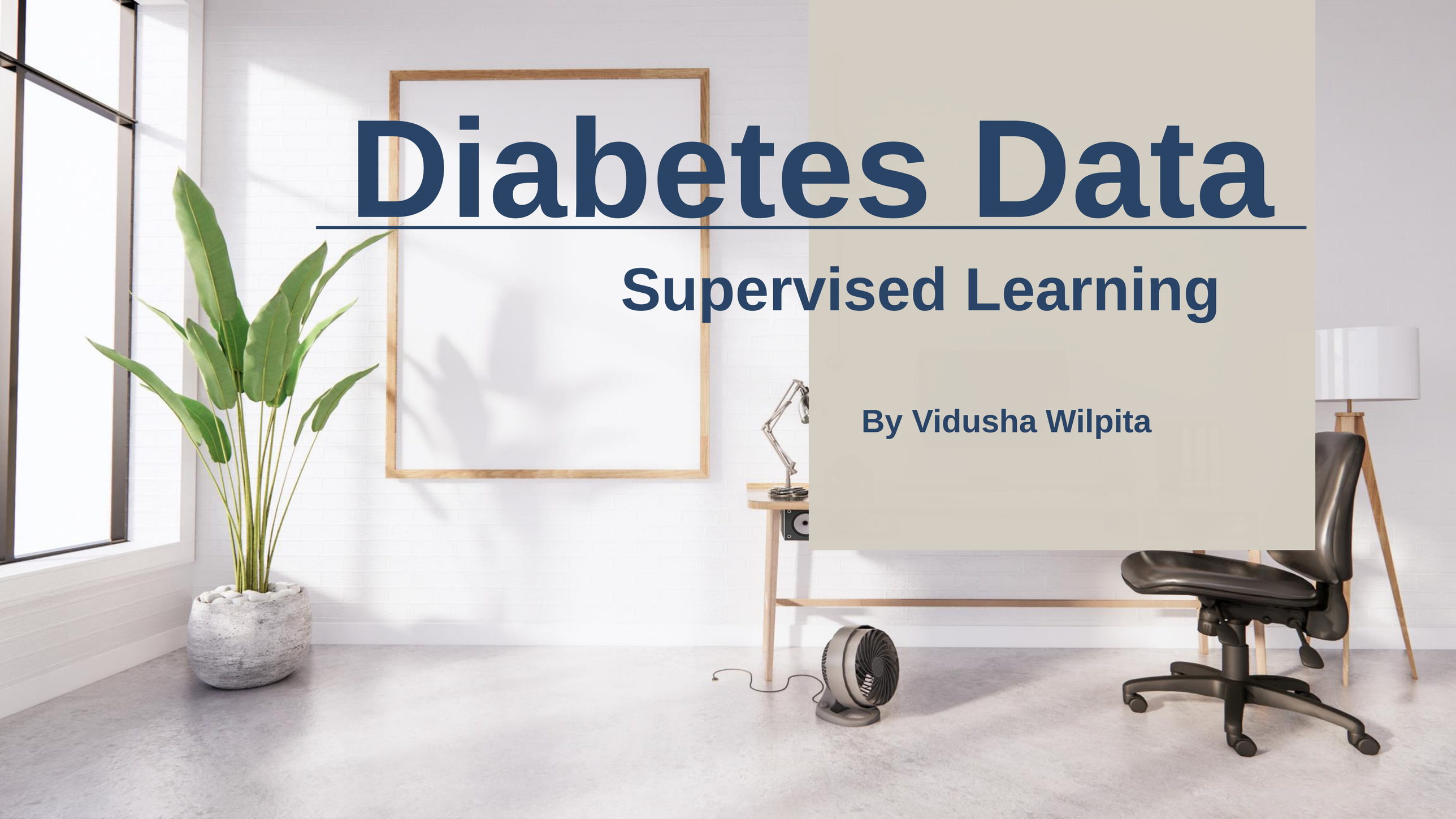


Diabetes Data

Supervised Learning

By Vidusha Wilpita

Contents

Introduction

- Data about 768 individuals (Numerical Data only)

Target Variable	Predictors
• Outcome	• Pregnancies • Glucose • Blood Pressure • Skin Thickness • Insulin • BMI • Diabetes Pedigree Function • Age

Flow Structure

Connected diabetes.csv to analyze the data

Created different visualizations to learn more about the data, Done data cleaning and preprocessing

