1. What is the source of your dataset?

The dataset is sourced from data.gov, the official open data portal of the U.S. Government. Specifically, it is based on publicly available health survey datasets such as:

• BRFSS (Behavioral Risk Factor Surveillance System) – CDC
• NHANES (National Health and Nutrition Examination Survey) – CDC / NIH

Source Website:
https://www.data.gov

Relevant Dataset Pages:
https://www.cdc.gov/brfss
https://www.cdc.gov/nchs/nhanes

The final dataset used in the project is a synthetic, structured version generated using the schema and distributions of these datasets.

2. Why did you choose this dataset for your problem statement?

This dataset was chosen because it is highly relevant to the problem of predicting early risk of silent diseases such as diabetes.

It contains key medical and lifestyle indicators like Age, BMI, Blood Pressure, Cholesterol, Physical Activity, and Smoking status, which are proven risk factors for diabetes.

The dataset supports binary classification (Diabetes risk vs. no risk), aligning perfectly with the project objective.

The data is public, anonymized, and ethical, ensuring privacy and fairness.

The structured format makes it ideal for machine learning models and Power BI visualization.

3. How was the data collected?

The data was collected from open government health datasets published by the CDC and NIH on data.gov.

Original data collection was done through large-scale national health surveys conducted by government health agencies.

These surveys gather information via questionnaires, interviews, and medical examinations.

The dataset used in this project is a synthetic dataset generated from these open datasets, preserving statistical distributions, feature relationships, and risk patterns.

No web scraping or personal data collection was performed.