



# "Diabetes" dataset

## **supervised Learning Machine**

By Hajar Kaddouri



# Structure

**1-Exploratory Data Analysis (EDA)**

**2-Data Preprocessing**

**3-Model Evaluation**

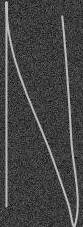
**4-Insights and Findings**

**5-Conclusion**



# Introduction

Full supervised learning machine is a learning project on a "Diabetes" dataset. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset.







# **Part I - EDA - Exploratory Data Analysis**

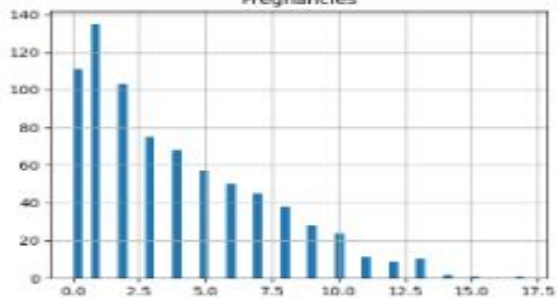


# EDA & Data Preprocessing:

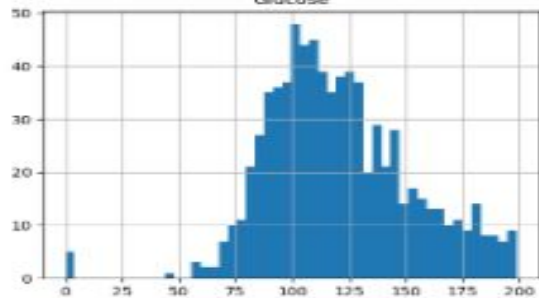
- **Data Understanding:** The dataset contains 9 columns, namely 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', and 'Outcome'.
- **Data Cleaning:** Some columns, specifically 'BMI' and 'SkinThickness', contained zero values, which are not meaningful in this context. These zero values were replaced with the respective column means.
- **Correlation Analysis and Averages:** Various visualization techniques have been used to show the distribution, correlations, and averages of the variables in the dataset.



