machine_learning_project-supervised-learning

- Machine Learning Project - Supervised learning

- By Ikenna Odinye

Project Outline

1

Project goal

2

Part I : EDA - Exploratory Data Analysis

3

Part II : Preprocessing & Feature Engineering

4

Part III : Training ML Model

5

Part IV : Conclusion

Project goal

- The goal is to create a Machine Learning Model that can diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset.

	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

```
data.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

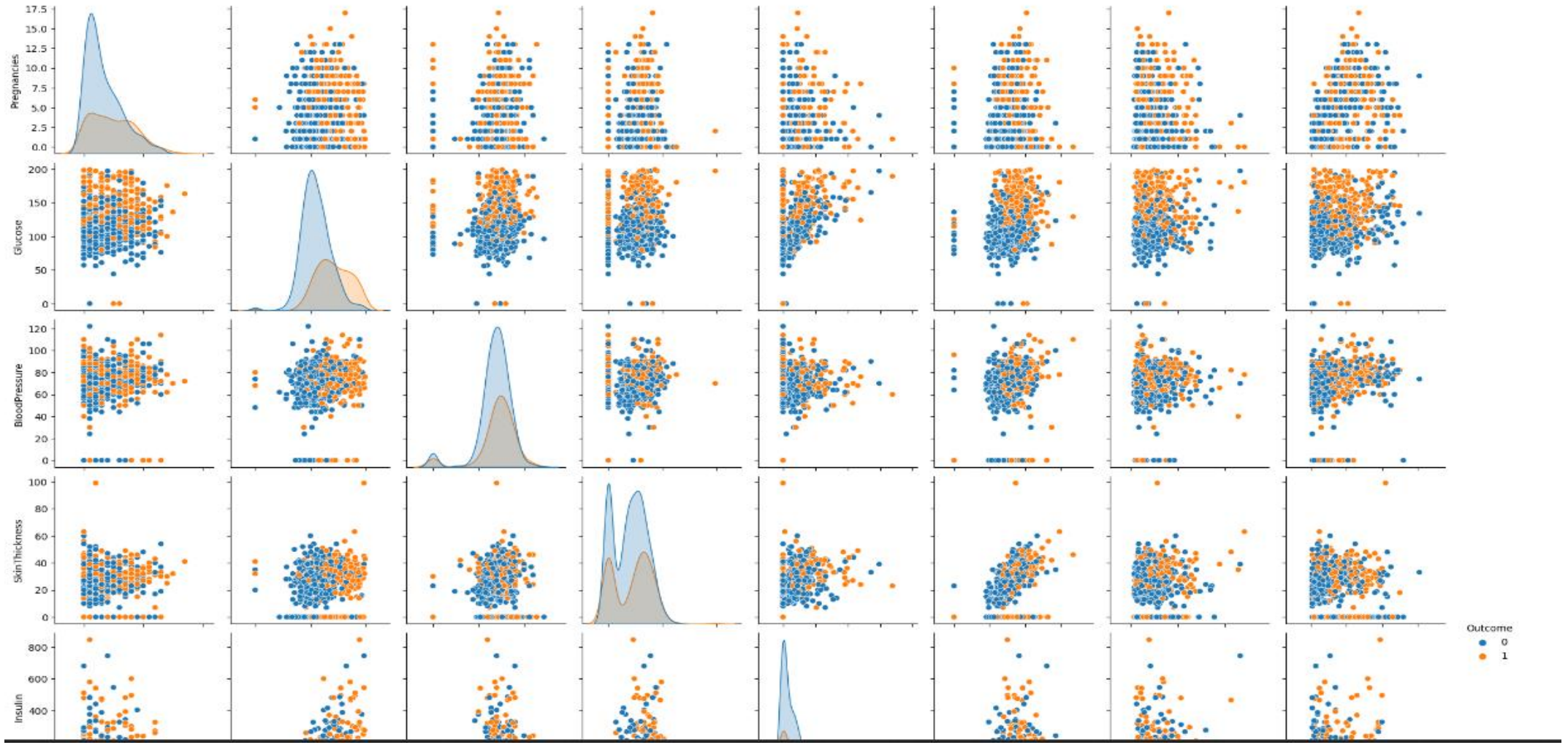
```
data.groupby('Outcome').count()
```

Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	500	500	500	500	500	500	500	500
1	268	268	268	268	268	268	268	268

Part I : EDA - Exploratory Data Analysis

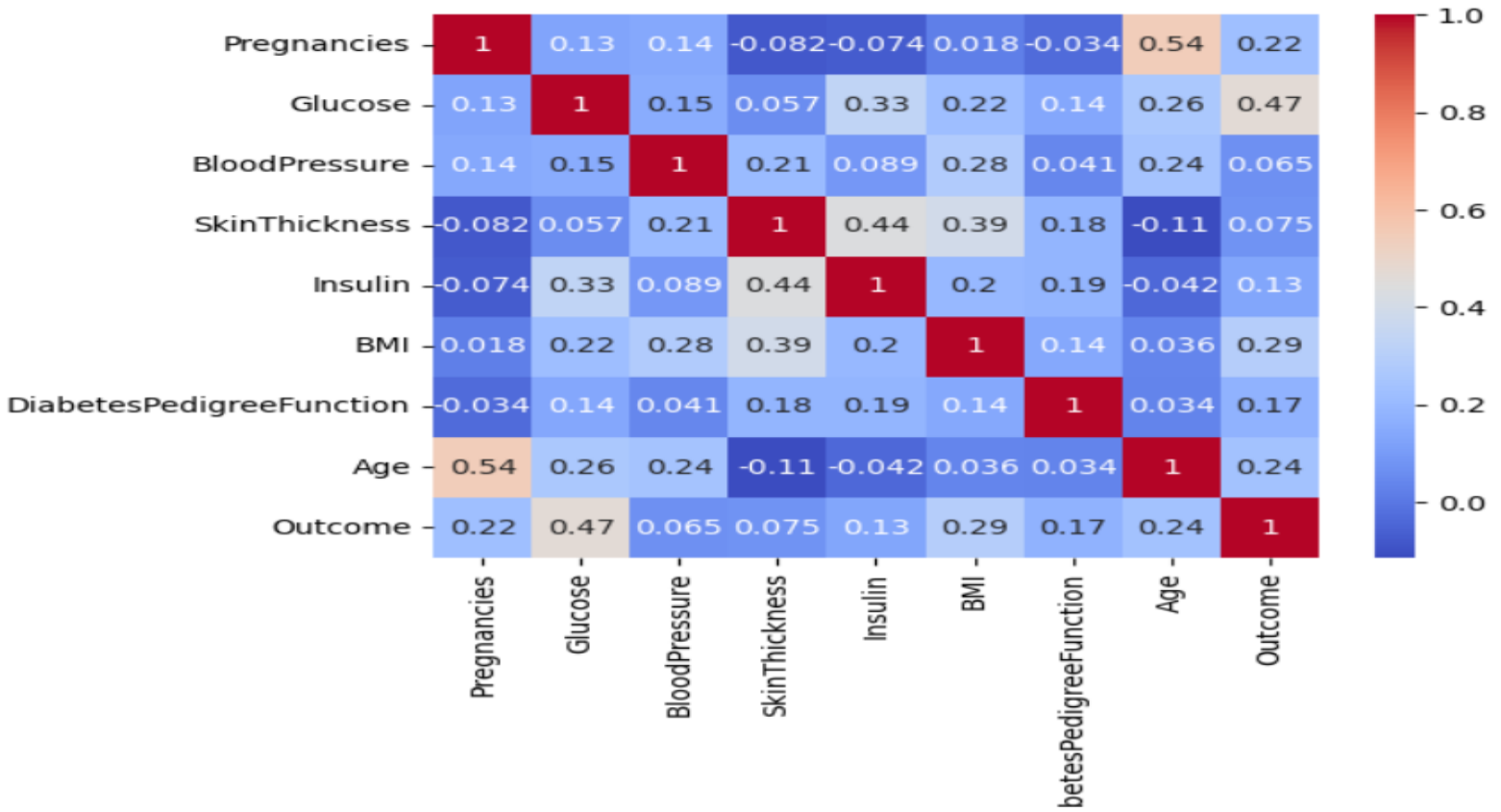
Data Visualization:Finding relationship between predictor variables and outcome variable