Created classifier Models to fit and predict the outcome

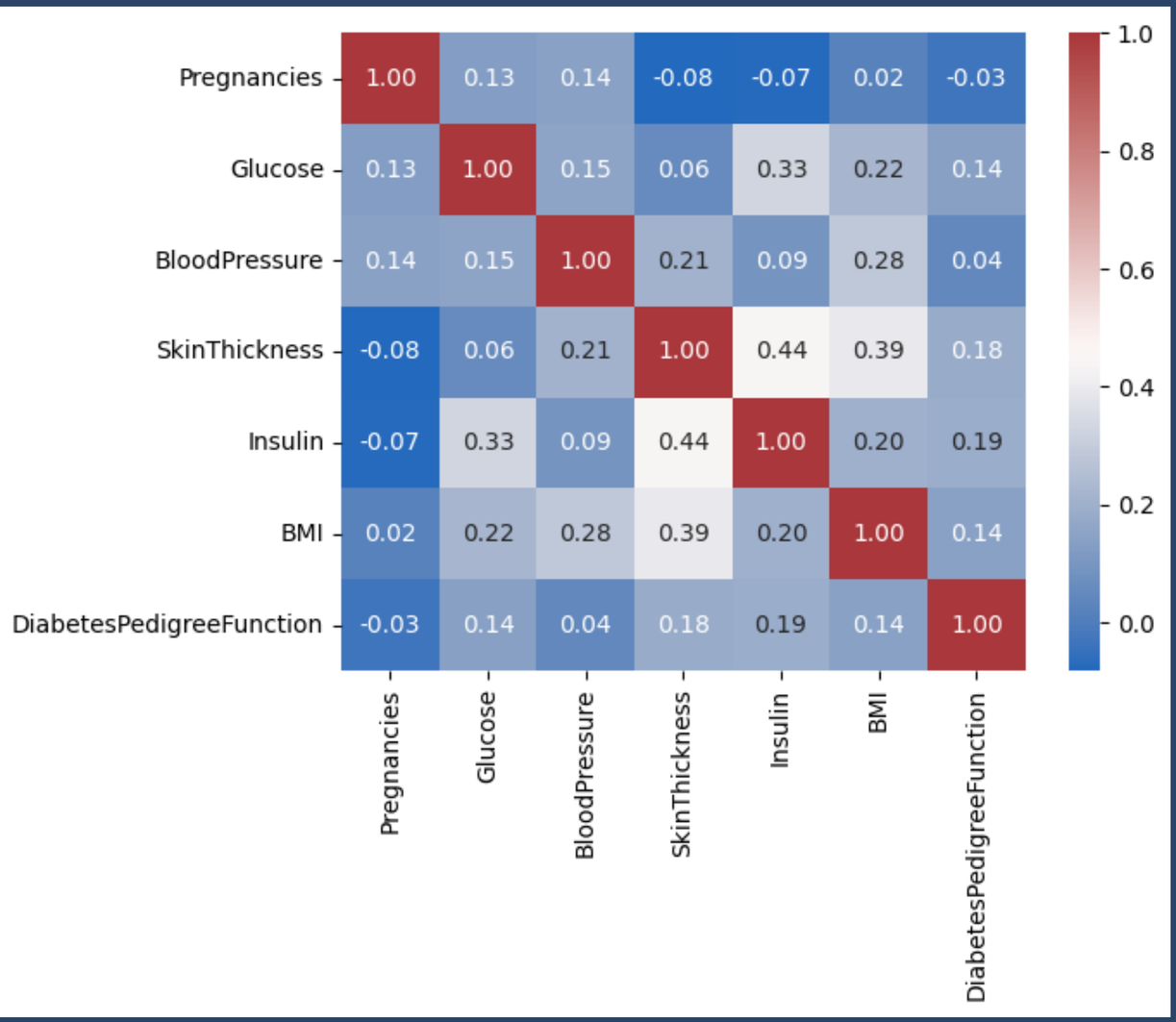
Coming up with conclusions using EDA, and machine learning models

