

# Diabetes Prediction Model



**TEAM:**  
**Data\_wizards**

**Anthony Ngatia**

**Naomi Rotich**

**Elsie Nduta**

**Mitchele Okubasu**

# Content Outline

Topics for discussion

**01** Project Overview

**02** Project Roadmap

**03** Introduction

**04** Visualizations

**05** Modelling

Key Challenges

Conclusion & Recommendation



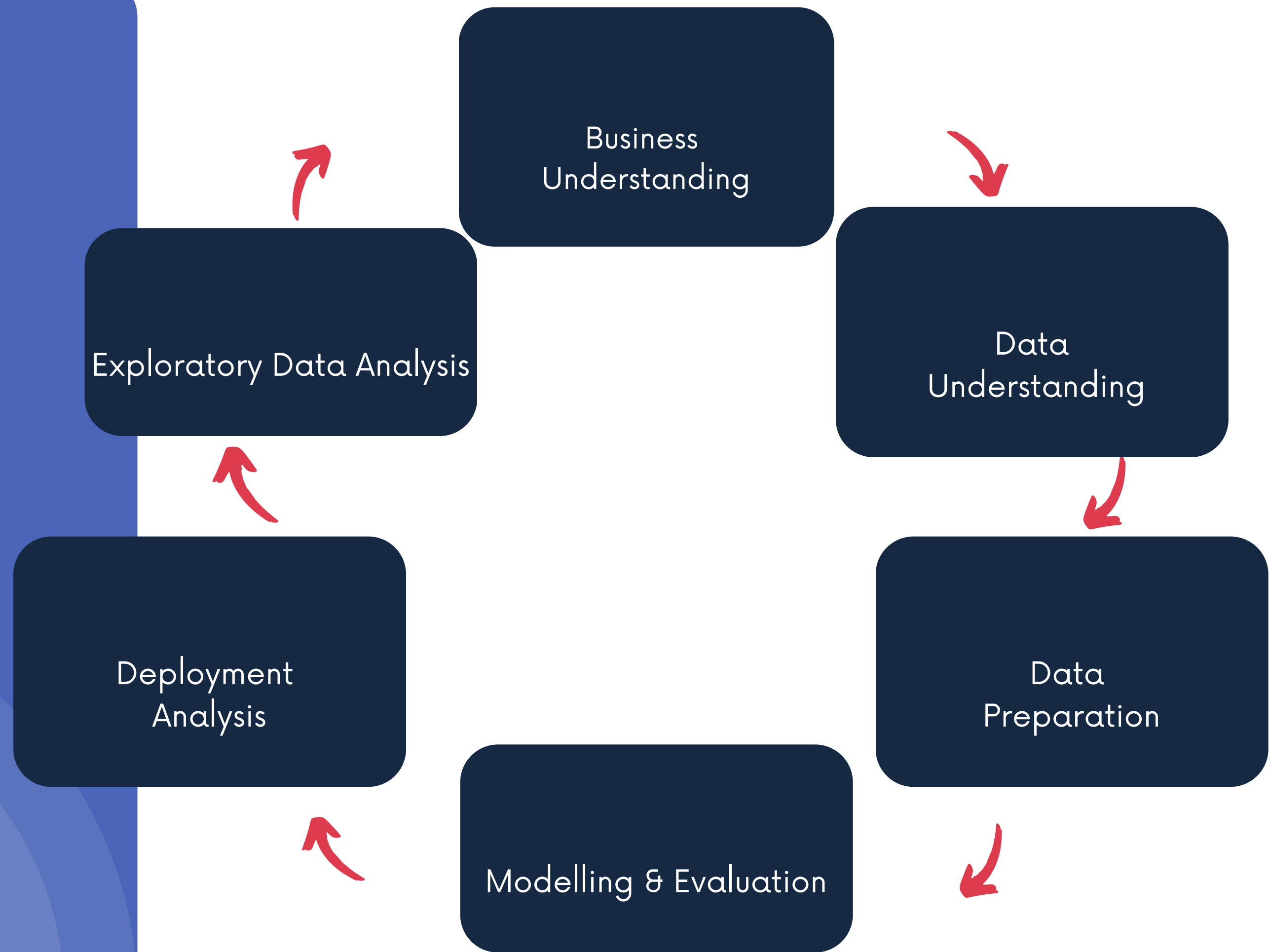
# Project Overview

- Dataset: Extensive medical records of patients, including demographics, clinical measurements, and medical history.
- Methodology: CRISP-DM framework – Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.
- Tools: Google Colab, Jupyter Notebook, Github, Spyder(Anaconda)etc.