```
sns.pairplot(data, hue="Outcome", diag_kind="kde")
plt.show()
```



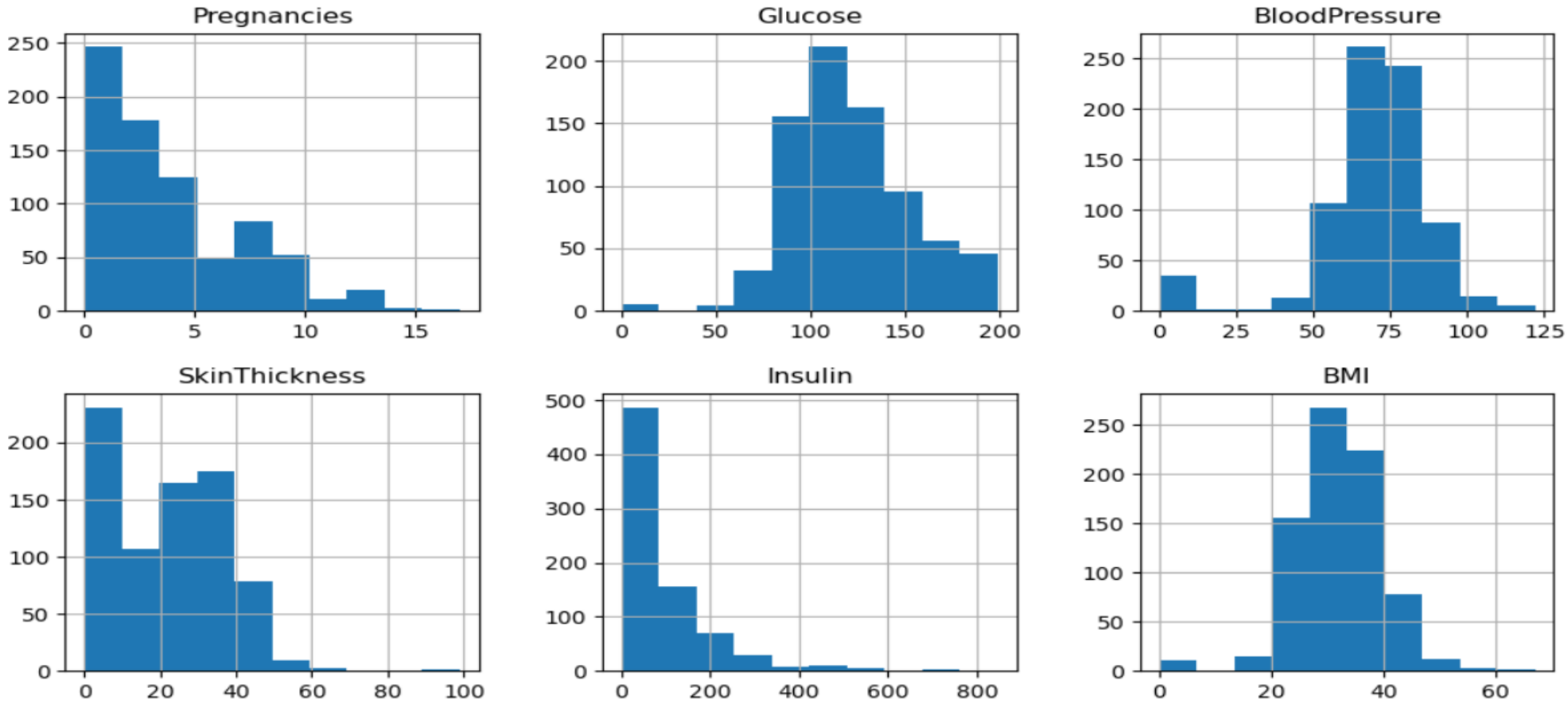
Correlation Matrix: Finding the correlation between the predictor variables

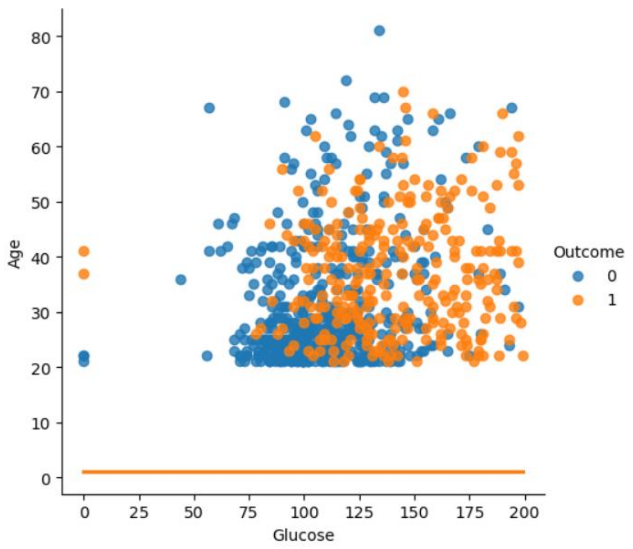
```
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm")
plt.show()
```

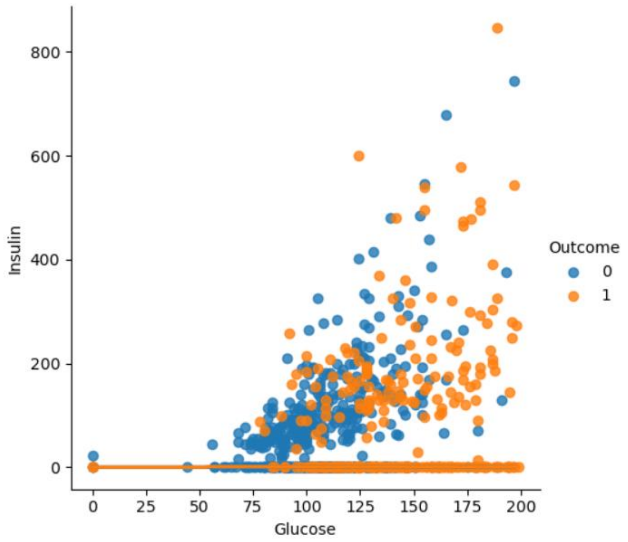


