Advanced Exploratory Data Analysis (EDA) for GlucoSense AI Diabetes Detection

A Data-Driven Approach for Early Diabetes Diagnosis

Presented by: Dhayanithi

01

Overview

Objective

Understanding key trends and insights from the dataset to aid diabetes prediction.

Dataset Overview

- The dataset has 1,00,000 records and 9 columns. There are no missing values in any column.

- Features: Gender, Age, Hypertension, Heart Disease, Smoking History, BMI, HbA1c, Blood Glucose, Diabetes (target).

Goal

Identify patterns, correlations, and anomalies for improved model performance.

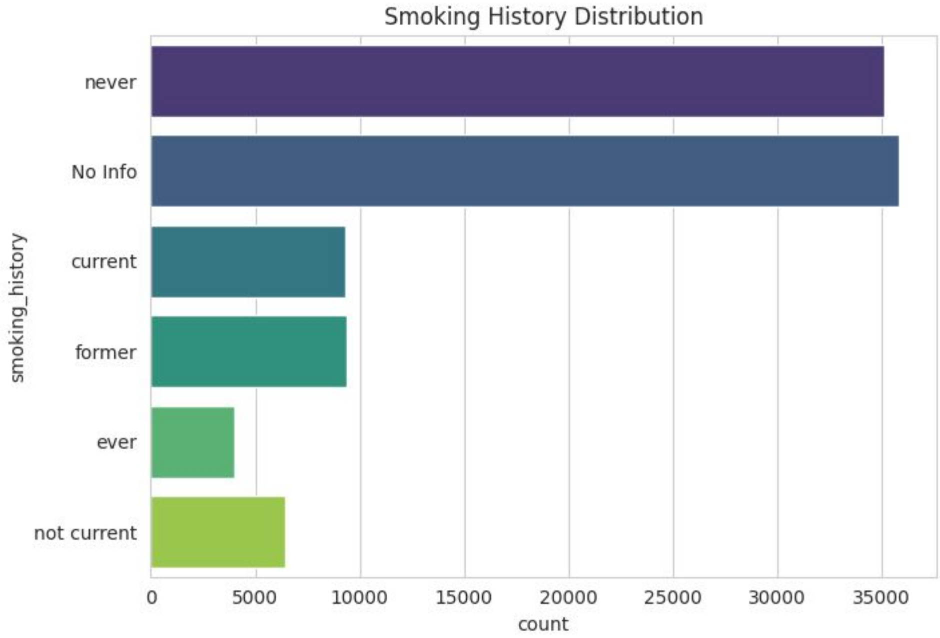
	A	B	C	D	E	F	G	H	I
1	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_lev	diabetes
2	Female	80	0	1	never	25.19	6.6	140	0
3	Female	54	0	0	No Info	27.32	6.6	80	0
4	Male	28	0	0	never	27.32	5.7	158	0
5	Female	36	0	0	current	23.45	5	155	0
6	Male	76	1	1	current	20.14	4.8	155	0
7	Female	20	0	0	never	27.32	6.6	85	0
8	Female	44	0	0	never	19.31	6.5	200	1
9	Female	79	0	0	No Info	23.86	5.7	85	0
10	Male	42	0	0	never	33.64	4.8	145	0
11	Female	32	0	0	never	27.32	5	100	0
12	Female	53	0	0	never	27.32	6.1	85	0
13	Female	54	0	0	former	54.7	6	100	0
14	Female	78	0	0	former	36.05	5	130	0
15	Female	67	0	0	never	25.69	5.8	200	0
16	Female	76	0	0	No Info	27.32	5	160	0
17	Male	78	0	0	No Info	27.32	6.6	126	0
18	Male	15	0	0	never	30.36	6.1	200	0
19	Female	42	0	0	never	24.48	5.7	158	0
20	Female	42	0	0	No Info	27.32	5.7	80	0
21	Male	37	0	0	ever	25.72	3.5	159	0
22	Male	40	0	0	current	36.38	6	90	0
23	Male	5	0	0	No Info	18.8	6.2	85	0

02

Data Quality Check

Missing Values

- smoking_history has many "No Info" entries (treat as missing).
- This means that for these records, the smoking status of the individual is unknown or not recorded.