# Project Roadmap

CRISP-DM Process





# Introduction

- **Objective:**

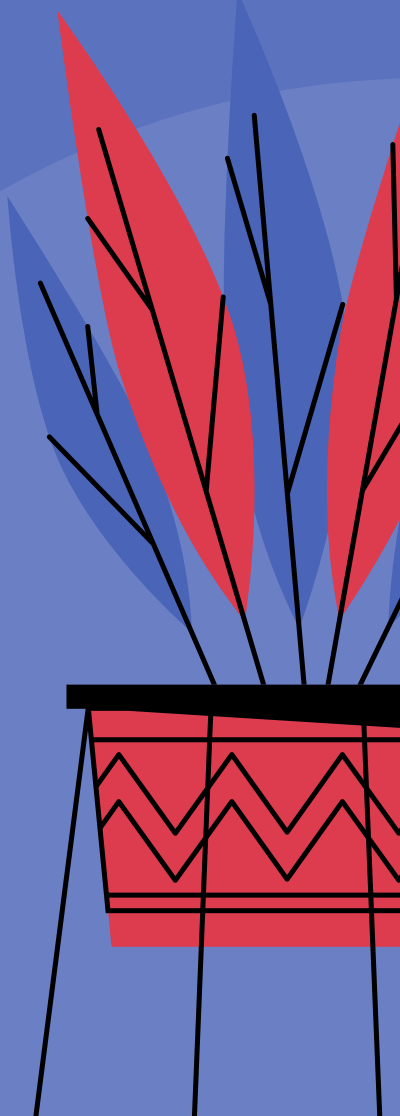
To develop an accurate and reliable predictive model for early detection of diabetes.

- **Why it matters:**

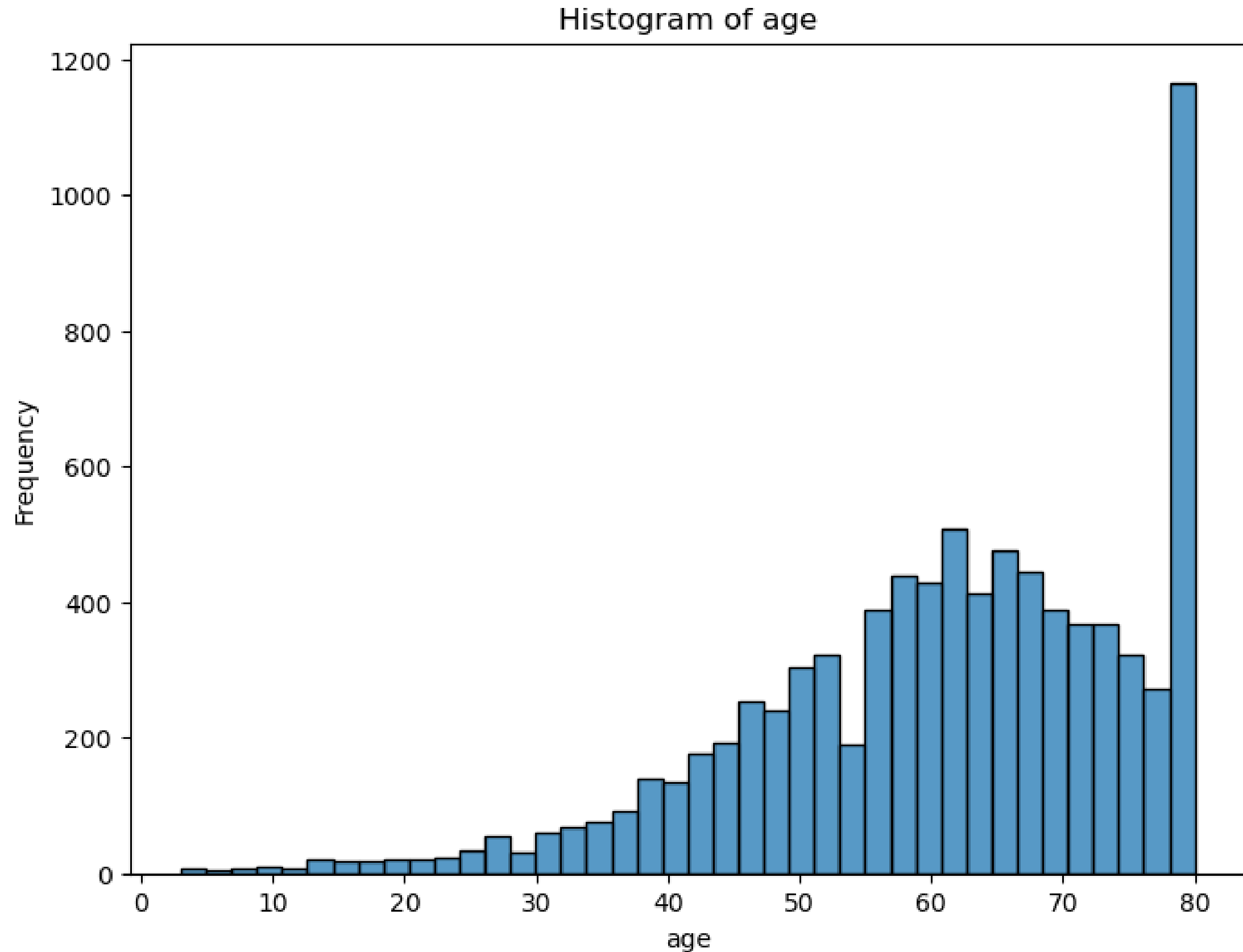
Early detection can lead to timely intervention and better patient outcomes.