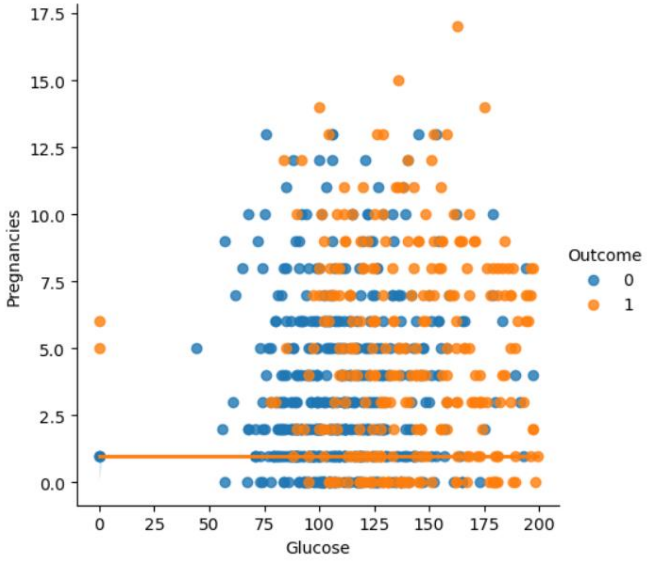
Distribution of each predictor variable

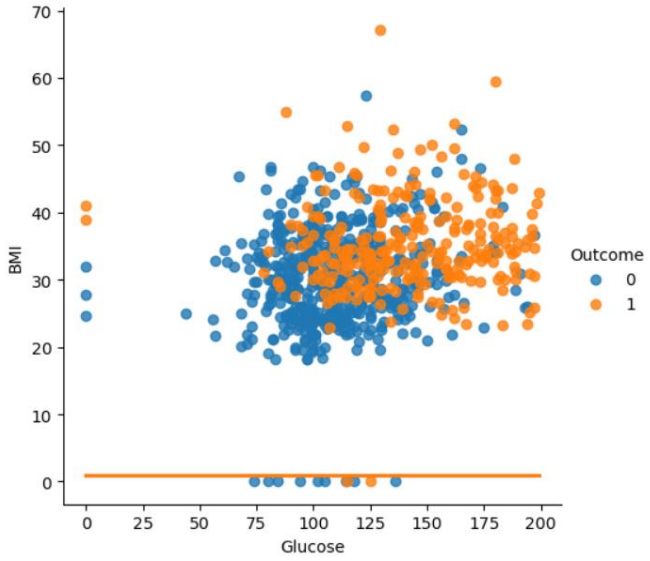
```
data.hist(figsize=(12, 10))
plt.show()
```





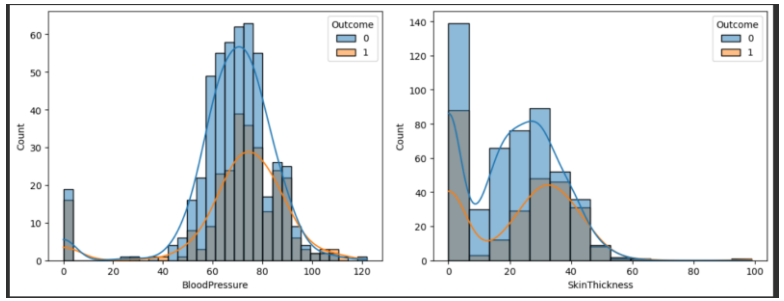


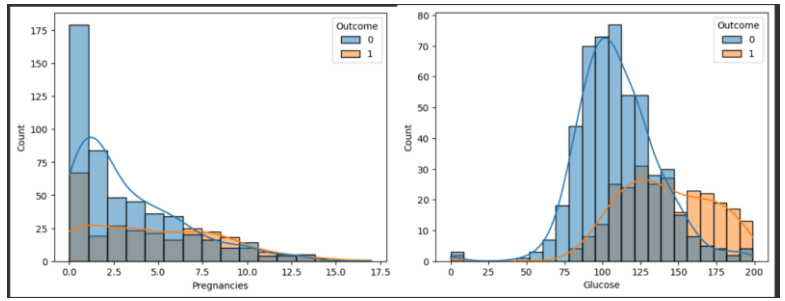


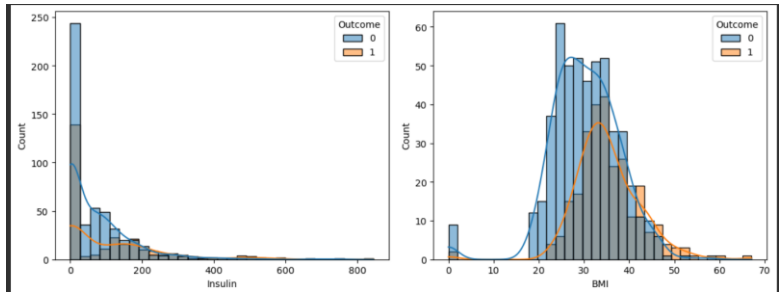


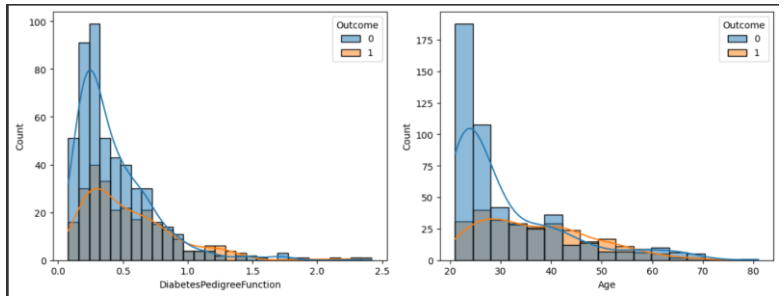