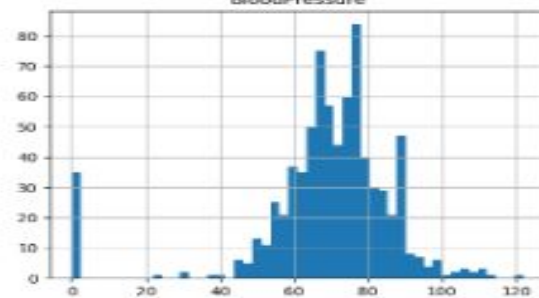
Pregnancies



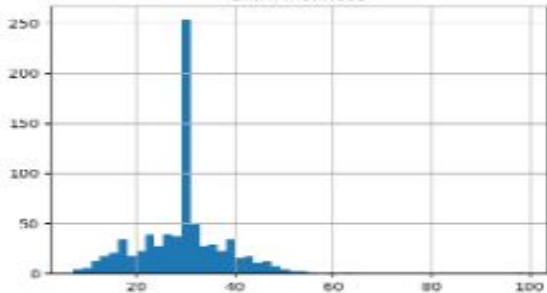
Glucose



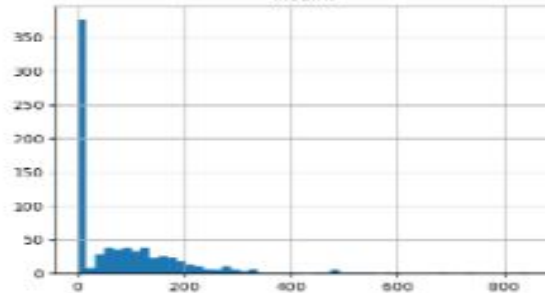
BloodPressure



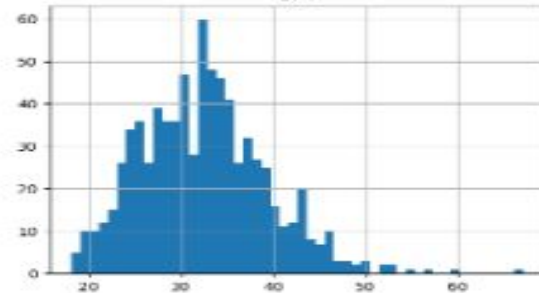
SkinThickness



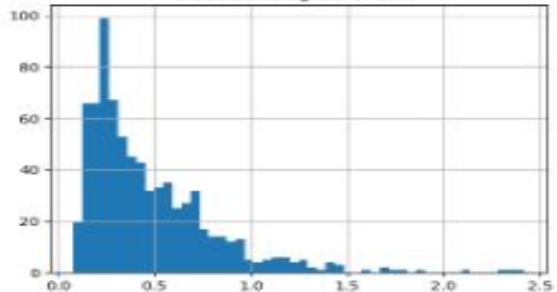
Insulin



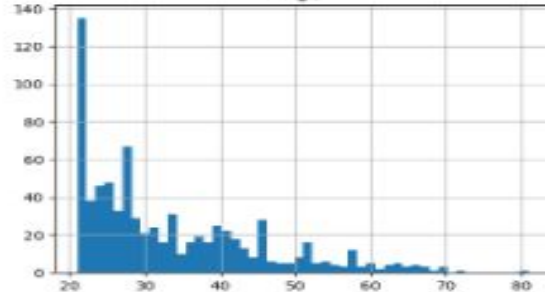
BMI



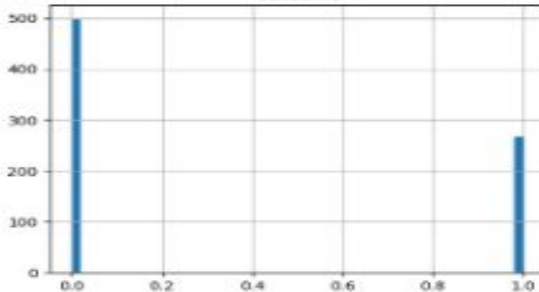
DiabetesPedigreeFunction