- **Data-driven approach:**

Utilizing advanced data analysis and machine learning techniques.

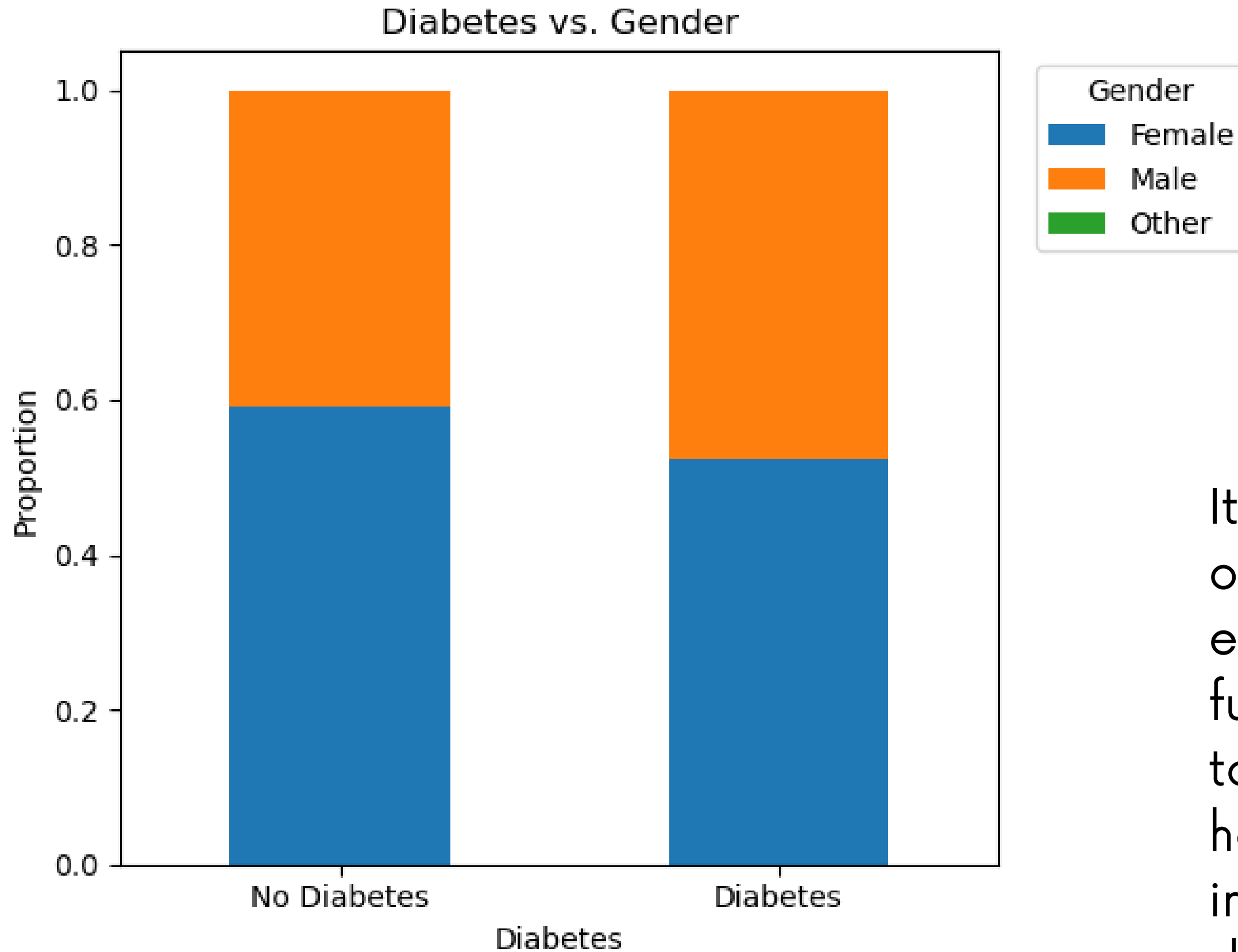


# From what age does Diabetes occurrence increase?



*It appears that from the age of 40 onwards most people have a likelihood of having diabetes*