- it can reduce the accuracy of insights and predictions.

Possible Action (optional)

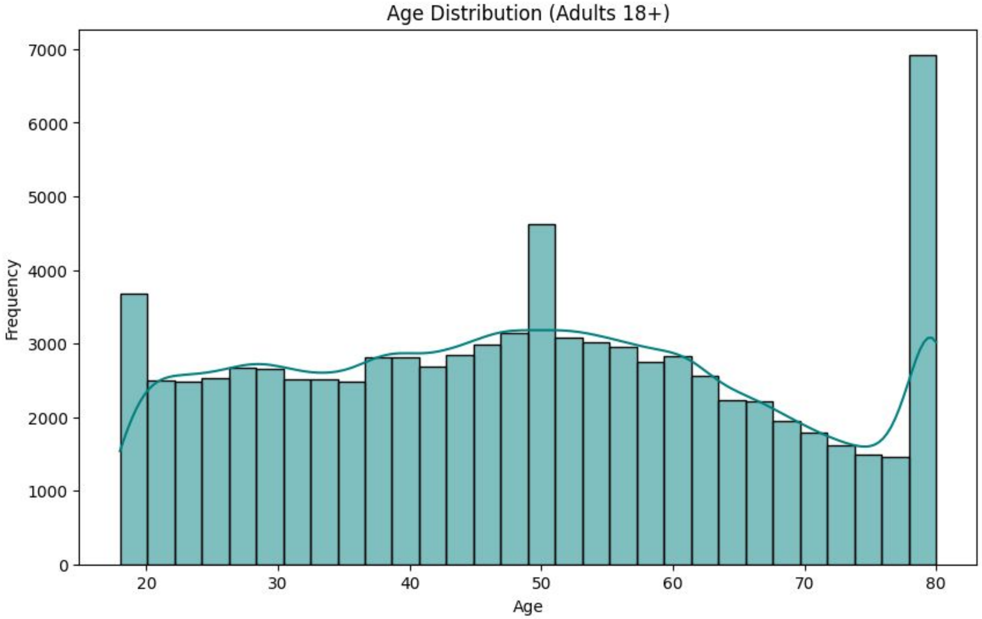
Imputation (Filling Missing Values)

- Replace "No Info" with "Non-Smoker" if most missing cases are likely non-smokers.
- Use statistical methods (like mode or KNN imputation) to estimate missing values.

Missing Values

- In the dataset, some records have implausible age values, such as 0.08 years (approximately 1 month old).

- Infants (age < 1 year) are not typically screened for diabetes in routine medical assessments.



Possible Action (optional)

Remove Records with Age < 1 Year

- Remove all records where age is less than 1 year to ensure valid data points for analysis.