EDA

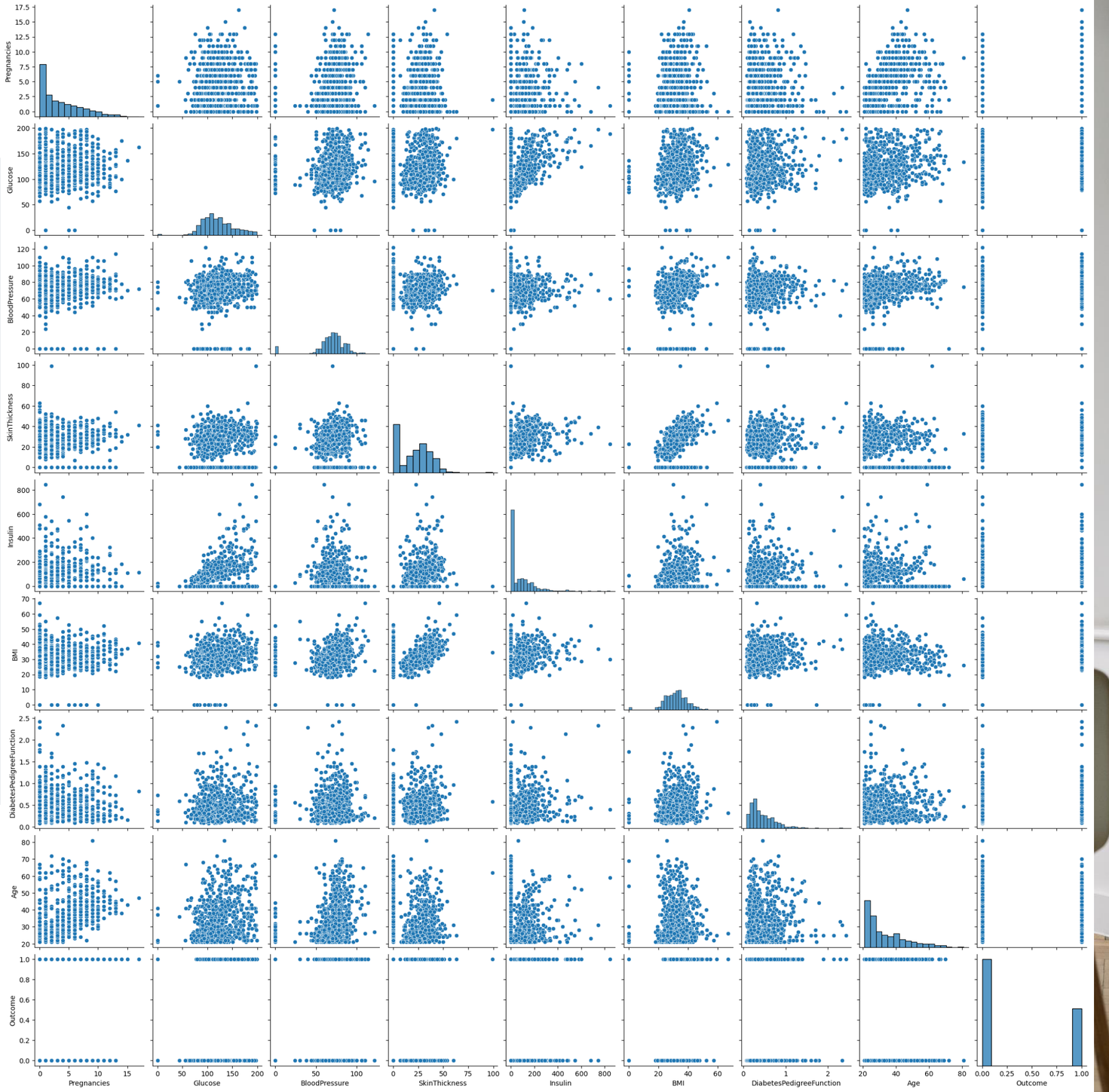
- Dataset does not contain any null values
- Dataset Description:

