# Dataset:-

The dataset used in this project is the Diabetes Health Indicators dataset, obtained in the form of CSV files and combined for analysis. The dataset comprises 21 input features and a single target label.

- **Datapoints:** Each record represents the health survey data of an individual adult

- **Features:**
  - Health conditions: High blood pressure, high cholesterol, stroke history, heart disease, Cholesterol check in past 5 years
  - Lifestyle factors: Smoking, physical activity, alcohol consumption, fruits intake, vegetable intake
  - Physiological status: BMI, difficulty walking
  - Mental/physical health condition: GenHlth, MentHlth, PhysHlth
  - Healthcare access: AnyHealthcare, NoDocbcCost
  - Demographics: Age, sex, education level, income level

- **Label (Target Variable):** The dataset includes Diabetes_012, which represents diabetes status:
  a. 0 → No Diabetes
  b. 1 → Prediabetes
  c. 2 → Diabetes

# Problem Statement:-

Diabetes is rapidly increasing across the globe, and its early detection is essential to reduce long-term health consequences. However, many individuals remain undiagnosed due to insufficient awareness and limited access to healthcare. This creates a significant challenge in identifying at-risk individuals at the right time.

To address this issue, the goal of this project is to analyze key health and lifestyle indicators to identify major risk factors associated with diabetes and to develop an accurate predictive model. This model will help estimate the likelihood of diabetes in individuals, enabling early detection and better healthcare decision-making.

**Importance of the Problem:**

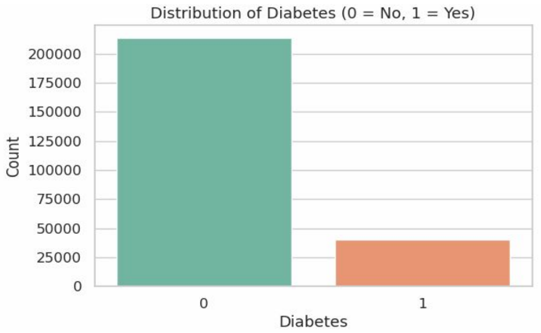
- Diabetes is a major global health concern, impacting millions of lives worldwide
- Early detection helps prevent:
  - Severe complications (heart disease, kidney damage, vision loss)
  - High healthcare and treatment costs
- Many diabetes cases are preventable through lifestyle modifications
- Identifying high-risk individuals allows:
  - Better healthcare planning
  - Targeted awareness programs
  - Improved quality of life through timely intervention

# EDA:-

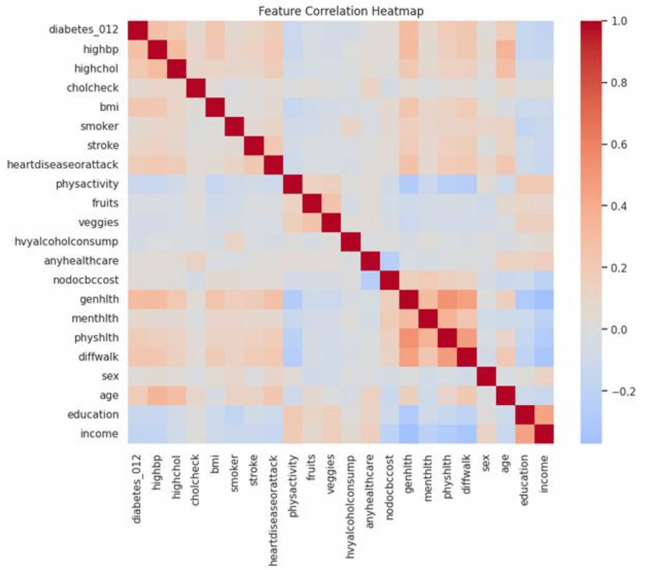
Objective: To understand dataset structure, variable distributions, and relationships affecting diabetes occurrence.