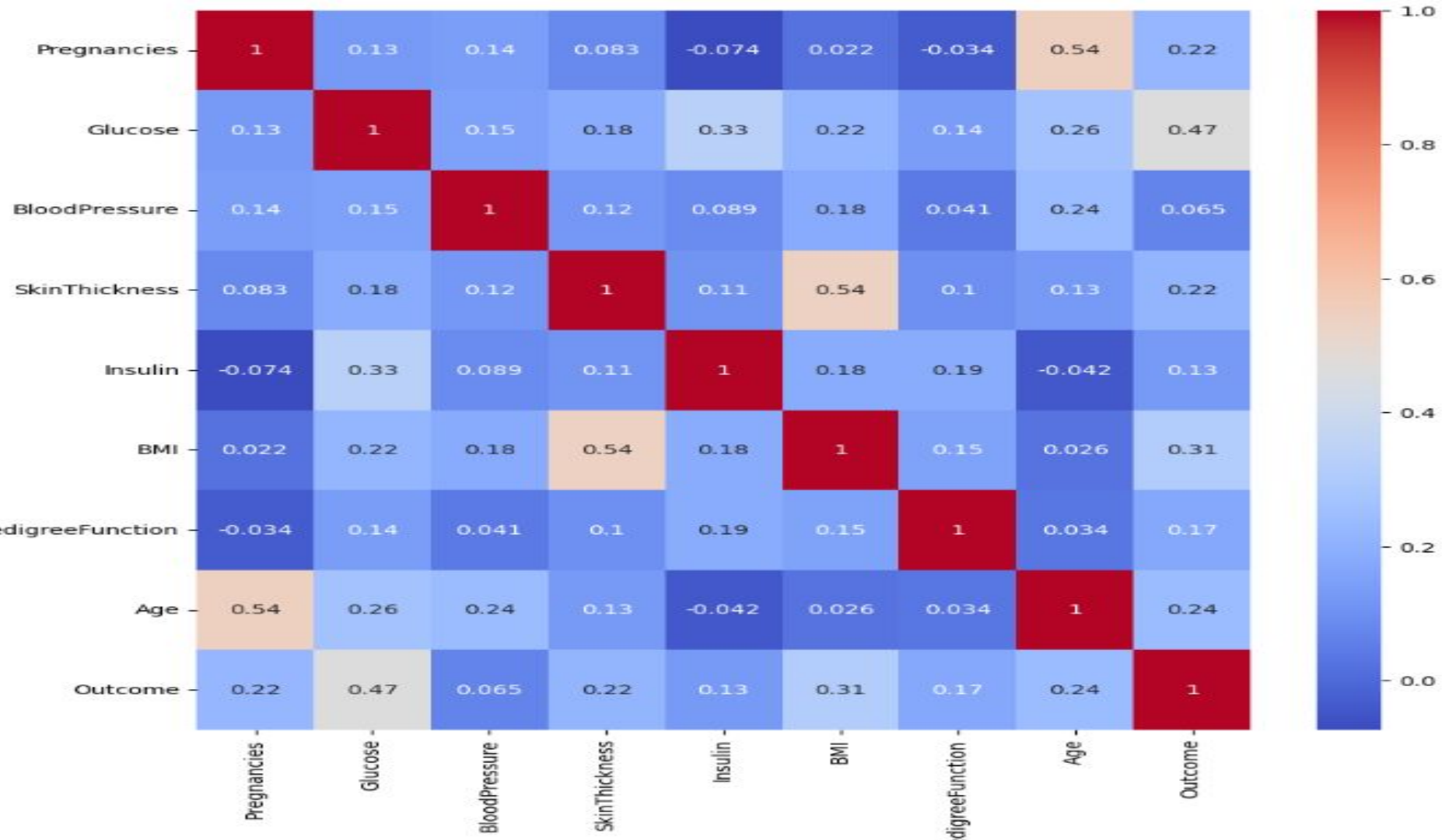


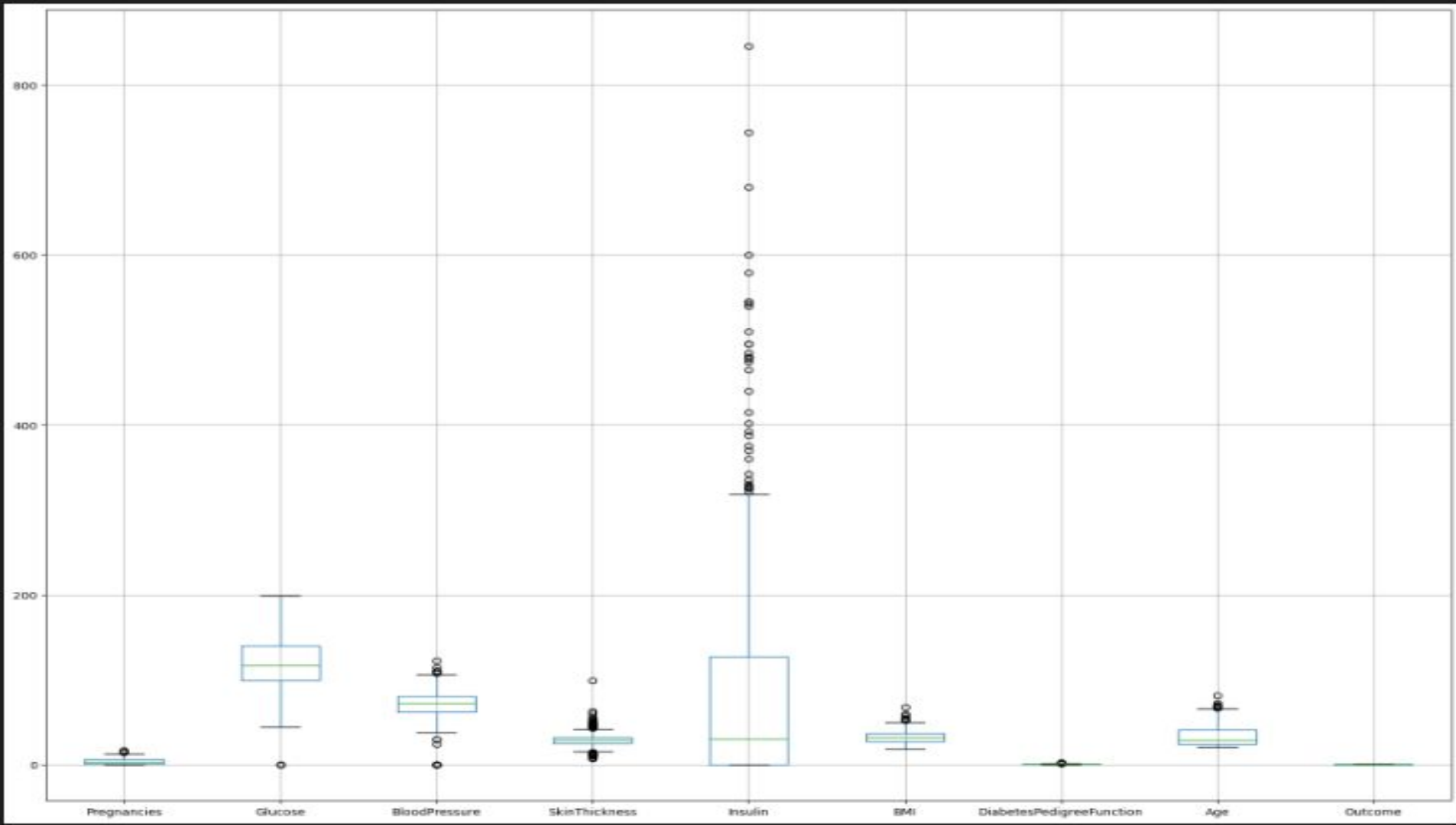
Age



Outcome









### **#Average Glucose Level for Individuals With and Without Diabetes:**

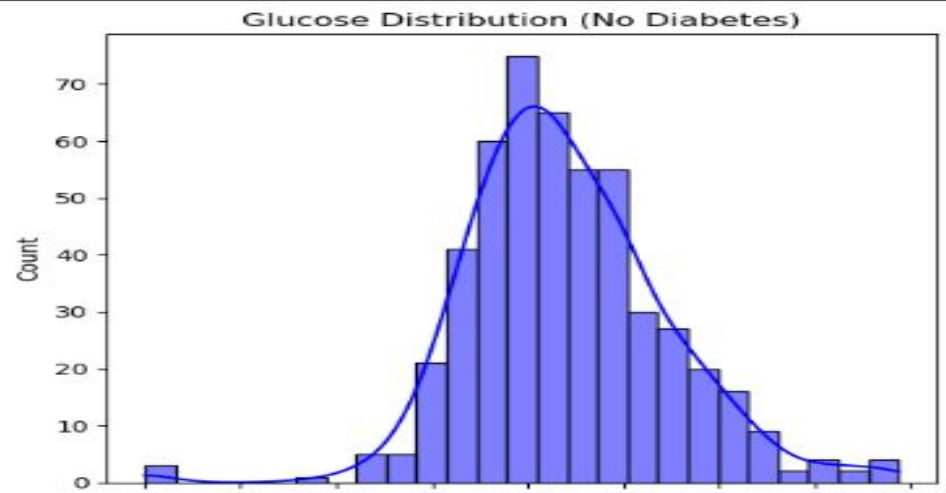
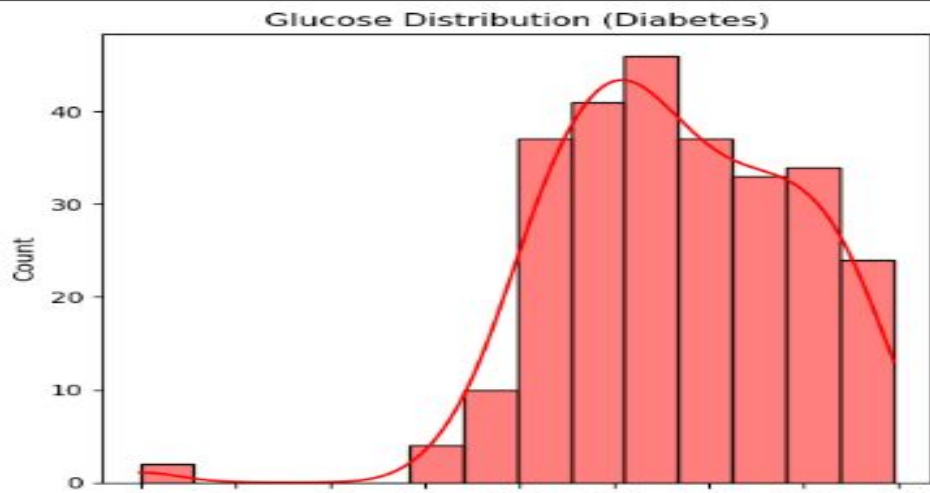
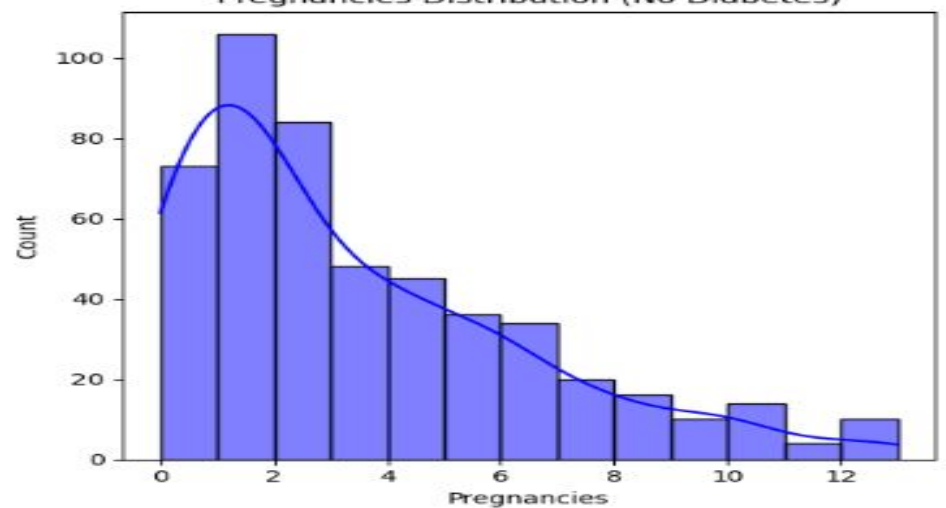
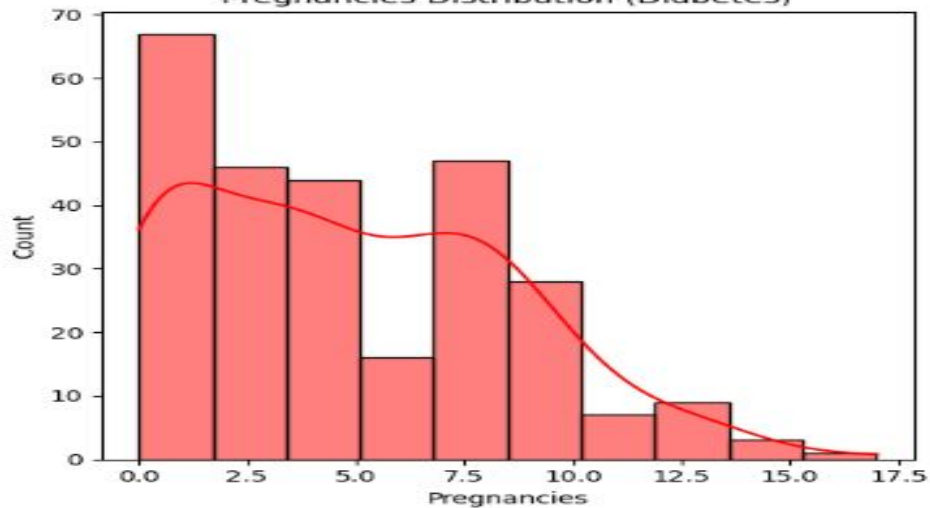
- ❖ **Average glucose level with diabetes: 141.25746268656715**