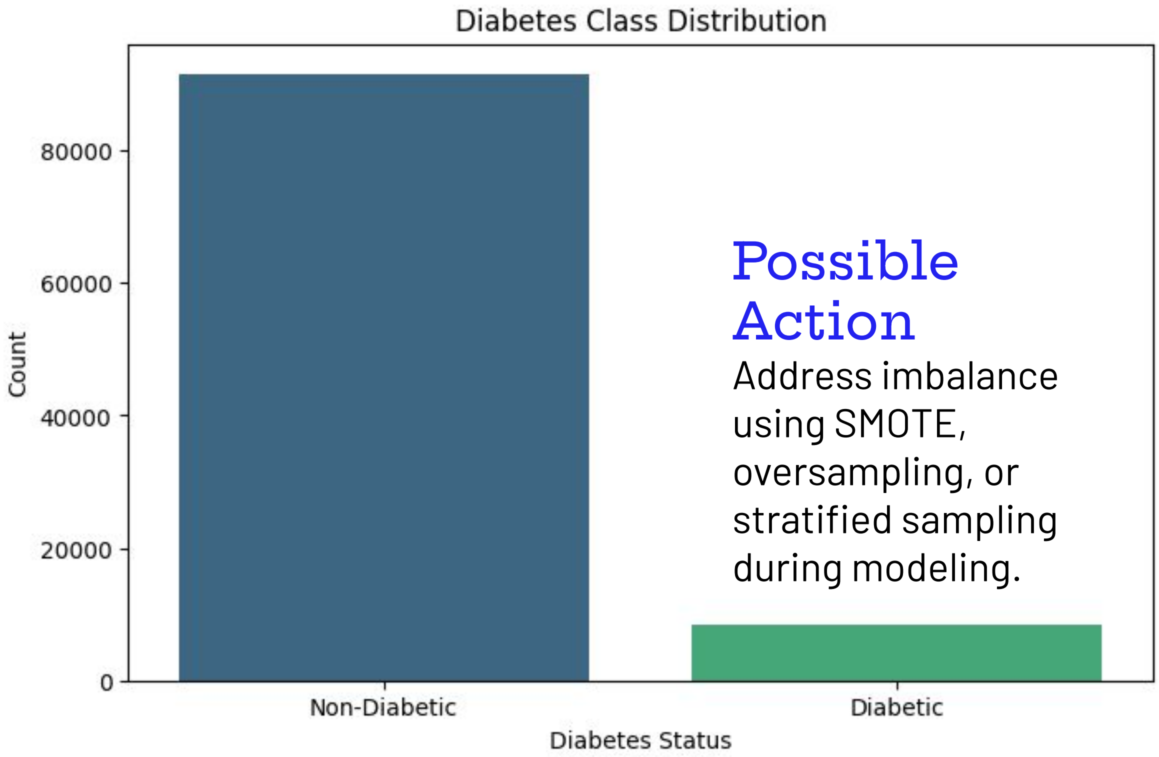
- Keep only records where age ≥ 18 to focus on adults, ensuring more meaningful and reliable insights.

03

Target Variable Distribution

Distribution of Diabetes

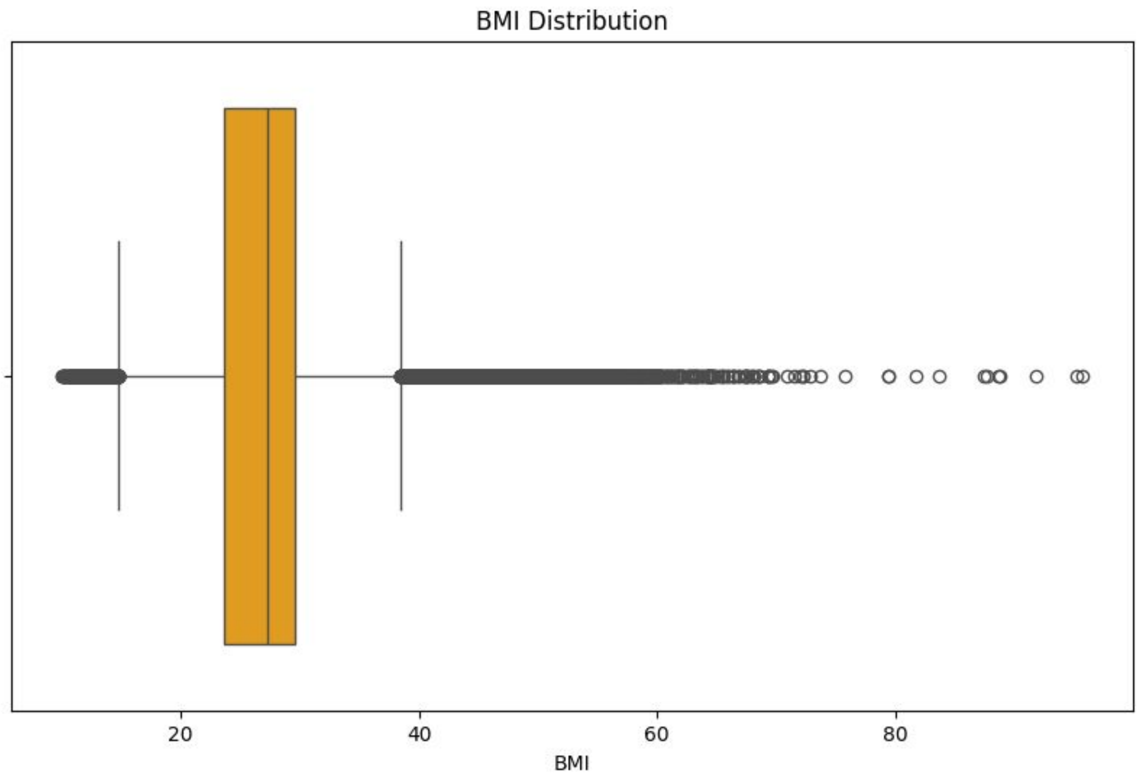
- The dataset is imbalanced, with only 12% of patients having diabetes.
- Example: 120 out of 1,000 patients are diabetic, while the remaining 880 are non-diabetic.
- This imbalance can cause the model to be biased toward the majority class (non-diabetic cases), leading to poor predictions for diabetic patients.