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

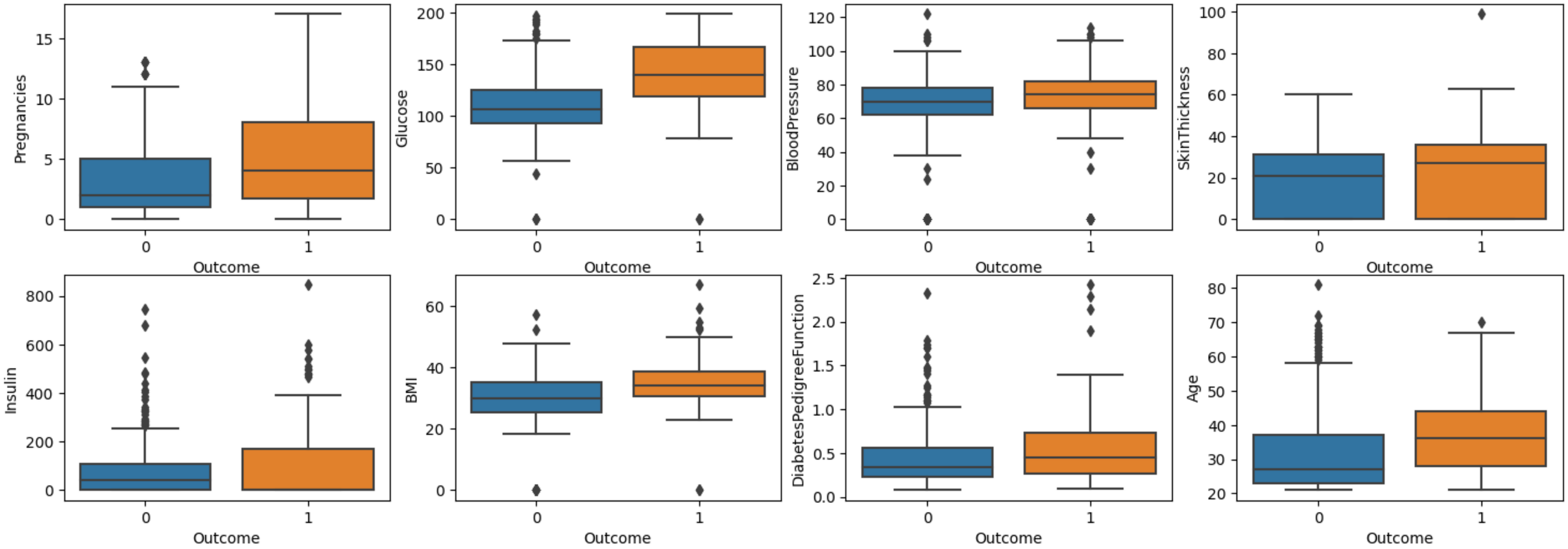
Heatmap of Predictors



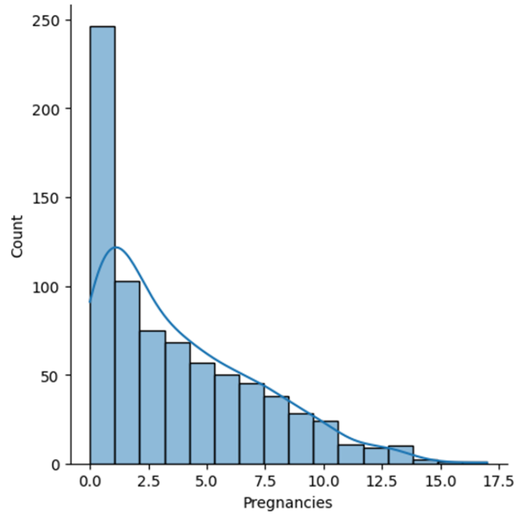
Pairplot for features

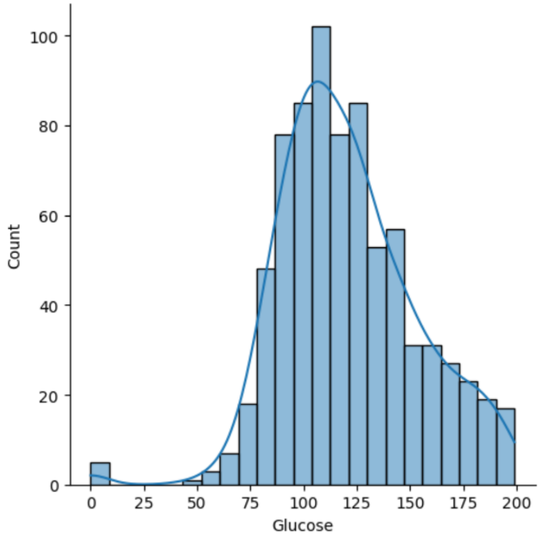


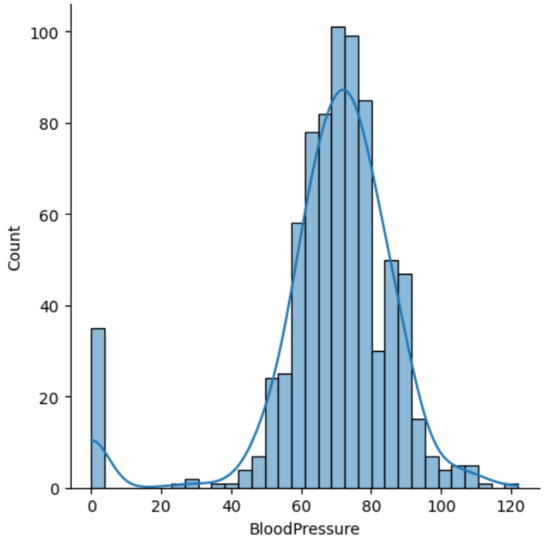