### 1. Dataset Overview

- Total Records: 253,680
- Total Features: 22
- Target Variable: diabetes_012 (converted to binary → 0 = No, 1 = Yes)
- Missing Values: 0
- Duplicates removed: 23,968

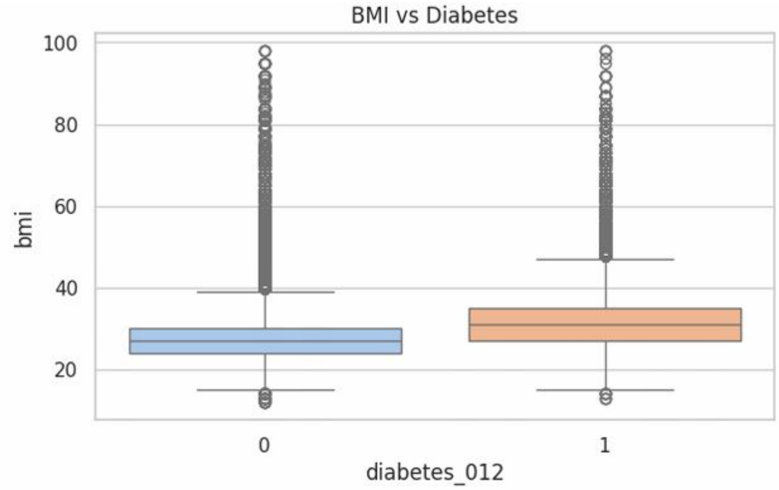


### 2. Feature Correlation Analysis

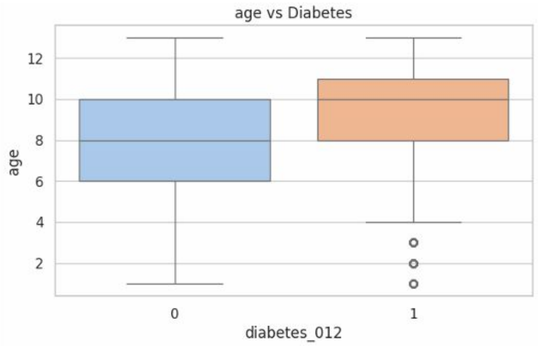


### 3. Feature Insights (Univariate & Bivariate)

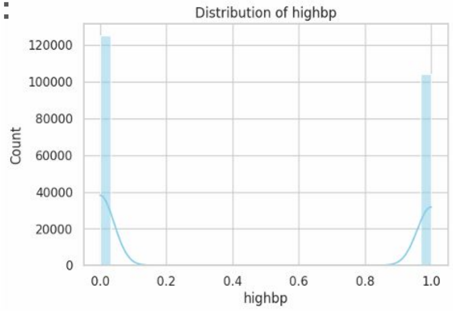
- **BMI vs Diabetes:**

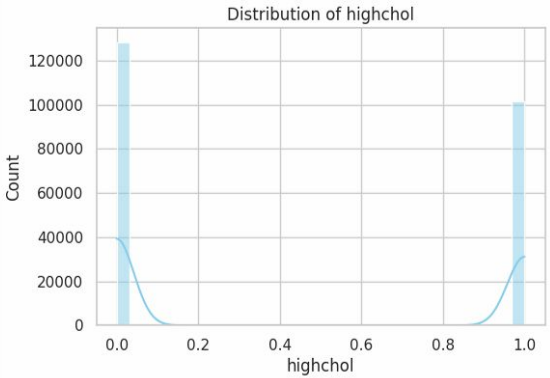


## Age vs Diabetes:

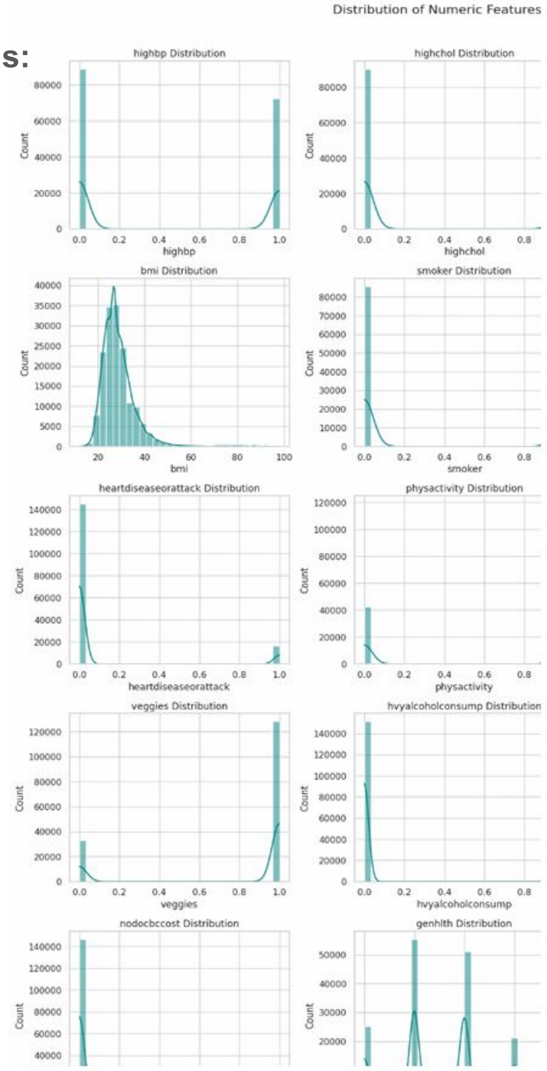


## Binary Features Comparison:





## Feature Distributions:



# Preprocessing Tasks:-

Objective: Prepare the dataset for modeling through cleaning, transformation, and standardization.

**1. Cleaning & Validation Steps**

Removed 23,968 duplicate records → ensures unbiased model training.

Verified no missing values → no imputation needed.

Dropped any constant columns → avoids redundant predictors.

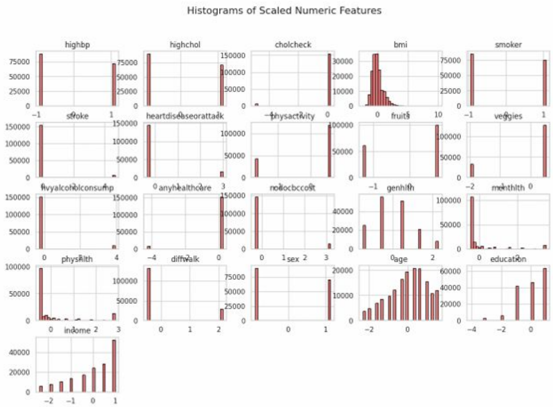
**2. Feature Preparation**

- Target Encoding:
    - Converted diabetes_012 into binary format (0 = non-diabetic, 1 = diabetic).
    - Ensures compatibility with binary classification models.
- Numeric Feature Scaling:
    - Applied StandardScaler for normalization.
    - Why? Prevents features with large numeric ranges from dominating smaller ones.
    - Effect: Improves gradient-based model performance & convergence speed.