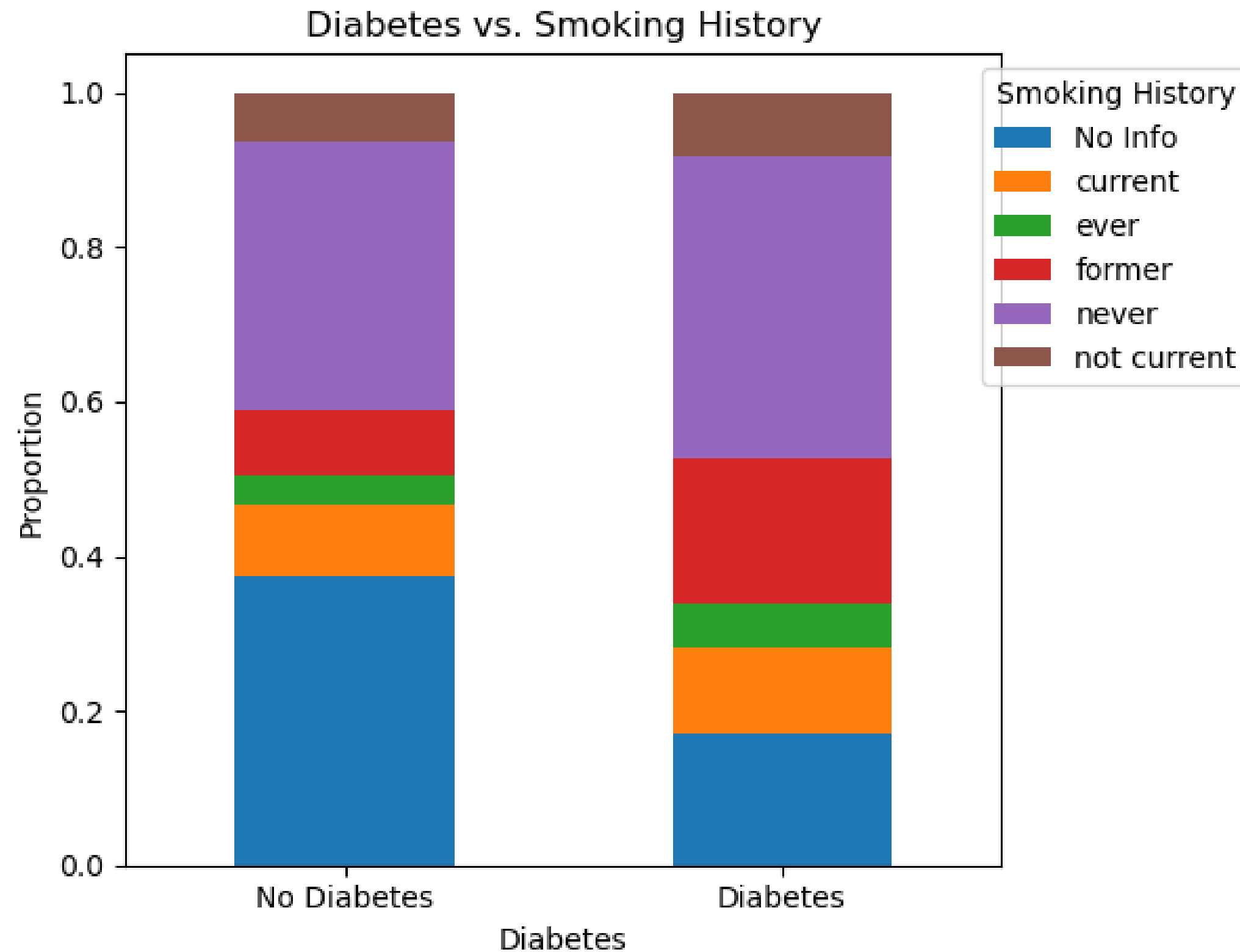
# What Gender has more occurrences of Diabetes ?