- ❖ **Average glucose level without diabetes: 109.98**

### **#Average BMI for Individuals With and Without Diabetes:**

- ❖ **Average BMI with diabetes: 35.38475719158496**
- ❖ **Average BMI without diabetes: 30.888434346103004**

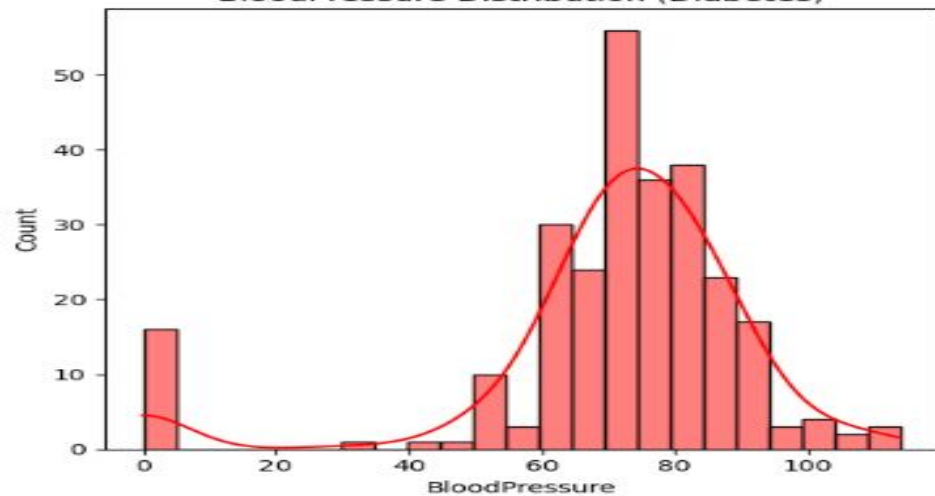
### **#Distribution of Predictor Variables for Individuals With and Without Diabetes:**