Possible Action

Address imbalance using SMOTE, oversampling, or stratified sampling during modeling.

Key Feature Distributions – BMI



- The mean BMI is ~27.5, which falls in the overweight category (BMI 25–29.9).
- Some extreme BMI values above 60 are detected, which may be data entry errors or unrealistic values.

Possible Action

Cap BMI at 40 to remove extreme outliers while retaining valid high BMI cases.
This prevents model bias due to unrealistic data points.

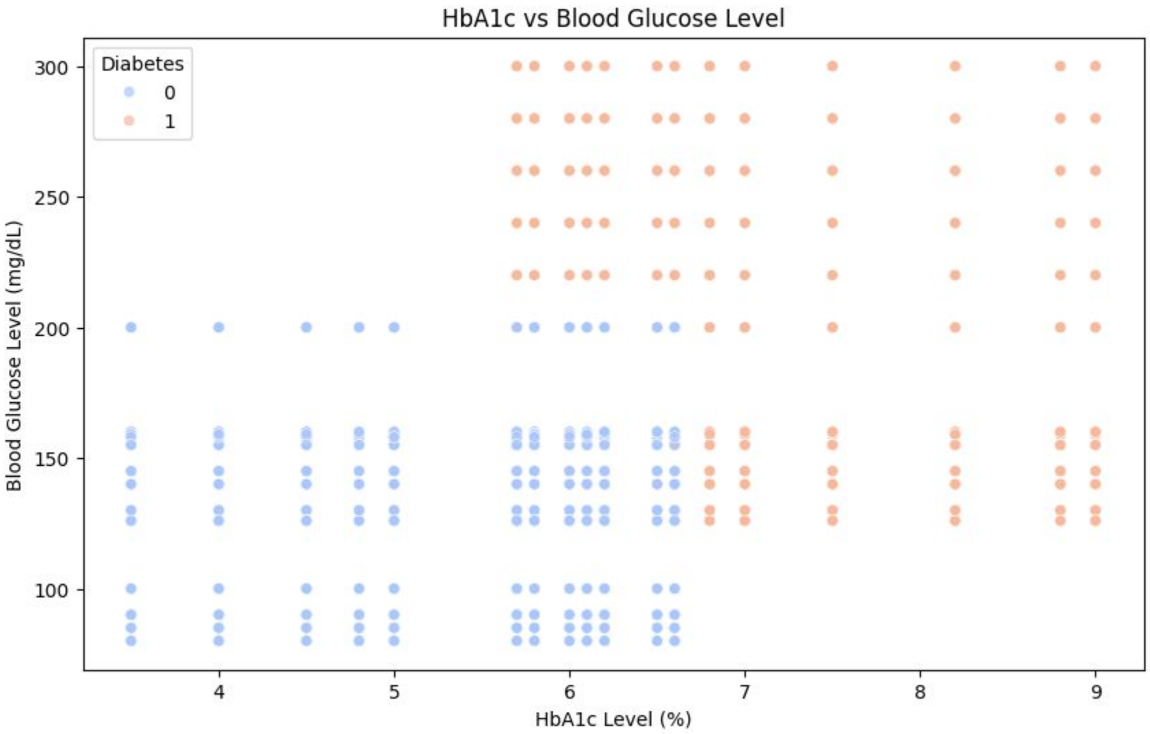
Key Feature Distributions – HbA1c & Blood Glucose

- Strong positive correlation: As HbA1c increases, Blood Glucose also rises.

- Diabetic patients cluster at:
  HbA1c ≥ 6.5% (diabetes threshold)
  Blood Glucose ≥ 200 mg/dL (common in diabetic individuals).