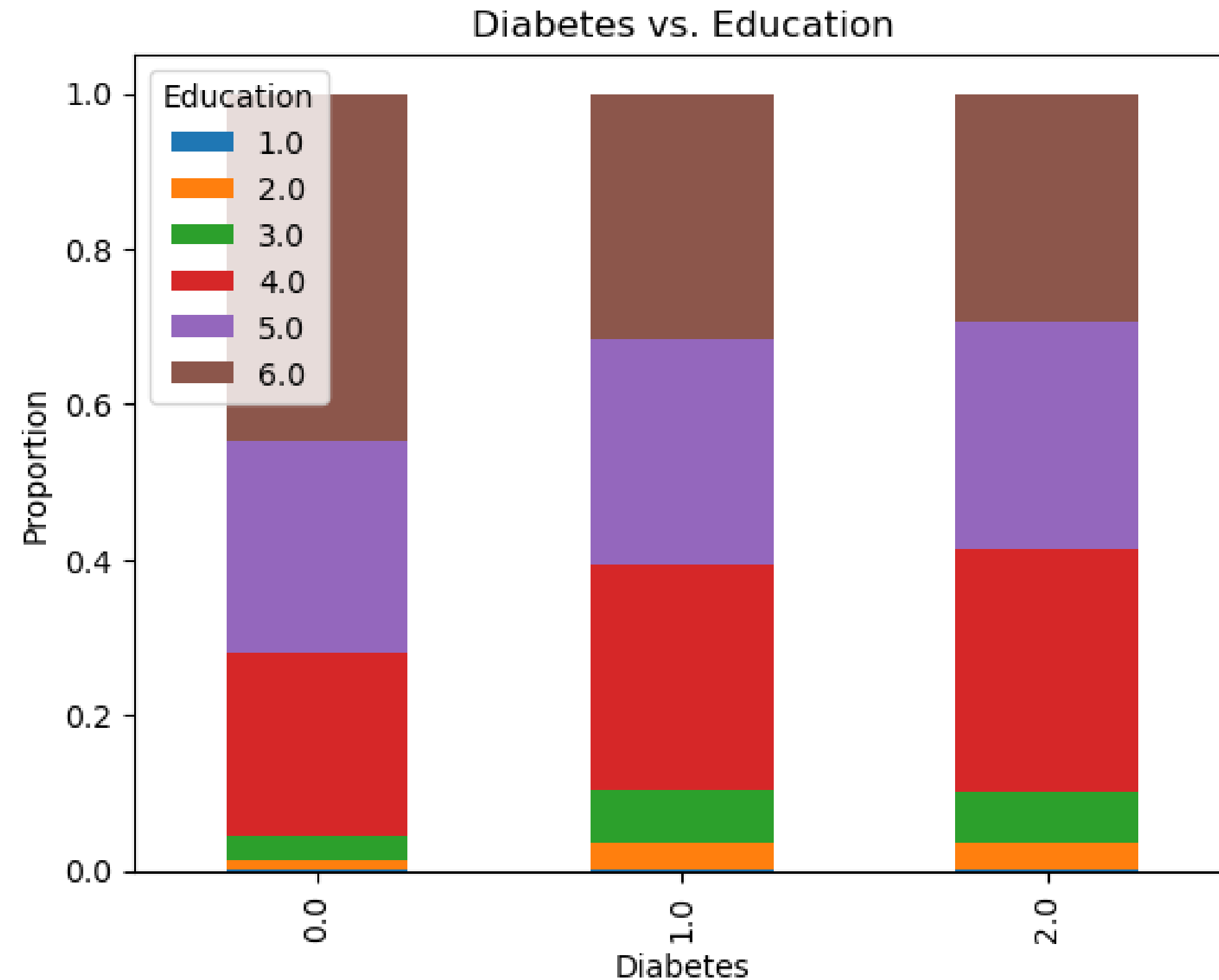
It was observed there were more cases of diabetes amongst men than women, especially after the age of 50. On further research, this could be attributed to the steady decline in the testosterone hormone, which has been linked to increased chances in diabetes diagnosis.