Interaction effect between the predictor variables

How the distribution of the predictor variables differ for individuals with diabetes and without diabetes

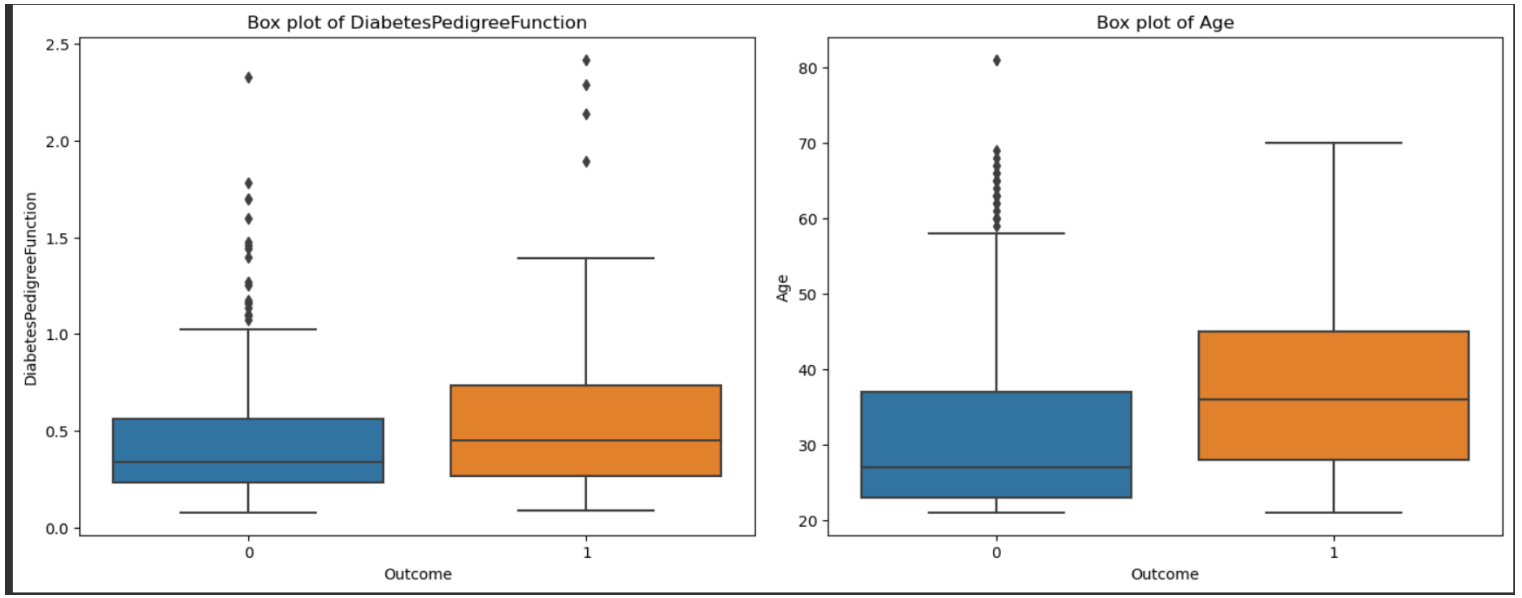


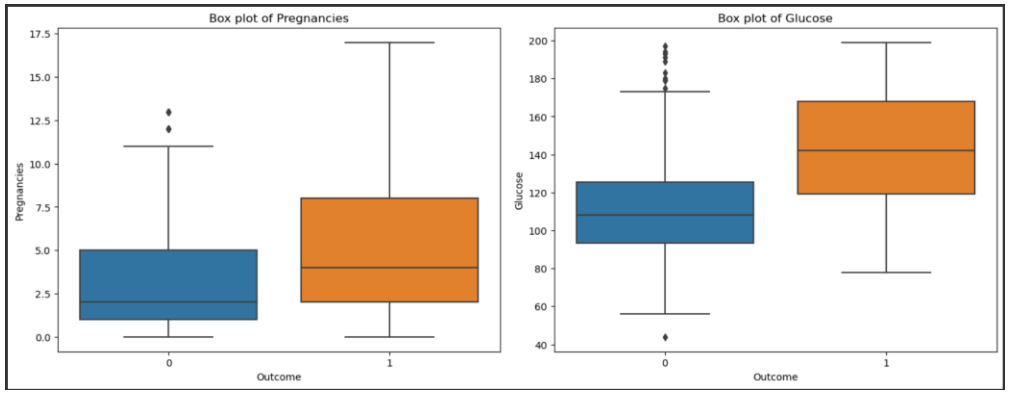


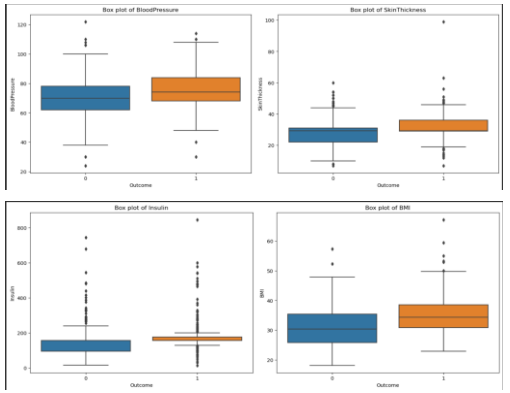




Using boxplot to check outlier