Add-on

Treat HbA1c and Blood Glucose as key predictive features for diabetes detection.



04

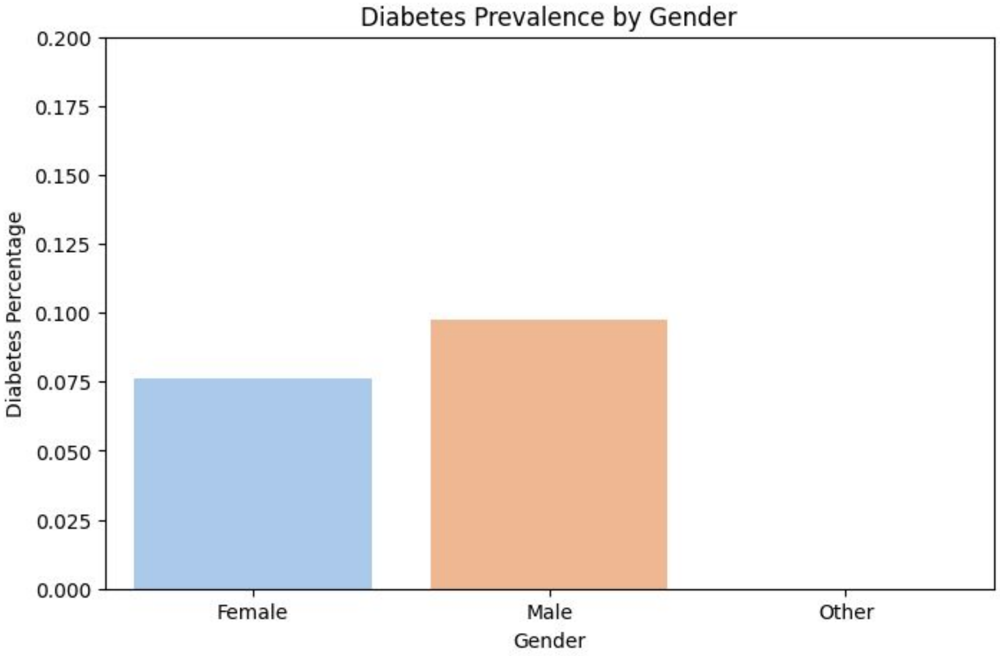
Categorical Features

Categorical Features – Gender & Diabetes

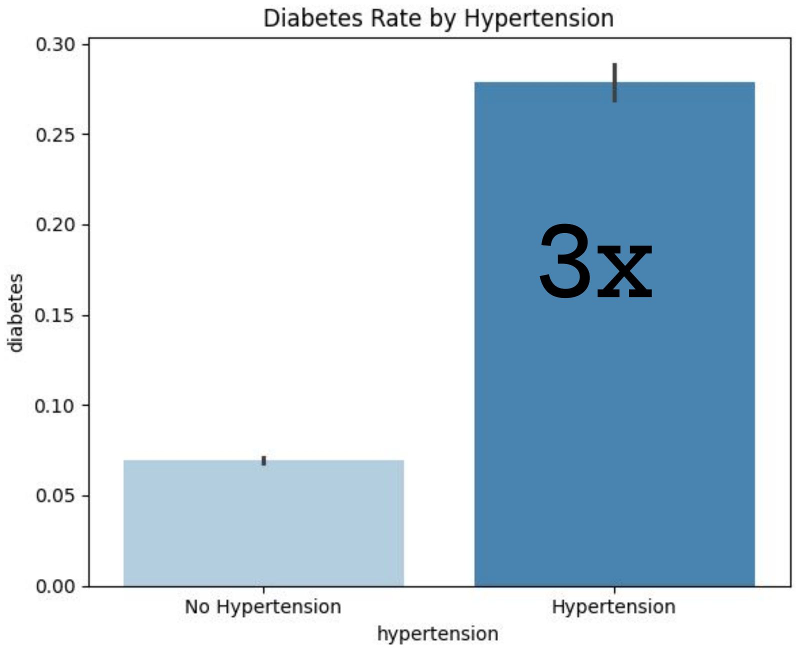
Diabetes prevalence:
14% in males
10% in females