Boxplot for Outcome vs predictors

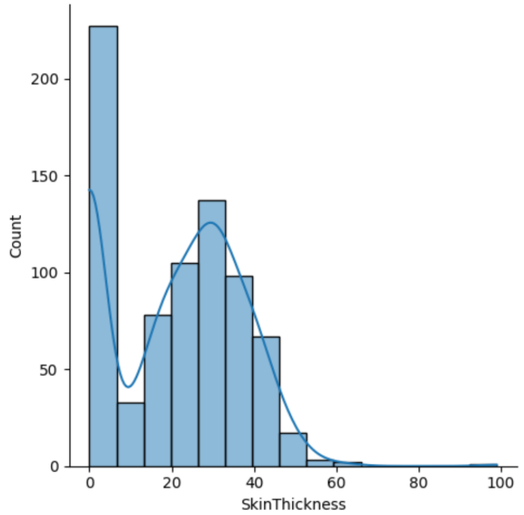


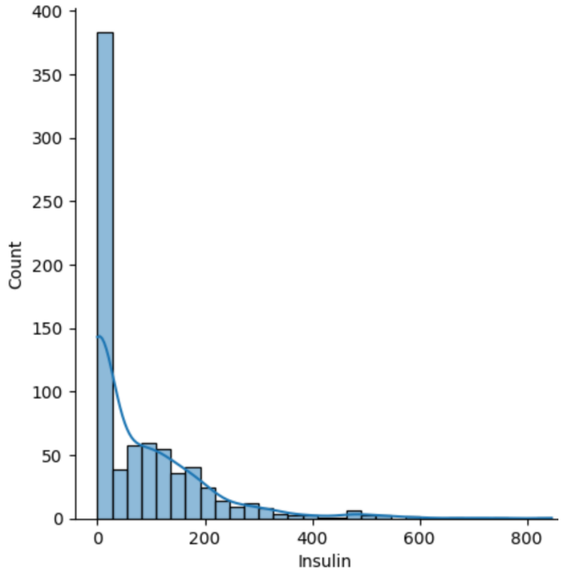
Distribution of Data

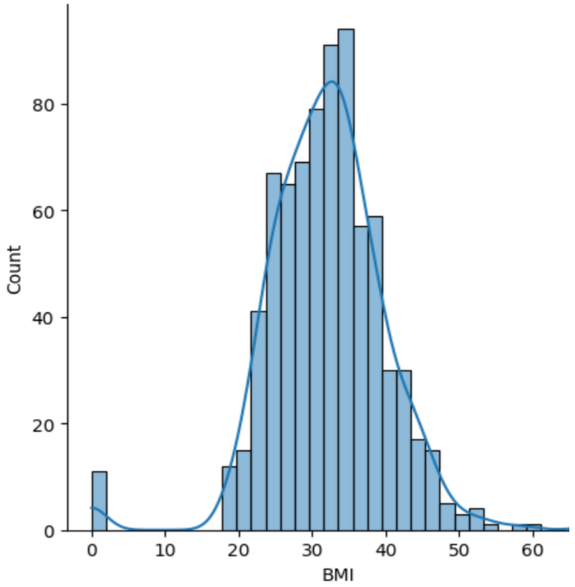


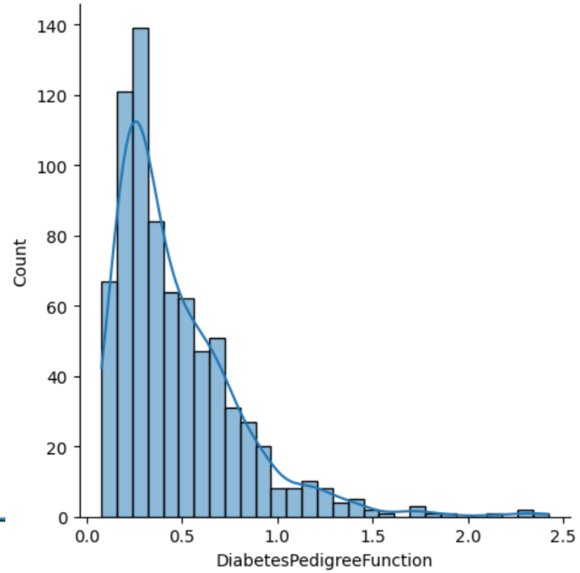


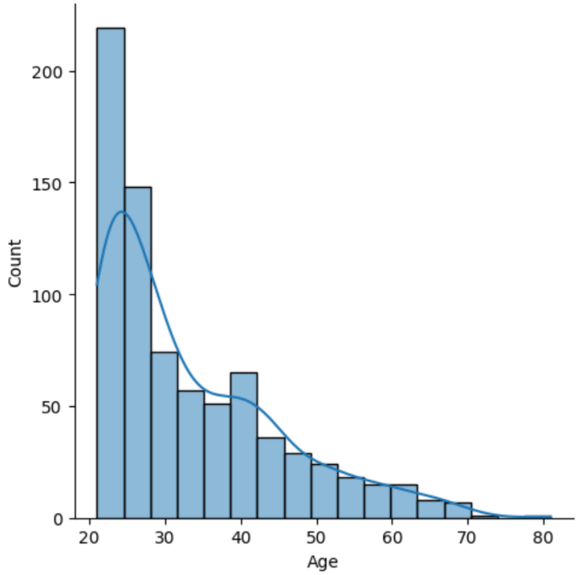