- ❖ **(See the next slides)**

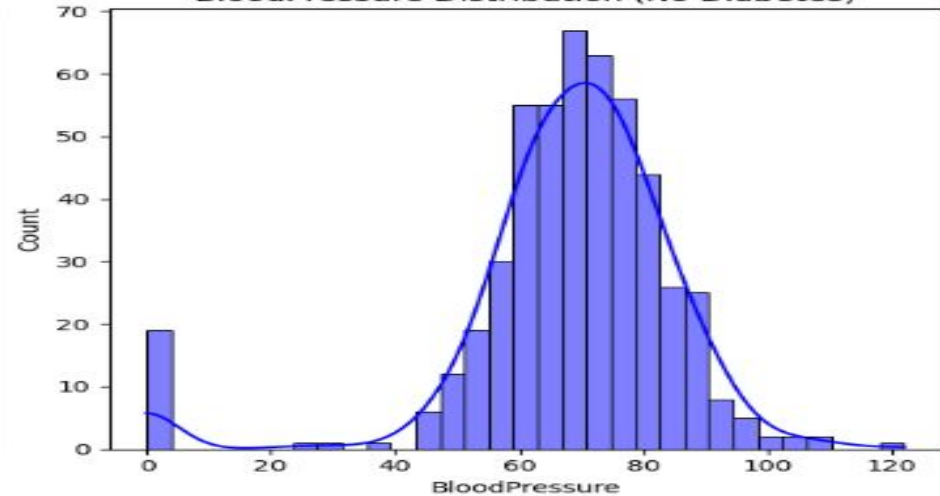




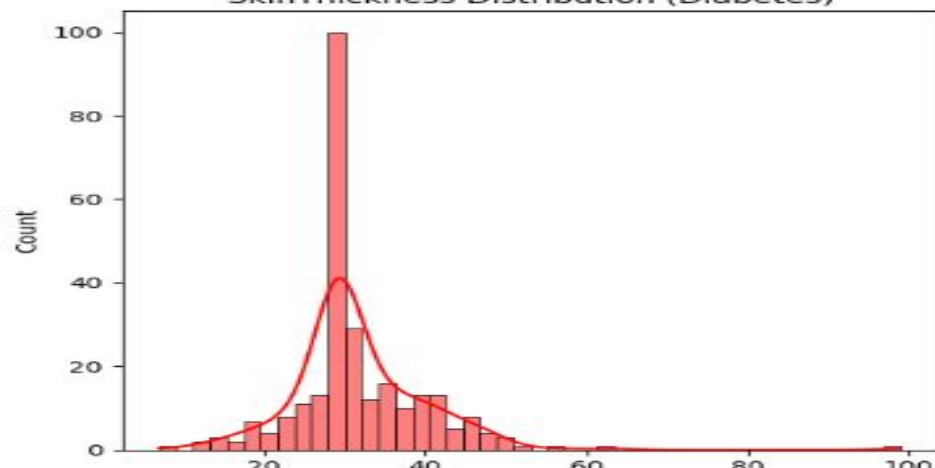
BloodPressure Distribution (Diabetes)



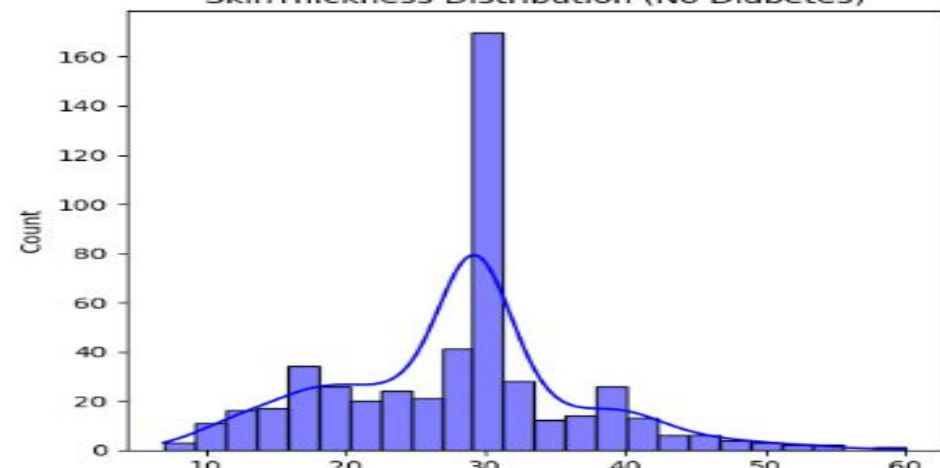
BloodPressure Distribution (No Diabetes)



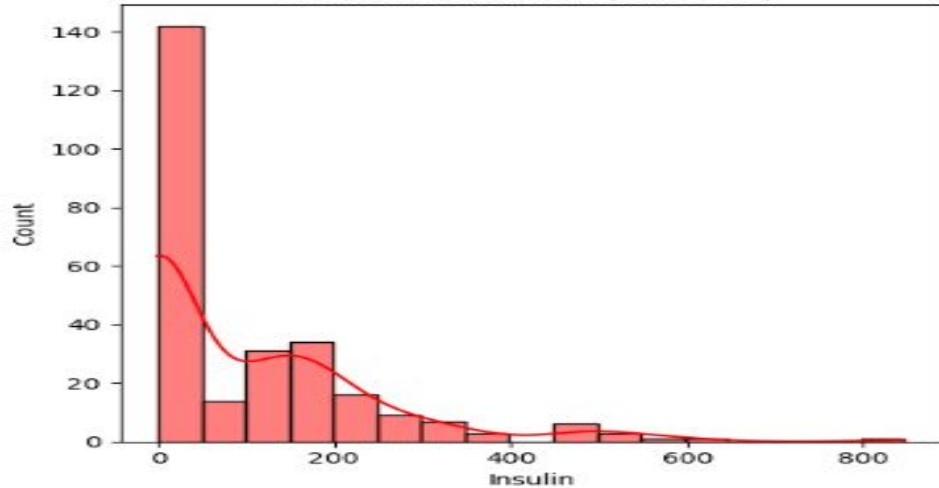
SkinThickness Distribution (Diabetes)