```
Num zeros in column Pregnancies is: 111
Num zeros in column Glucose is: 5
Num zeros in column BloodPressure is: 35
Num zeros in column SkinThickness is: 227
Num zeros in column Insulin is: 374
Num zeros in column BMI is: 11
Num zeros in column DiabetesPedigreeFunction is: 0
Num zeros in column Age is: 0
Num zeros in column Outcome is: 500
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	724.000000	724.000000	724.000000	724.000000	724.000000	724.000000	724.000000	724.000000	724.000000
mean	3.866022	121.882597	72.400552	29.182331	156.056122	32.467127	0.474765	33.350829	0.343923
std	3.362803	30.750030	12.379870	9.018907	87.395294	6.888941	0.332315	11.765393	0.475344
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	25.000000	118.250000	27.500000	0.245000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	29.182331	156.056122	32.400000	0.379000	29.000000	0.000000
75%	6.000000	142.000000	80.000000	33.000000	156.056122	36.600000	0.627500	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

```
Shape before dropping NAs (768, 9)
Shape after dropping NAs for Glucose, BMI, and BloodPressure columns (724, 9)
```

PART II: Processing and Feature Engineering – Handling Null values (average imputation) which will address outliers

Handling imbalanced data: As the number of samples for Class 0 is significantly higher than Class 1, the dataset is imbalanced