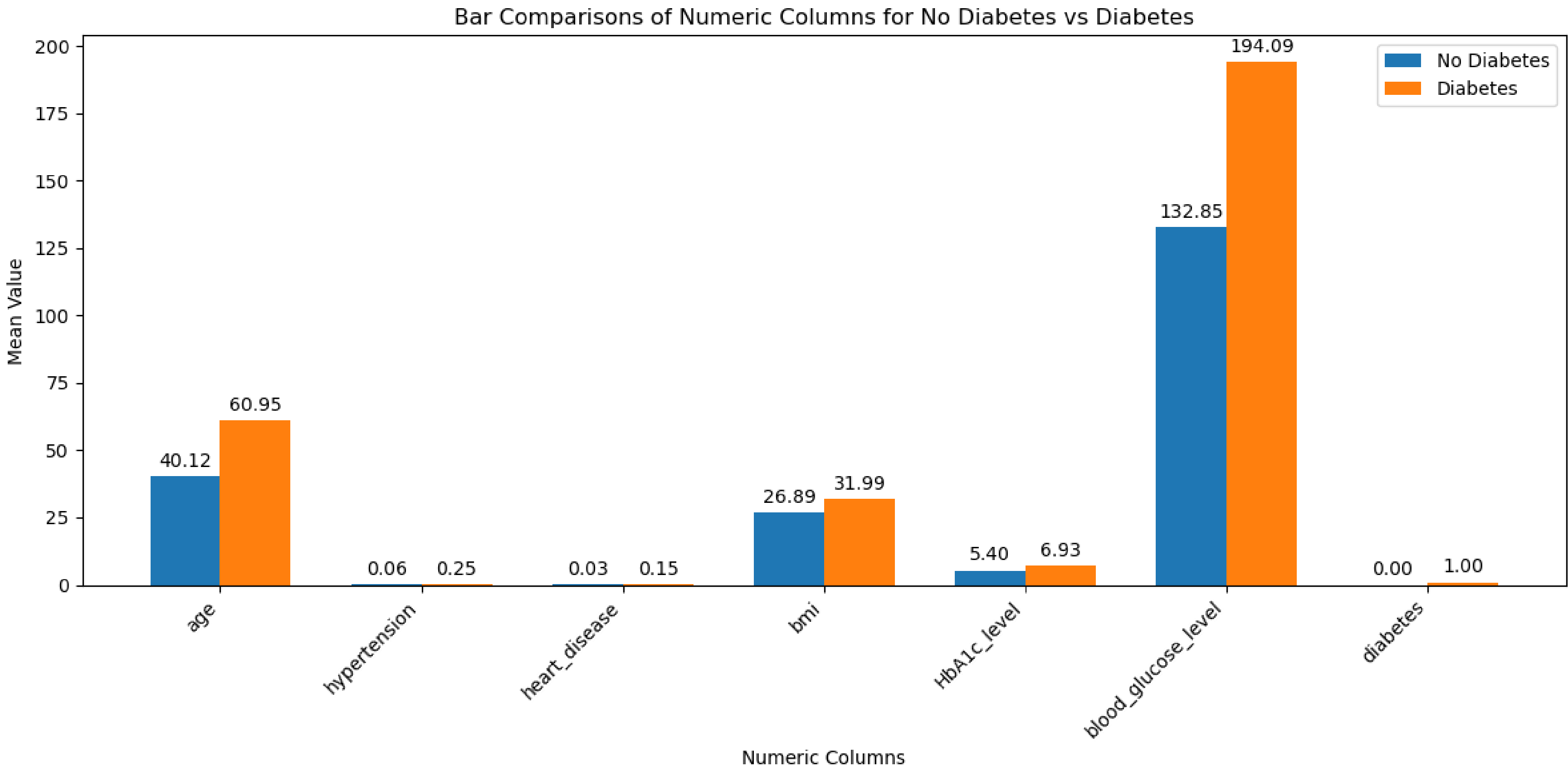
# What's the smoking comparison between those with diabetes and no\_Diabetes ?