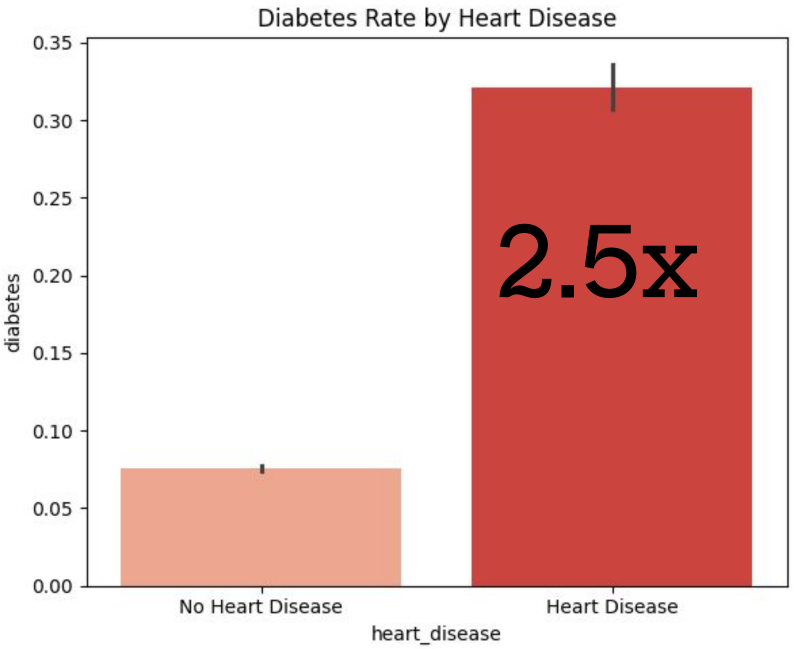
Key Takeaway:

Gender plays a role in diabetes risk, and it should be considered in data analysis and predictive modeling.