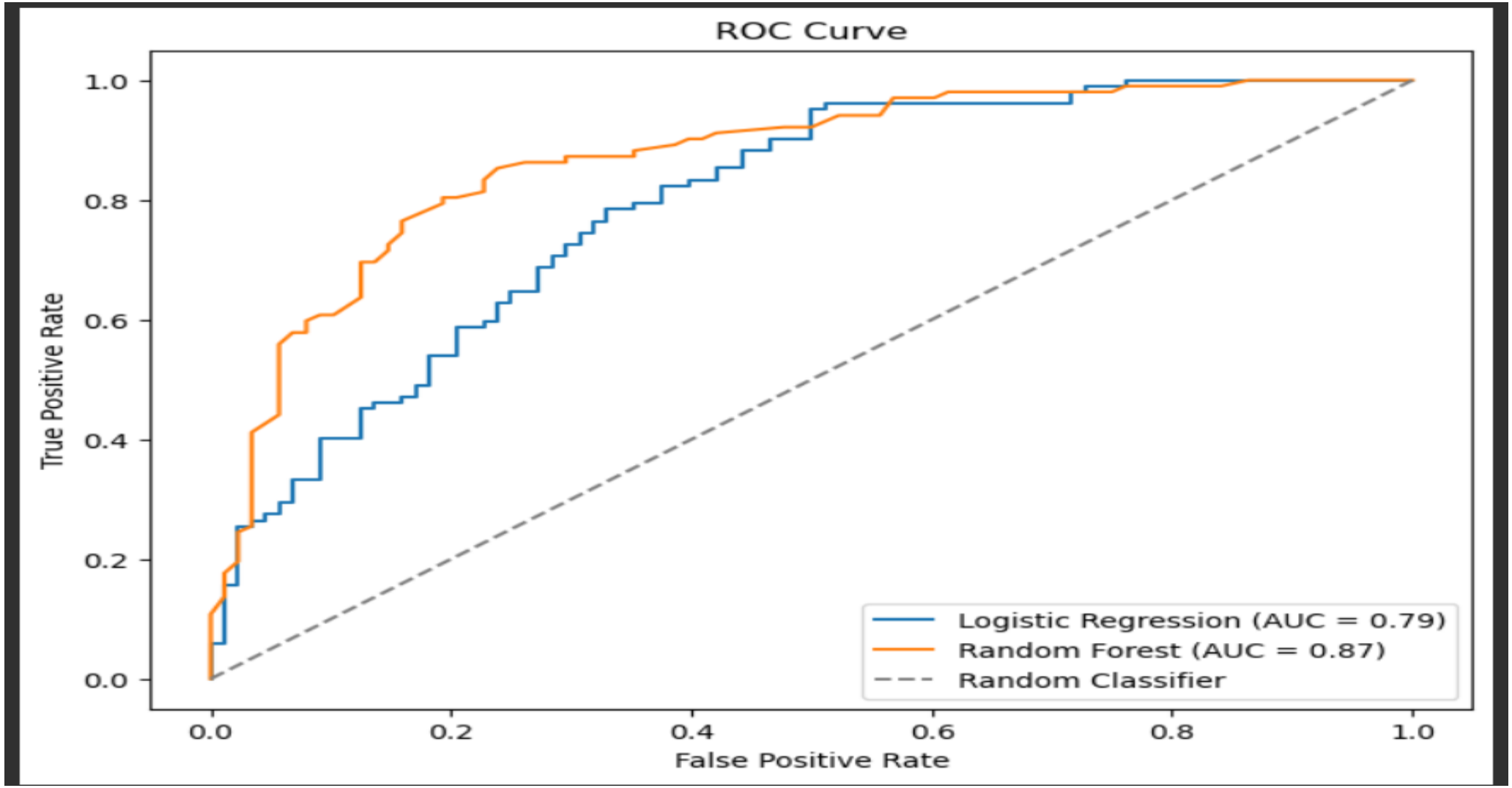
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Outcome								
0	475	475	475	475	475	475	475	475
1	249	249	249	249	249	249	249	249