# What's the comparison of education levels and those with diabetes and no\_Diabetes ?

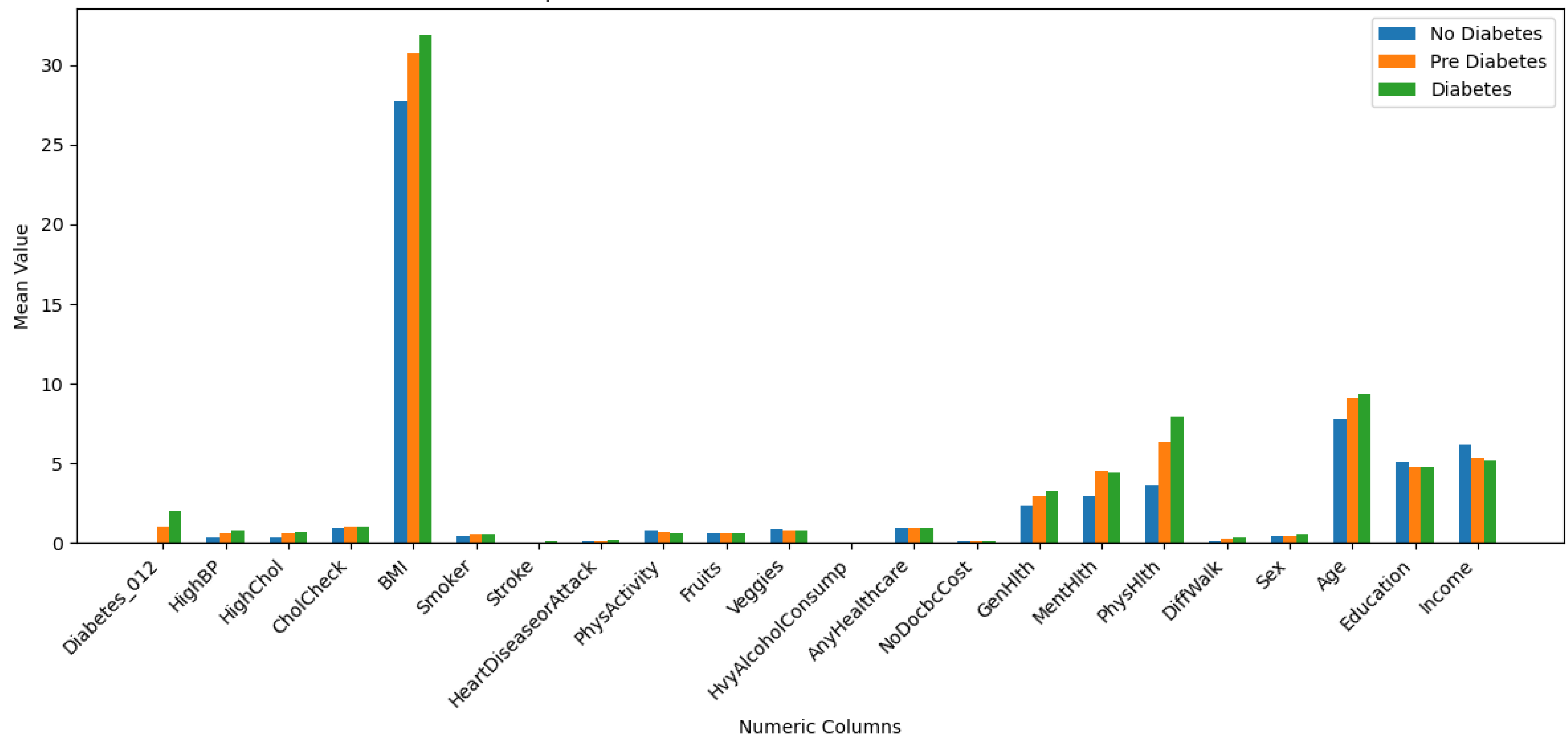


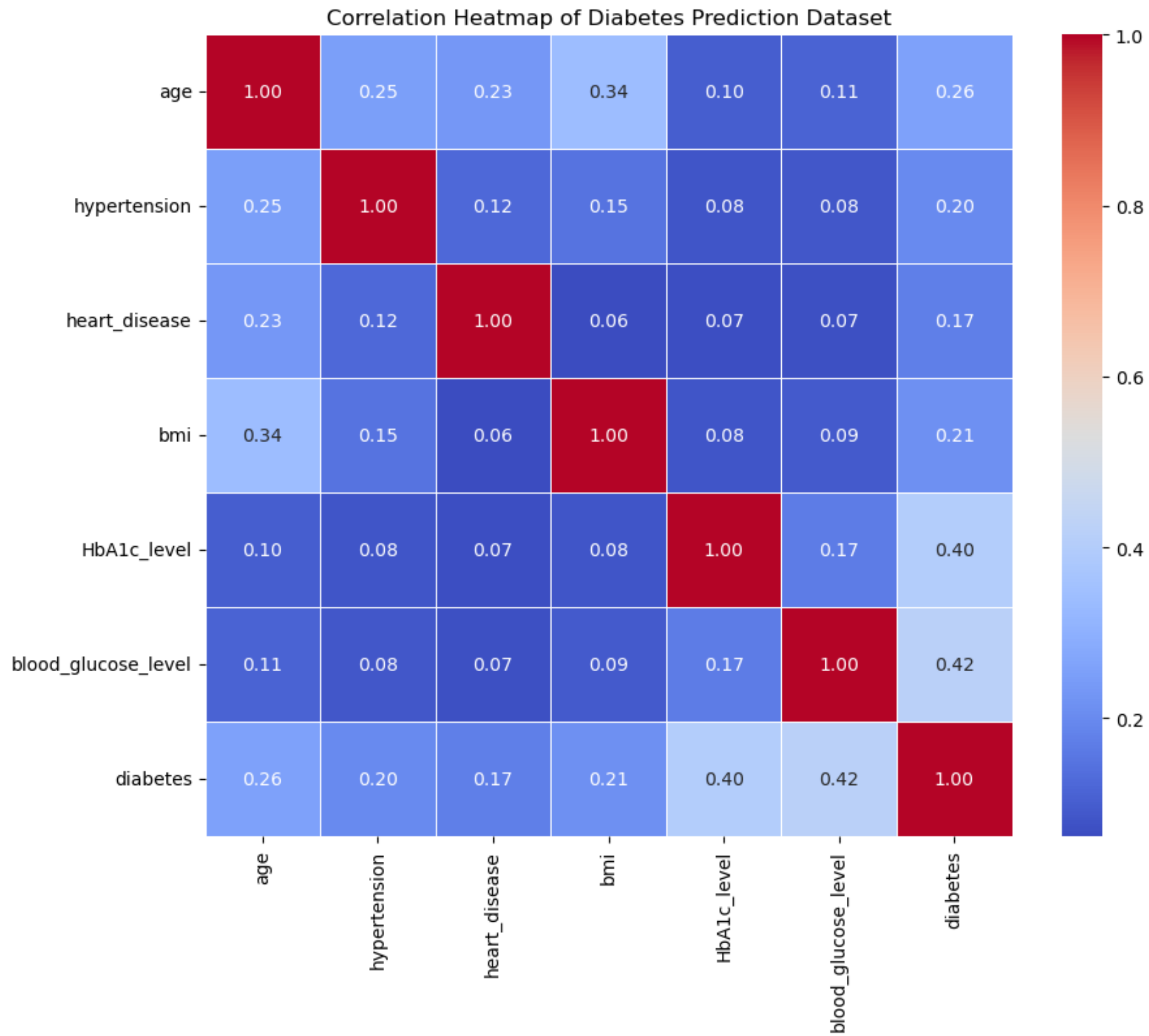
# Comparisons between those with diabetes and no\_Diabetes ?



# Comparisons between those with diabetes and no\_Diabetes ?

Bar Comparisons of Numeric Columns for Pre Diabetes vs Diabetes





**Correlation of  
all the factors**

# Modelling

XGBOOST

CNN MODEL

SVM MODEL

Decision  
Tree



# Model Evaluation



## XGBOOST

Precision :

**0.99**

Recall :

**0.84**

F1 Score :

**0.84**

## CNN

Precision :

**0.98**

Recall :

**0.67**

F1 Score :

**0.80**

## SVM MODEL

Precision :

**0.89**

Recall :

**0.89**

F1 Score :

**0.89**

## Decision Tree

Precision :

**0.96**

Recall :

**0.96**

F1 Score :

**0.96**

**Function that prompts user to input gender, age,...and returns a prediction based on the input**

**model : Decision Tree**

```
predict_diabetes(model)
```

```
Enter gender (Male/Female): Female
```

```
Enter age: 60
```

```
Enter hypertension (0 for No, 1 for Yes): 0
```

```
Enter heart disease (0 for No, 1 for Yes): 0
```

```
Enter smoking history (never/No Info/current/former/ever/not current): former
```

```
Enter BMI: 9
```

```
Enter HbA1c level: 6.5
```

```
Enter blood glucose level: 200
```

```
Based on the input, the person is predicted to NOT have diabetes.
```

```
predict_diabetes(model)
```

```
Enter gender (Male/Female): Female
```

```
Enter age: 60
```

```
Enter hypertension (0 for No, 1 for Yes): 1
```

```
Enter heart disease (0 for No, 1 for Yes): 0
```

```
Enter smoking history (never/No Info/current/former/ever/not current): ever
```

```
Enter BMI: 27
```

```
Enter HbA1c level: 7
```

```
Enter blood glucose level: 150
```

```
Based on the input, the person is predicted to have diabetes.
```

# Key Challenges

**Imbalanced dataset**

**Model generalization to  
diverse patient populations**

# CONCLUSIONS

The gender with more occurrences of diabetes are Females.

Most smokers seem to quit smoking after being diagnosed with diabetes this is because smoking habits have a probability of causes smoking is a significant risk factor for developing type 2 diabetes.

As income increases there are lesser people with diabetes these could be correlated with their ability to lead healthier lifestyles .

Those with higher level of education also have lesser occurrences of diabetes this could be associated by their ability to do better research or an understanding of the importance to lead healthier lifestyles.



# RECOMMENDATIONS

Collection of a more balanced data(no diabetes and diabetes) would produce a better performing model that wont be biased after training.