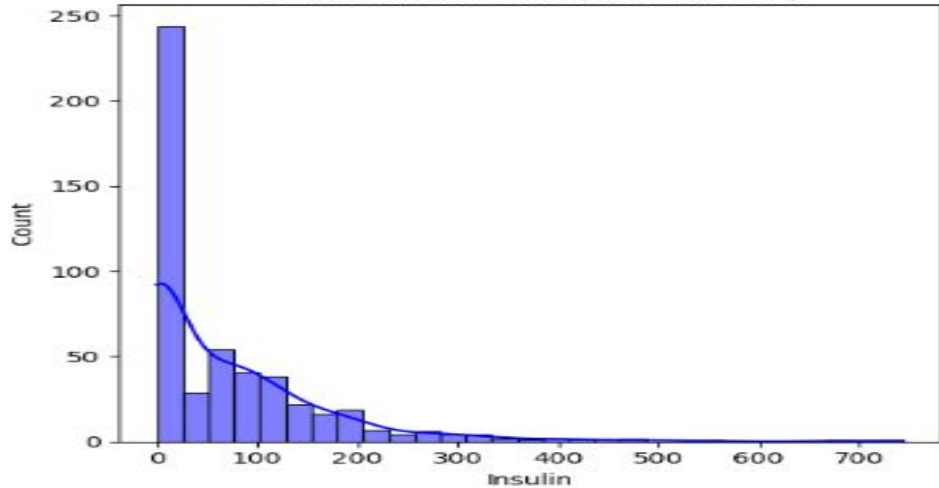
SkinThickness Distribution (No Diabetes)



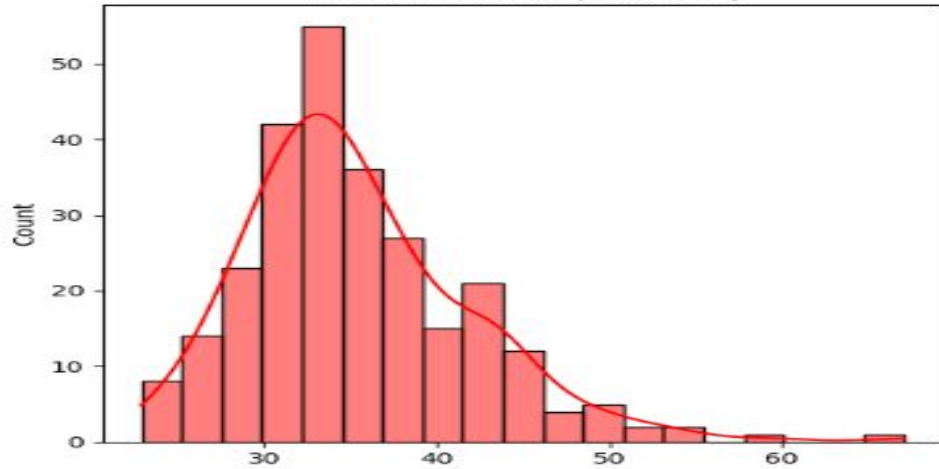
Insulin Distribution (Diabetes)



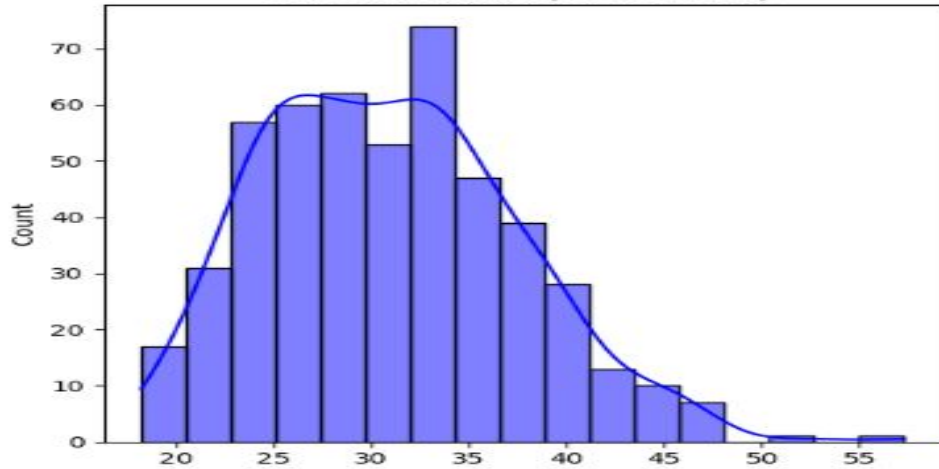
Insulin Distribution (No Diabetes)



BMI Distribution (Diabetes)

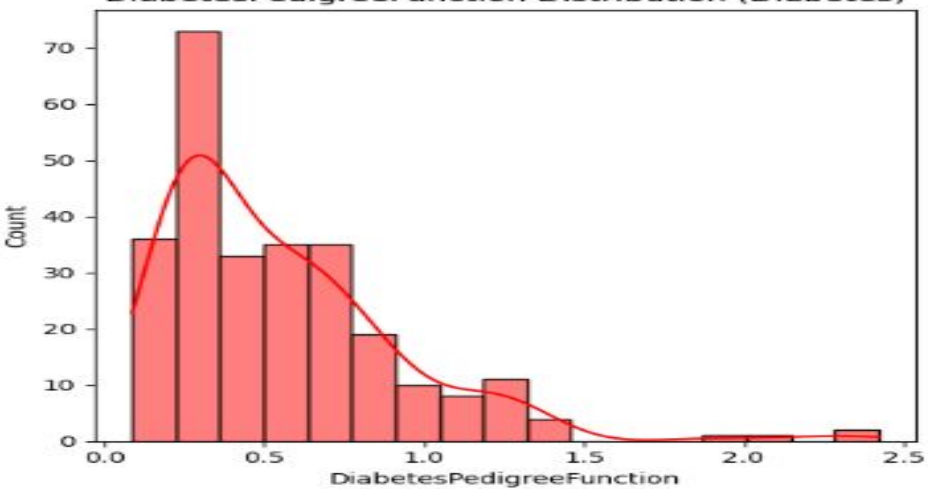


BMI Distribution (No Diabetes)