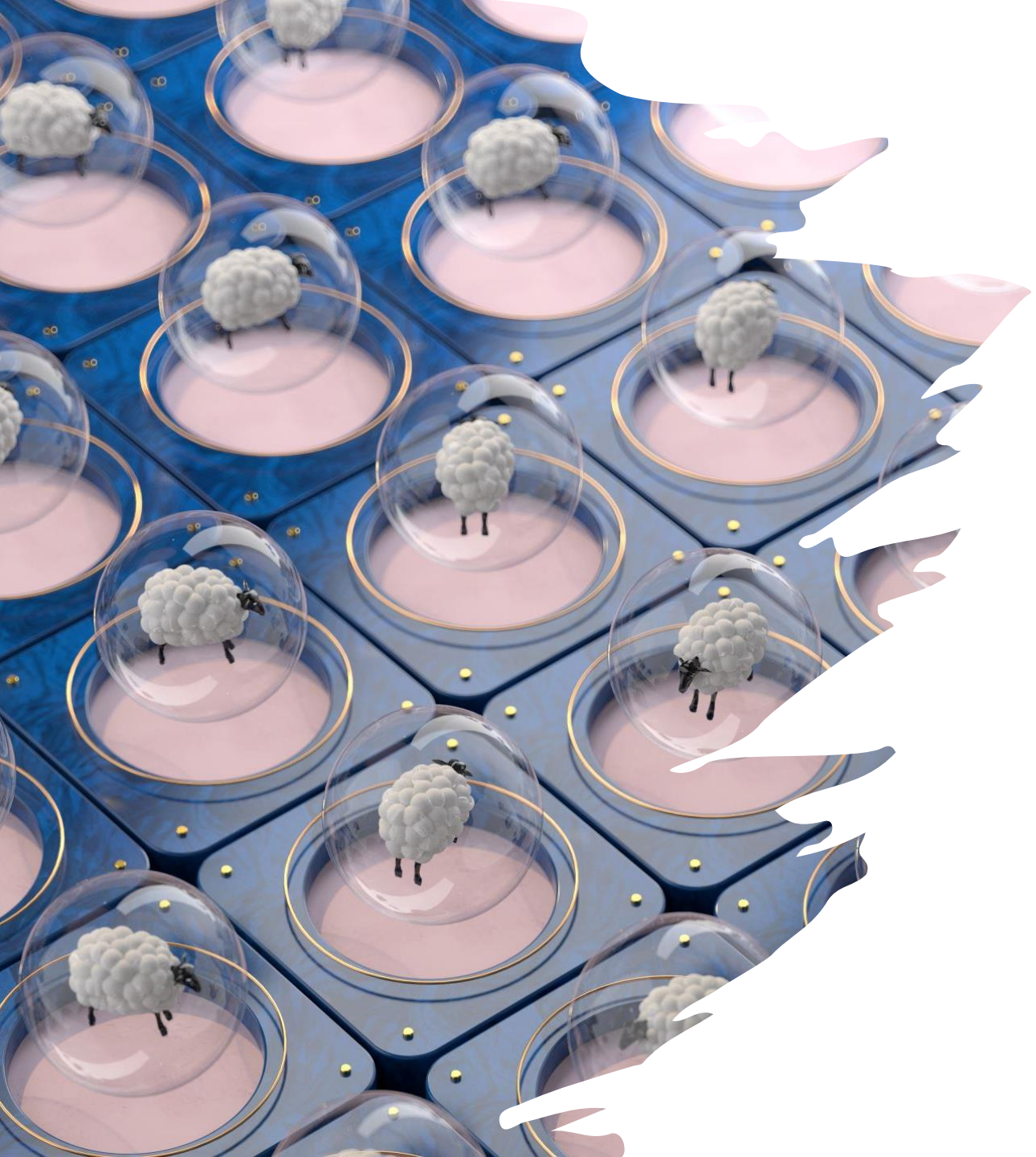
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Outcome								
0	475	475	475	475	475	475	475	475
1	475	475	475	475	475	475	475	475