**3. Train-Test Split**

- Split into 80% train and 20% test sets (stratified by diabetes status). → Maintains class balance across datasets.

**Visualization During Preprocessing:-**



# Hypothesis Tests:-

Based on our Exploratory Data Analysis, we identified High Blood Pressure and BMI as strong potential indicators. We formulated two hypotheses to statistically test these observations.

**Test 1: High Blood Pressure**

- Is there a statistically significant association between having High Blood Pressure and having Diabetes?
- Methodology: Chi-Squared ($\chi^2$) Test of Independence.
  - This is the standard test to determine if an association exists between two categorical variables.
- Hypotheses:
  - Null Hypothesis ($\boldsymbol{H_0}$): High Blood Pressure and Diabetes status are independent (i.e., there is no association).
  - Alternative Hypothesis ($\boldsymbol{H_1}$): High Blood Pressure and Diabetes status are dependent (i.e., there is an association).

**Test 2: BMI**

- Is the mean Body Mass Index (BMI) significantly different for patients with diabetes compared to those without diabetes?
- Methodology: Independent Two-Sample T-Test.
  - This test is used to compare the means of a numerical variable between two independent groups.
- Hypotheses:
  - Null Hypothesis ($\boldsymbol{H_0}$): The mean BMI of the non-diabetes group is equal to the mean BMI of the diabetes group.
  - Alternative Hypothesis ($\boldsymbol{H_1}$): The mean BMI of the two groups is not equal.