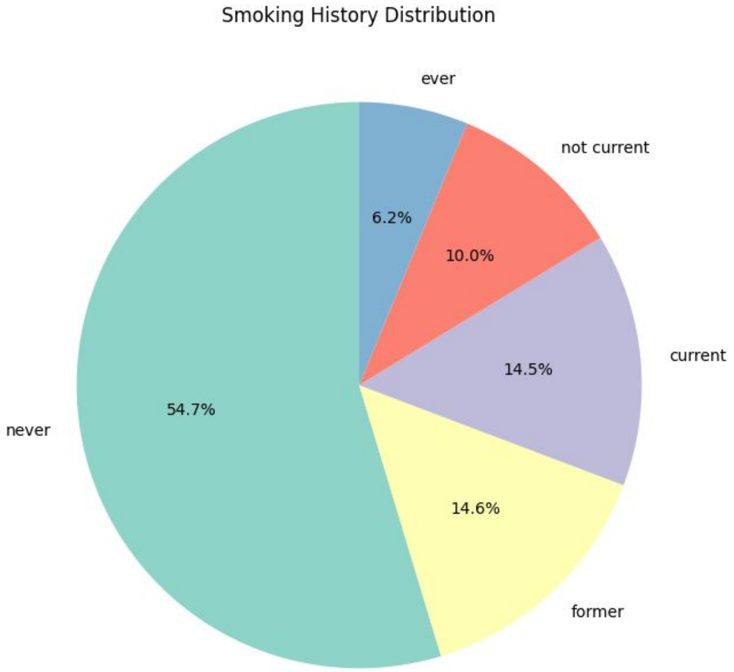


Categorical Features – Hypertension & Heart Disease





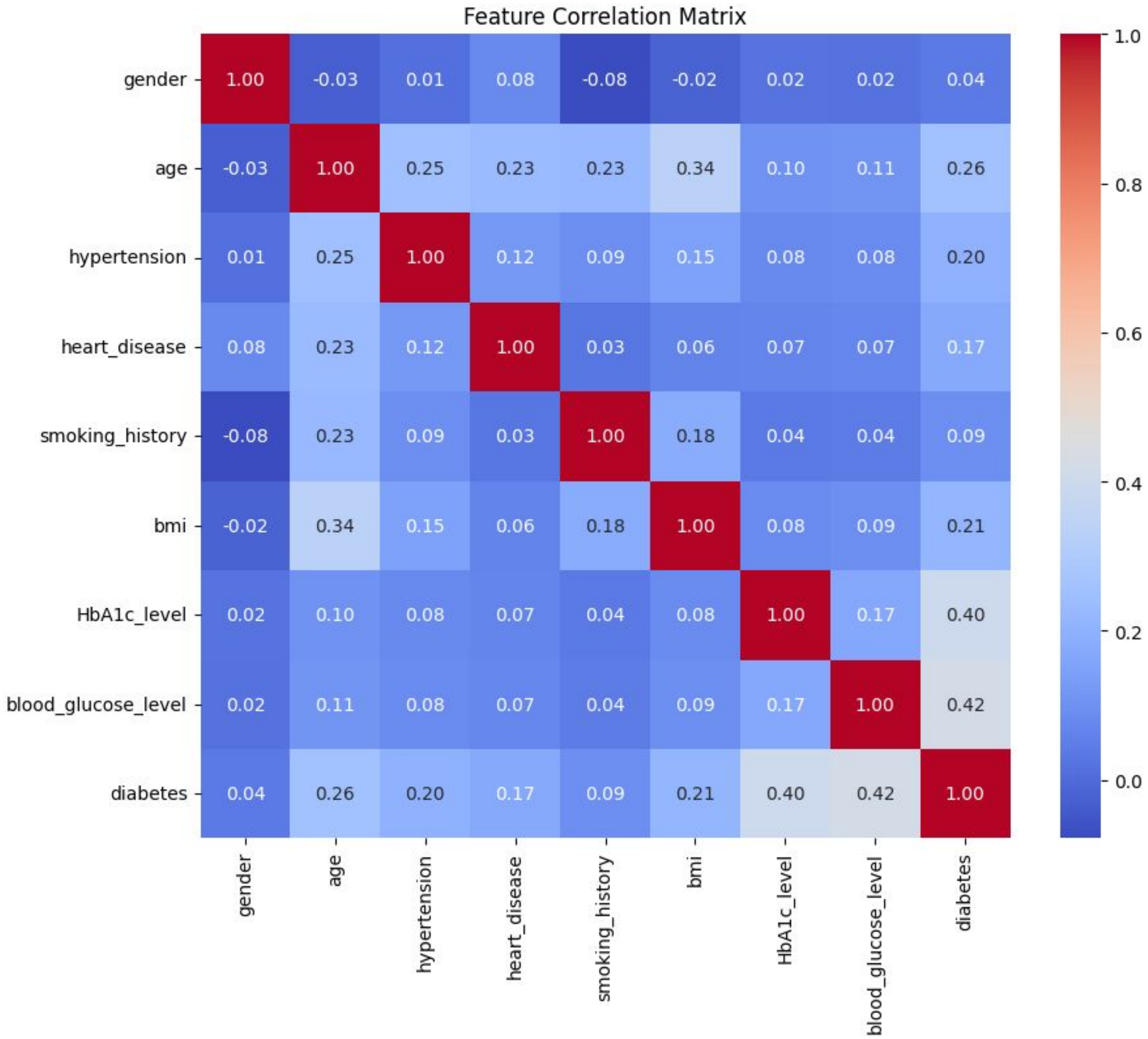
Categorical Features – Smoking History



Former smokers have the highest diabetes rate (**18%**), followed by current smokers (**15%**).
suggests that past smoking may have long-term effects on diabetes risk.

05

Correlation Analysis



Key Correlations:

HbA1c & Blood Glucose: +0.75 (Strong Positive Correlation)
Higher HbA1c is strongly associated with higher Blood Glucose levels.
Age & Hypertension: +0.4 (Moderate Positive Correlation)
Older individuals are more likely to have hypertension.
BMI & Diabetes: +0.3 (Moderate Positive Correlation)
Higher BMI slightly increases diabetes risk.

06

Action Plan

Feature Selection:
HbA1c, Blood Glucose, Age, Hypertension, and BMI are key predictors.

Data Preprocessing:
Handle missing smoking data.
Normalize/cap outliers (BMI, Age).

Feature Engineering:
Categorize BMI into underweight/normal/overweight/obese.

Let’s build and refine the GlucoSense AI Diabetes Detection Model