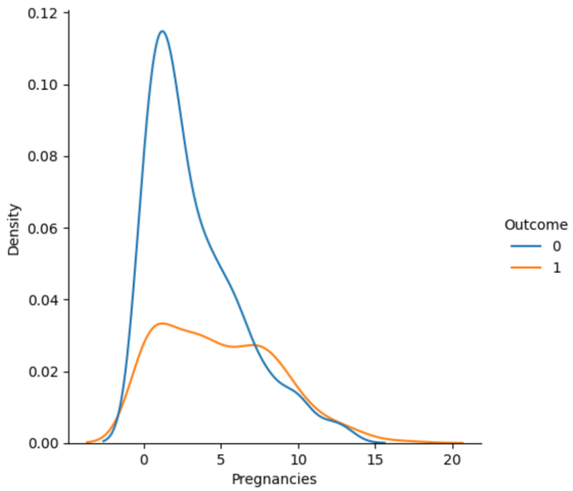


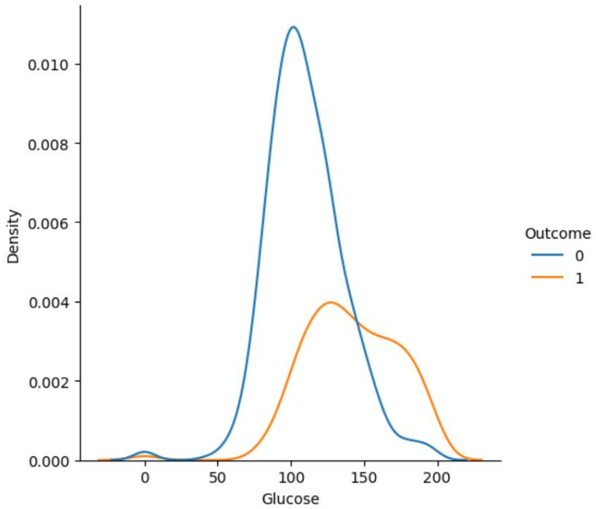


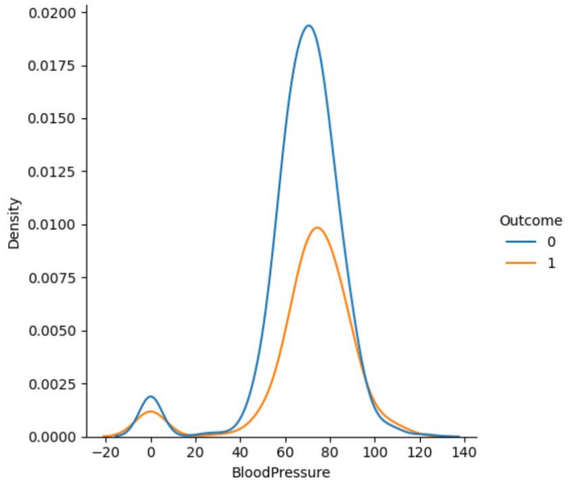


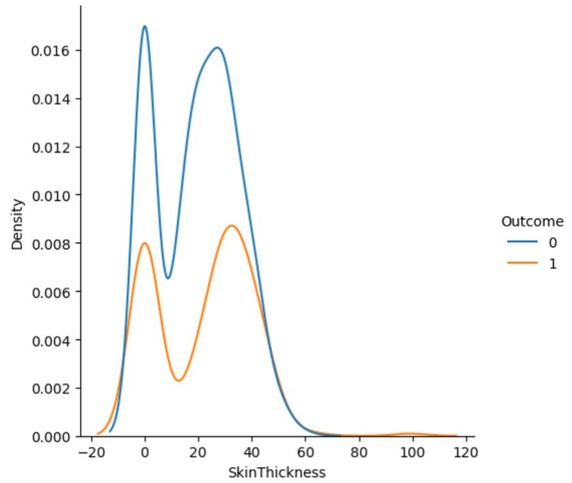


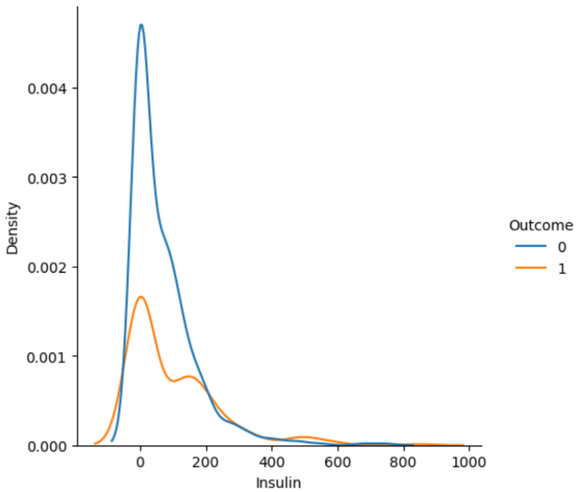
Distribution of Diabetic vs Non-diabetic

