# Diabetes Prediction Using Medical and Demographic Factors

1st Kanika Saxena
ksaxena2@buffalo.edu

2nd Rajashree Shanmuganathan
rajashre@buffalo.edu

3rd Alankriti Dubey
adubey2@buffalo.edu

## I. Problem Statement

This report aims to investigate the relationship between demographic and medical factors and diabetes risk, identifying key patterns that can inform early intervention and prevention strategies. Specifically, the analysis focuses on understanding how obesity, age, and lifestyle choices like smoking contribute to the onset and progression of diabetes, and how these insights can be used to target at-risk populations more effectively.

### A. Background

Diabetes affects millions of people worldwide, and many of those cases are driven by medical and demographic factors. Predictive models using basic medical and demographic data can offer an early warning system, allowing timely detection and reducing patient complications. This is significant, especially in settings where routine testing may not be feasible, as well as early detection and timely treatment of potential patients.

### B. Contribution to the Problem Domain

Contribution Potential: This project has the potential to significantly enhance diabetes management by:

- Early Risk Identification: Helping healthcare providers focus on individuals at high risk, enabling preventive procedures.
- Healthcare Efficiency: Reducing the reliance on extensive testing through accurate, data-driven predictions using basic medical and demographic data.
- Personalized Treatment: Supporting tailored healthcare plans based on each individual's unique risk profile.
- Research Insights: Advancing understanding of the relationships between medical and demographic factors and diabetes risk.

## II. Data Source

The dataset used for this project is the Diabetes Prediction Dataset obtained from Kaggle. It contains around 100,000 rows and 9 columns, with the target variable being Diabetes, where a value of 1 indicates the presence of diabetes and 0 indicates its absence. The dataset includes several medical and demographic features such as age, gender, BMI, hypertension, heart disease, HbA1c level, and blood glucose level.

This dataset has been utilized in over 240 Kaggle projects, and has also been referenced in various research papers focusing on diabetes prediction and machine learning applications in healthcare. Its large size and diverse features make it a valuable resource for understanding the factors contributing to diabetes risk.

The dataset can be accessed at: Kaggle Diabetes Prediction Dataset

## III. Exploratory Data Analysis (EDA)

### A. Distribution by Gender

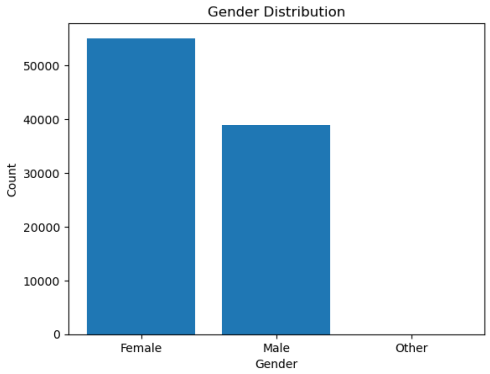


Fig. 1. Distribution by Gender

The histogram shows that a higher proportion of diabetic patients are female compared to males, suggesting a potential gender-related factor in diabetes commonness.

### B. Distribution of BMI

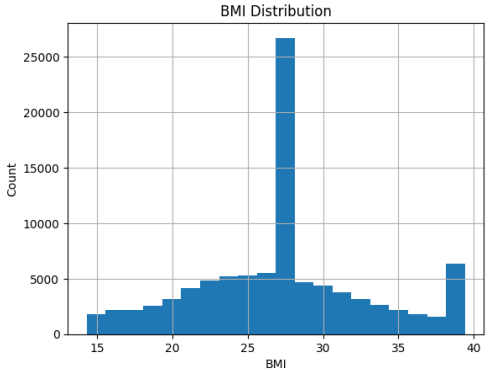


Fig. 2. Distribution of BMI

Approximately 70% of patients in the dataset are classified as either overweight or obese, highlighting a significant relationship between weight and diabetes risk. Given this concerning observation, it is crucial to maintain a healthy weight in order to reduce the likelihood of developing diabetes and similar health complications.

Below is the weight distribution of patients in the table:

| Classification | Count |
|---|---|
| Overweight | 41,701 |
| Obesity | 23,524 |
| Healthy Weight | 21,770 |
| Underweight | 7,055 |

### C. Smoking Habits of Patients

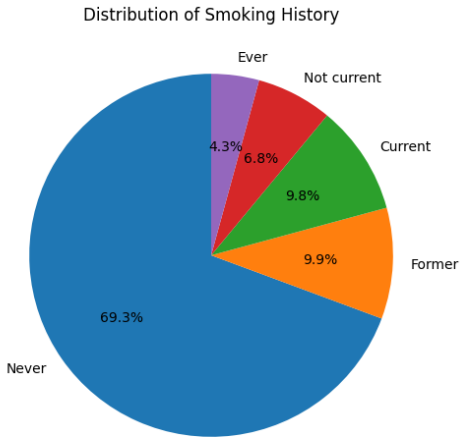


Fig. 3. Smoking Habits of Patients

The visualization of smoking history shows that approximately 70% of patients have never smoked, while 9.7% are current smokers. While the data shows that many patients have never smoked, this doesn't mean that smoking is irrelevant to diabetes risk. Current and former smokers may face higher complications related to diabetes.