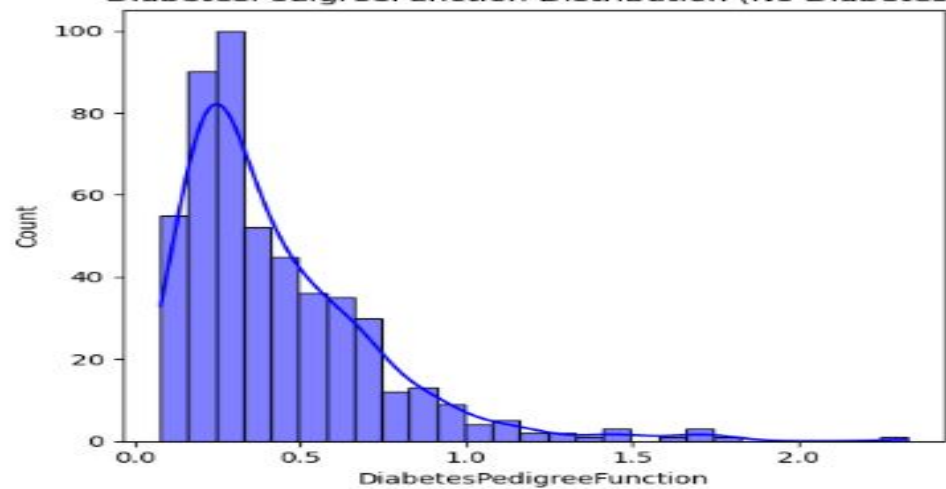




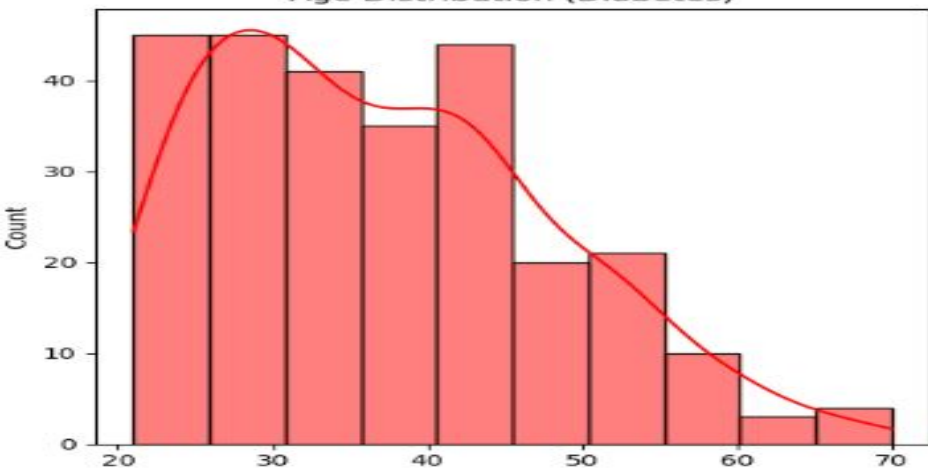
DiabetesPedigreeFunction Distribution (Diabetes)



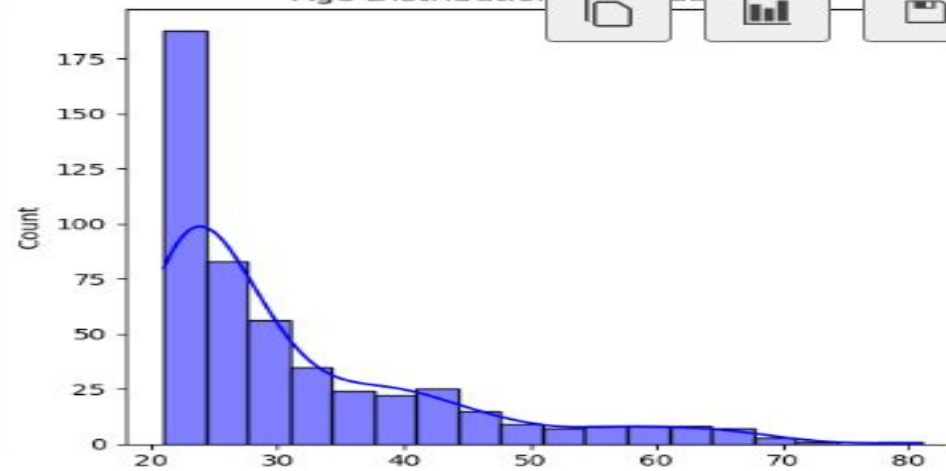
DiabetesPedigreeFunction Distribution (No Diabetes)



Age Distribution (Diabetes)