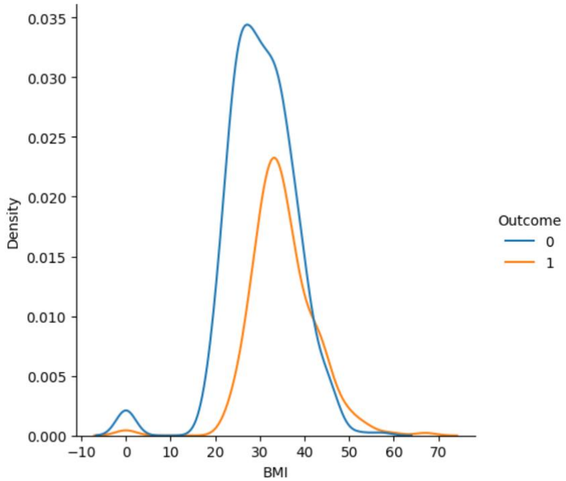


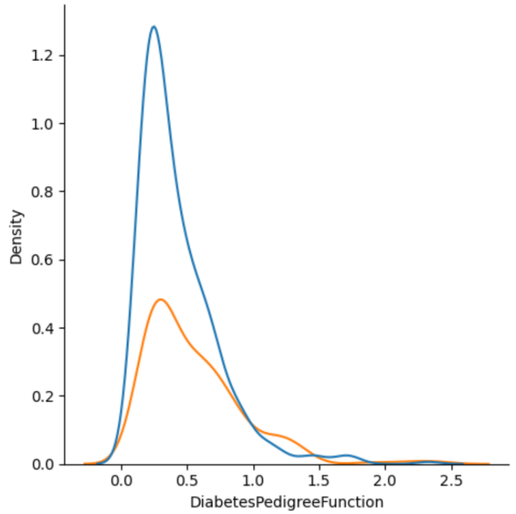


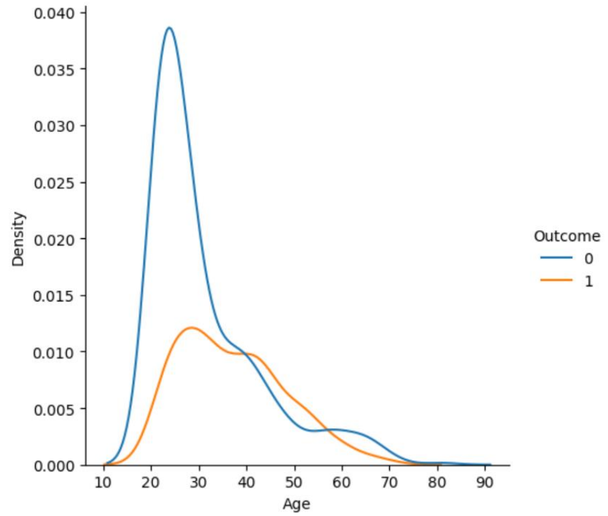






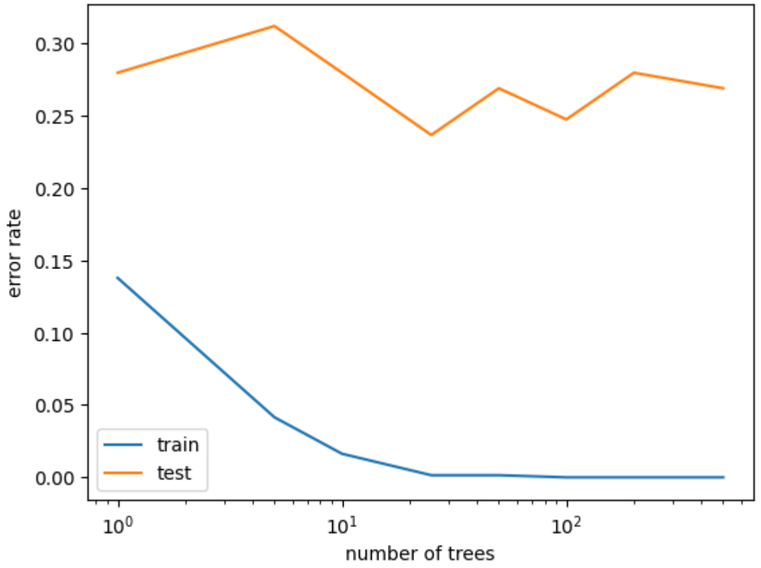






Results/ Conclusions

Random Forest Classifier Model

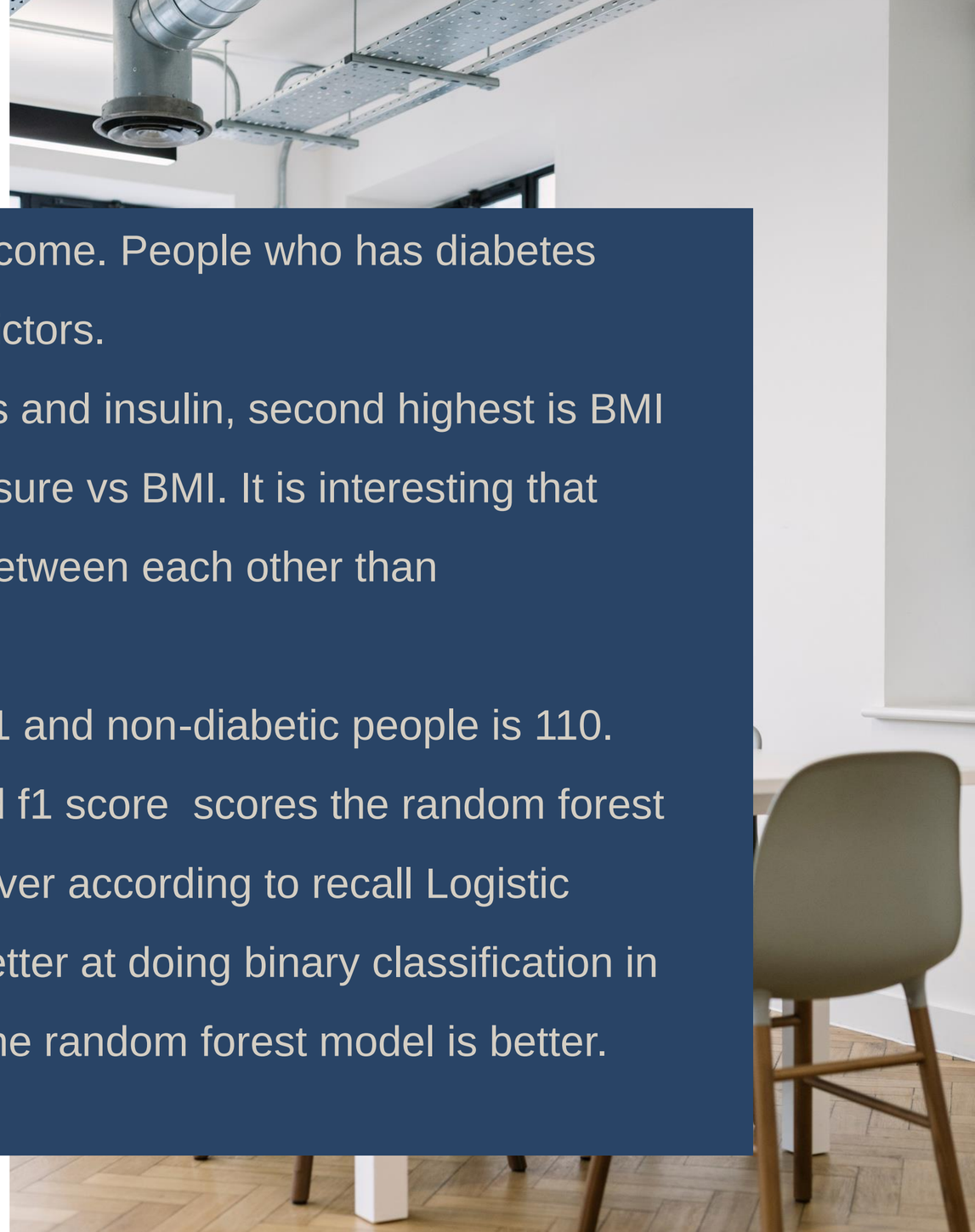



```
Accuracy: 0.752688
Precision: 0.628571
Recall: 0.687500
F1 score: 0.656716
ROC-AUC score: 0.737193
```

Logistic Regression Model

```
Accuracy: 0.698925
Precision: 0.547619
Recall: 0.718750
F1 score: 0.621622
ROC-AUC score: 0.703637
```

Conclusions



1. All the predictors have a relationship with the outcome. People who has diabetes seem to have a higher value for each of the predictors.
2. The highest correlation is between Skin thickness and insulin, second highest is BMI vs skin thickness, and third highest is Blood pressure vs BMI. It is interesting that most of the predictors have positive correlation between each other than pregnancies.
3. The mean glucose level for diabetic people is 141 and non-diabetic people is 110.
4. -According to accuracy, precision, ROC-AUC and f1 score scores the random forest classification model is better at classifying. However according to recall Logistic regression model is better. As roc-auc curve is better at doing binary classification in this case (TPR and FPR) we can conclude that the random forest model is better.

Challenges

- Limited time

Future Goals

- More EDA to learn more about the dataset
- More data cleaning

Thank you!