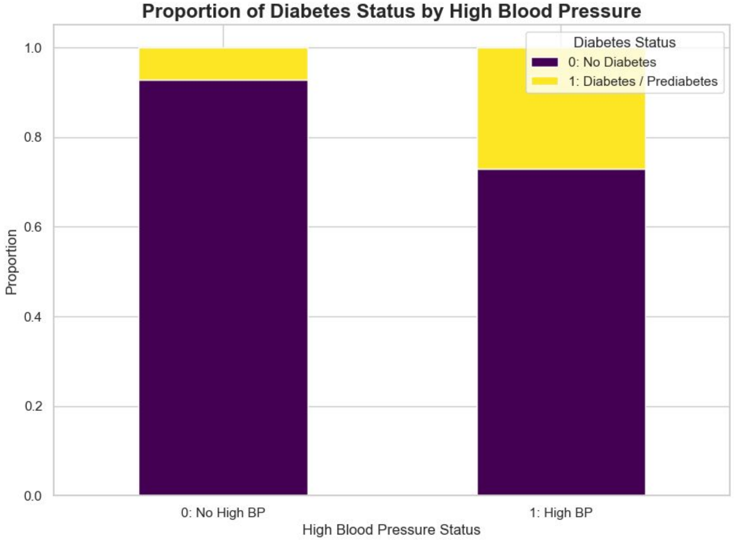
**Significance Level ($\boldsymbol{\alpha}$)**: For both tests, we set our significance level, $\alpha$ = 0.05.

# Experiments Conducted:-

We conducted the two tests on our dataset of 253,680 records.

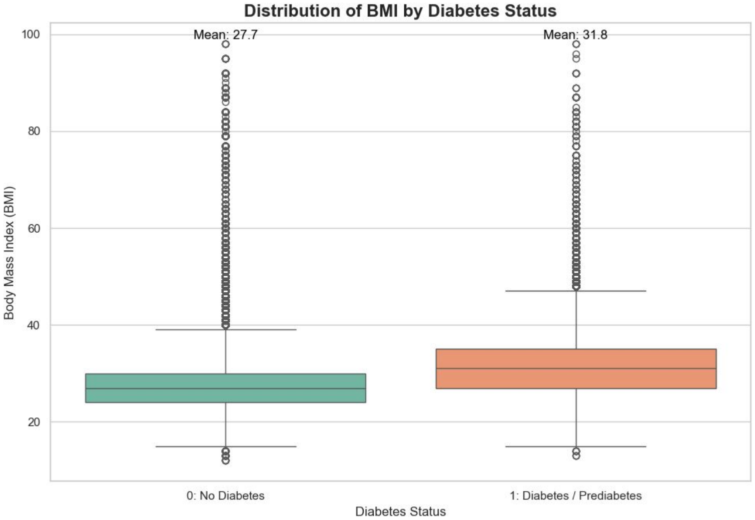
### Test 1: High BP ($\chi^2$ - Test)

- **Results:** Chi-Squared Statistic: 18537.57, p-value: 0.0
- **Conclusion:** The p-value (0.0) is less than our $\alpha$ (0.05). We reject the null hypothesis.
- **Validation:** The experiment confirms a highly significant association between high blood pressure and diabetes.
- **Visual Evidence:** As the plot below shows, the proportion of individuals with diabetes (yellow) is substantially larger in the 'High BP' group on the right.



### Test 2: BMI (T-Test)

- **Results:** Mean BMI (No Diabetes): 27.74, Mean BMI (Diabetes): 31.80, p-value: 0.0
- **Conclusion:** The p-value (0.0) is less than our $\alpha$ (0.05). We reject the null hypothesis.
- **Validation:** The experiment confirms a highly significant difference in mean BMI between the two groups.
- **Visual Evidence:** The box plot shows the entire BMI distribution for the diabetes group is shifted significantly higher, confirming it as a key indicator.


Distribution of BMI by Diabetes Status
Mean: 27.7
Mean: 31.8
Body Mass Index (BMI)
0: No Diabetes
1: Diabetes / Prediabetes
Diabetes Status