### D. Glucose Levels vs BMI

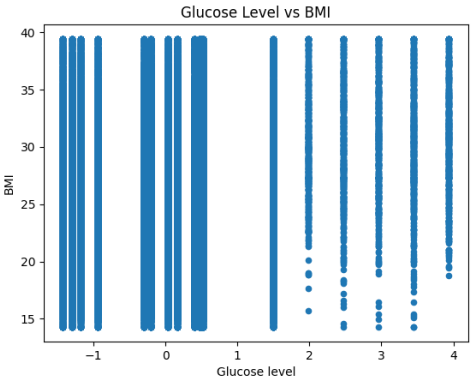


Fig. 4. Glucose Levels vs BMI

This graph concludes that glucose levels can vary independently of BMI values, as shown by the similar glucose levels across a range of BMI classifications. However, higher glucose levels are more commonly observed in patients with a high BMI, indicating a potential association between obesity and glucose dysregulation.

### E. Diabetes Rates with Age

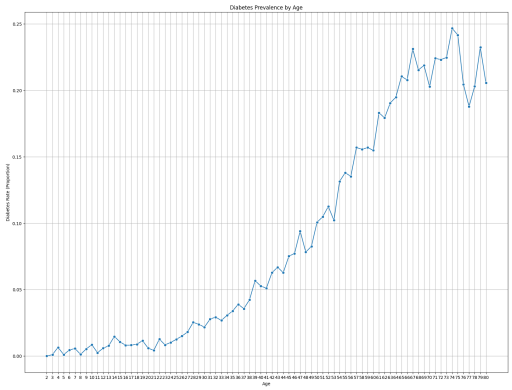


Fig. 5. Diabetes Rates with Age

The line chart clearly demonstrates that the rate of diabetes increases with age, showing a general upward trend despite some slight fluctuations. This highlights the growing risk of diabetes among older populations.

### F. BMI Distribution Across Age Groups

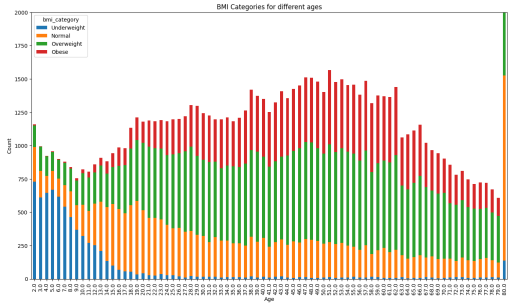


Fig. 6. BMI Distribution Across Age Groups

The line chart clearly demonstrates that the rate of diabetes increases with age, showing a general upward trend despite some slight fluctuations. This highlights the growing risk of diabetes among older populations.

### G. BMI vs Diabetes Risk

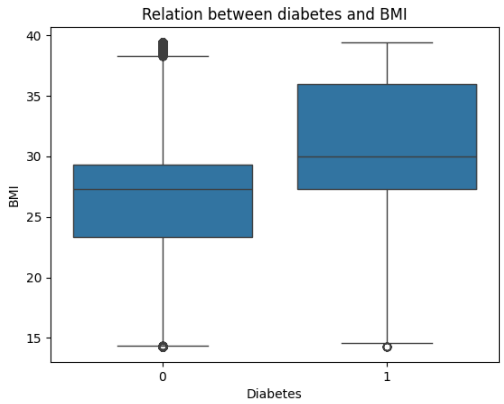


Fig. 7. BMI vs Diabetes Risk

The box plot illustrates that diabetic patients tend to have higher BMIs, with the median BMI falling between 30 and 35. In contrast, non-diabetic individuals have a lower median BMI, around 25 to 30. This suggests that a higher BMI is strongly associated with an increased likelihood of diabetes.

## IV. Key Outcomes and Insights

The analysis shows that higher BMI (27-37 range), older age (especially 40+), and female gender are strongly linked to a higher risk of diabetes. Additionally, a portion of diabetic patients have a history of smoking.

**Key measures** include promoting weight management, early screening for older populations and women, and reinforcing smoking cessation programs to reduce diabetes risk.

This analysis will inform the feature selection process for predictive modeling, with BMI, age, gender, and smoking history identified as the most significant predictors of diabetes risk.