	Logistic Regression	Random Forest
Training Accuracy	0.7723684210526316	1.0
Test Accuracy	0.7052631578947368	0.8052631578947368

PART III: Training ML models (Logistics Regression & Random Forest)

Plot ROC curve





Part IV : Conclusion

- Overfitting:

- There were no significant signs of overfitting observed in either model. The testing accuracy of the Random Forest model was close to its training accuracy, indicating that the model generalized well to unseen data. And Logistic Regression model also showed reasonable consistency between training and testing accuracy, suggesting that it avoided overfitting.

- Feature Importance:

- Features such as "Glucose," "BMI," and "Age" emerged as significant contributors to predicting diabetes. And these features can provide valuable insights for medical practitioners to focus on critical indicators during diagnosis.

Part IV : Conclusion

- Performance score:

- The Random Forest model outperformed the Logistic Regression model in all key evaluation metrics, including accuracy, precision, recall, F1 score, and ROC AUC score.

- The higher accuracy and other metrics of the Random Forest model indicate its superior ability to predict the presence or absence of diabetes compared to the Logistic Regression model.

- Room for Improvement:

- While the models achieved reasonable accuracy, there is still room for further improvement.

- Additionally, obtaining more diverse and larger datasets may enhance the model's predictive capabilities. For example, diabetes mellitus has two types – Type 1 and Type 2 and they have different ethological causes such as, Hereditary and diet, which should be captured in the data. Note that Type 1, Insulin-dependent diabetes, generally develops in childhood or adolescence..