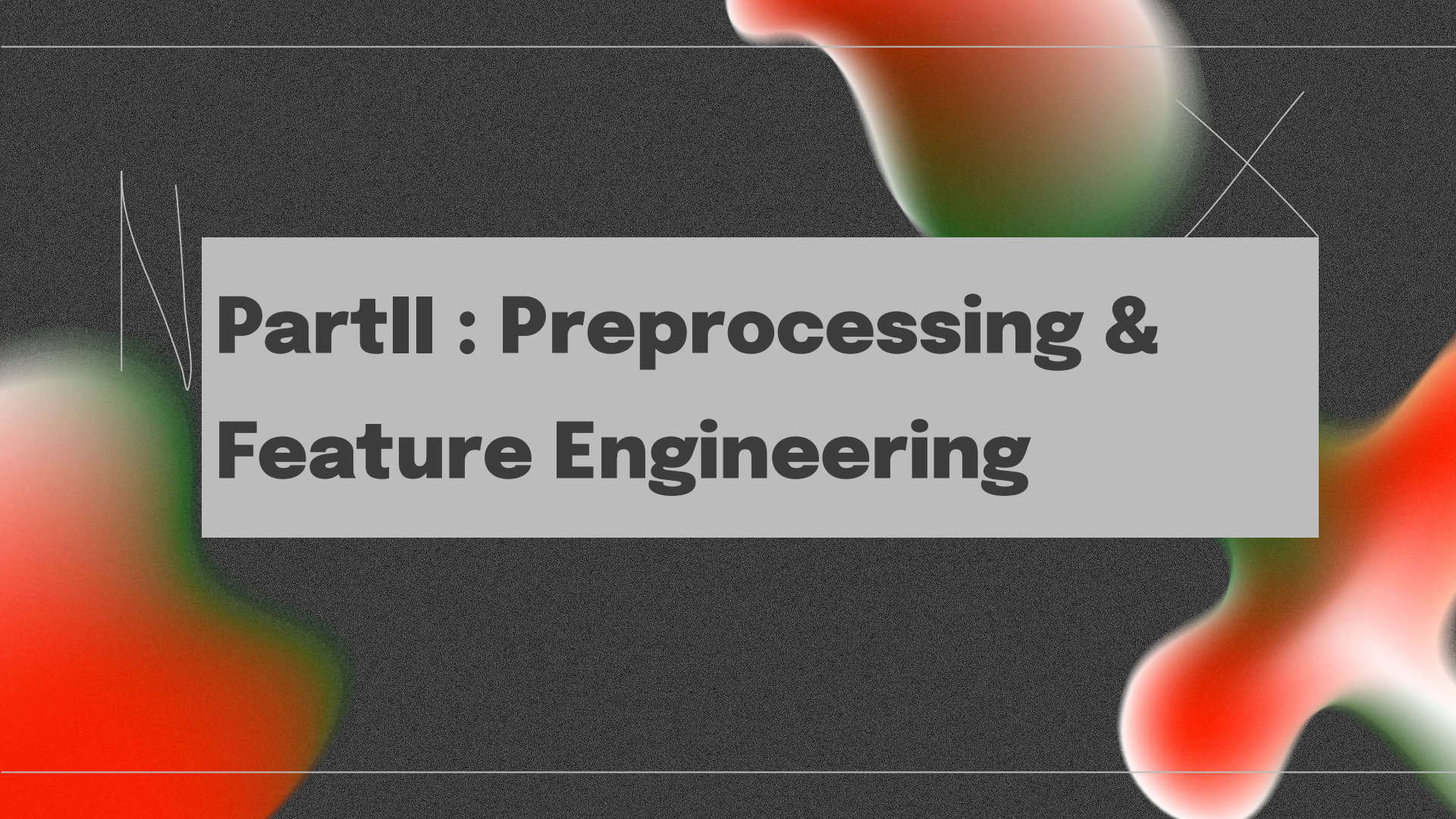


Age Distribution (No Diabetes)



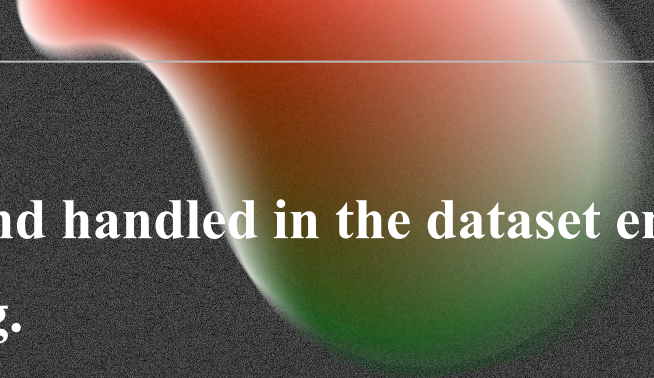
**Correlation Between Variables: The EDA reveal that some predictor variables like glucose level, bloodPressure, and BMI were found to have a strong correlation with the diabetes outcome. This suggests these factors play a significant role in diabetes diagnosis.**





# **PartII : Preprocessing & Feature Engineering**



- 
- ❑ **#Outliers : Outliers were found and handled in the dataset ensuring a cleaner dataset for model building.**
  - ❑ **#Feature ingeniering performed by creating new Bmi categories**
  - ❑ **#Scaling and normalization of the data for more accuracy**
  - ❑ **#Data Imbalance: The original dataset was imbalanced, with the majority class being those without diabetes. This imbalance was handled using oversampling technique, which helped in creating a more balanced dataset for model training.**





# **PartII :Training ML Model**





---

The selected models:

- LogisticRegression
- RandomForest



# Evaluation

- Evaluation metrics for logistic model : accuracy 0.725 precision : 0.7346938775510204 recall : 0.7128712871287128 f1\_logistic: 0.7236180904522613 roc\_auc: 0.7251225122512251 Evaluation
- Evaluation metrics for forest model : accuracy 0.805 precision : 0.78181818181819 recall : 0.8514851485148515 f1\_logistic: 0.8151658767772512 roc\_auc: 0.8045304530453046

the best-performing model based on the evaluation metrics is  
Random Forest model (accuracy 0.805)



# Conclusion

- Best Score: 0.8640000000000001 Best Hyperparameters: {'max\_depth': None, 'max\_features': 'log2', 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 50}
- Model Performance: Among the models developed, the Random Forest Classifier showed the best performance. After hyperparameter tuning, it delivered superior results compared to the Logistic Regression model based on evaluation metrics such as accuracy, precision, recall, and F1-score.





**THANKS**



# RESOURCES

- <https://github.com/hajar-kaddouri/ml-project-supervised-learning>