 |
| HbA1c\_level | 5.53 | 1.07 | 3.5 | 9.0 |
| blood\_glucose\_level | 138.06 | 40.71 | 80.0 | 300.0 |

## 4. Categorical Insights

Examples of most common values:

- gender: 3 unique values (most frequent: Female)  
- smoking\_history: 5 unique values (most frequent: never)  
- diabetes: 2 values (0 = No diabetes, 1 = Diabetes), majority class = 0 (~91.5%)

