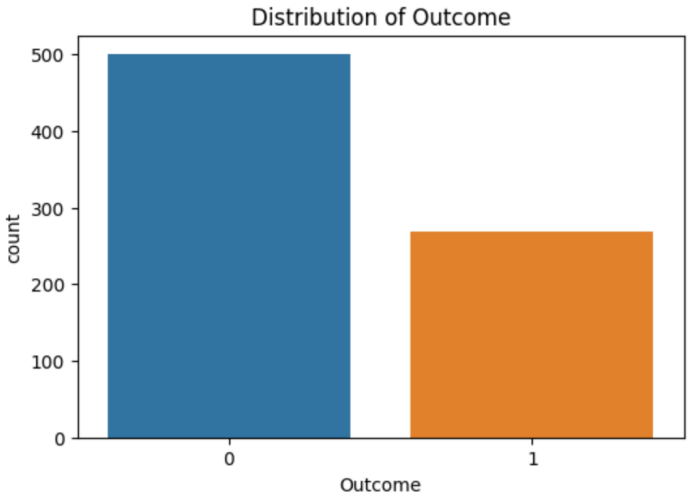
Supervised Learning Project
— based on the “Diabetes” dataset

Part1: Exploratory Data Analysis

1. Get some basic information about the dataset, including shape, columns, data types, null values.
2. Visulaize the relationships between the different variables.

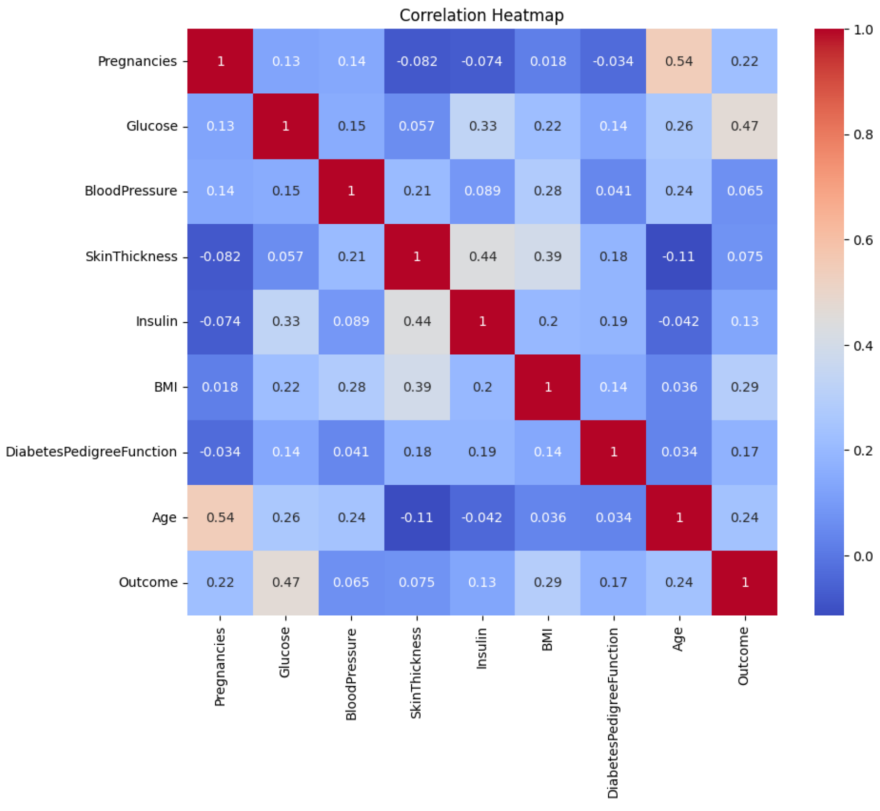
Visualization

2.1 Outcome Distribution



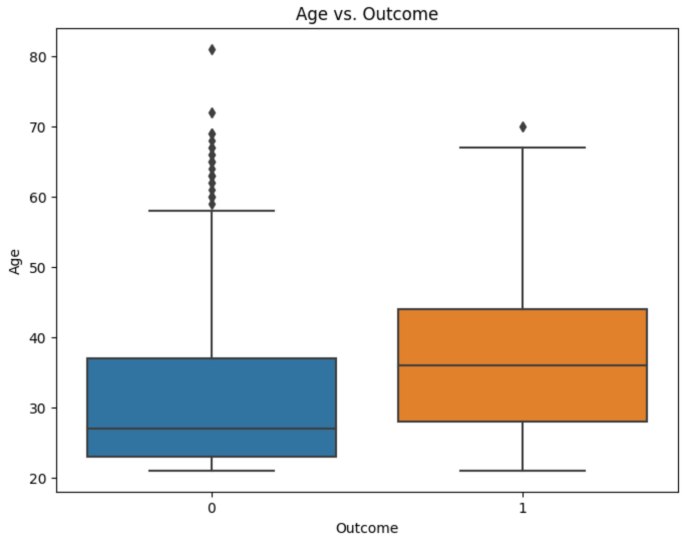
Attention: imbalanced data

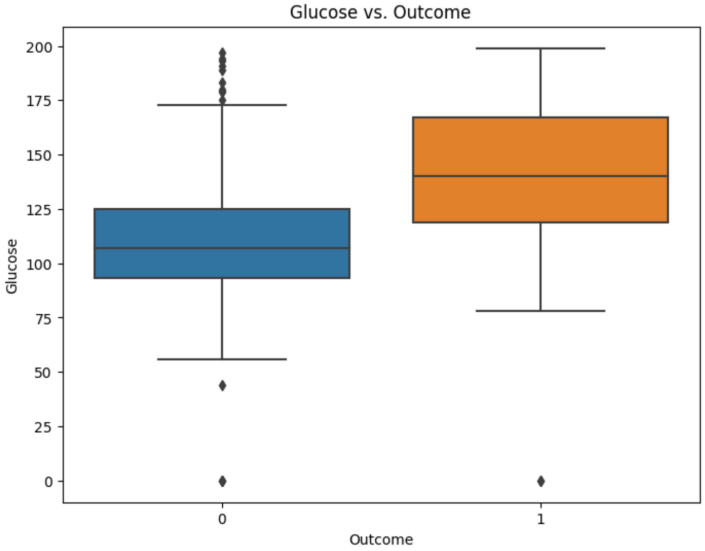
2.2 Correlation Heatmap

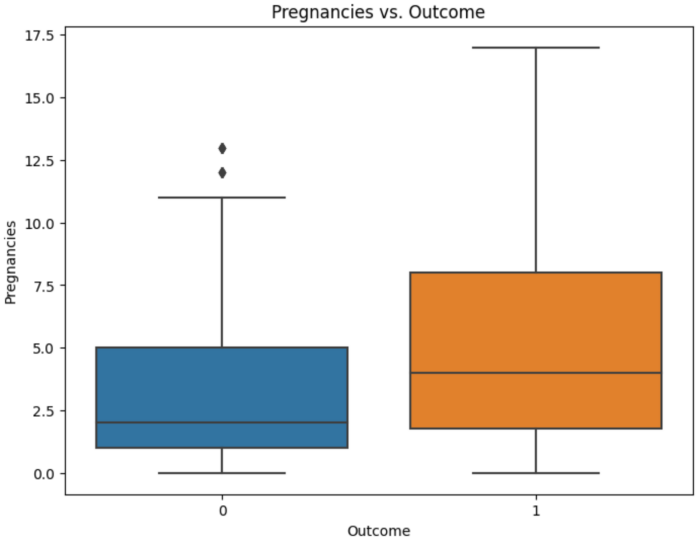


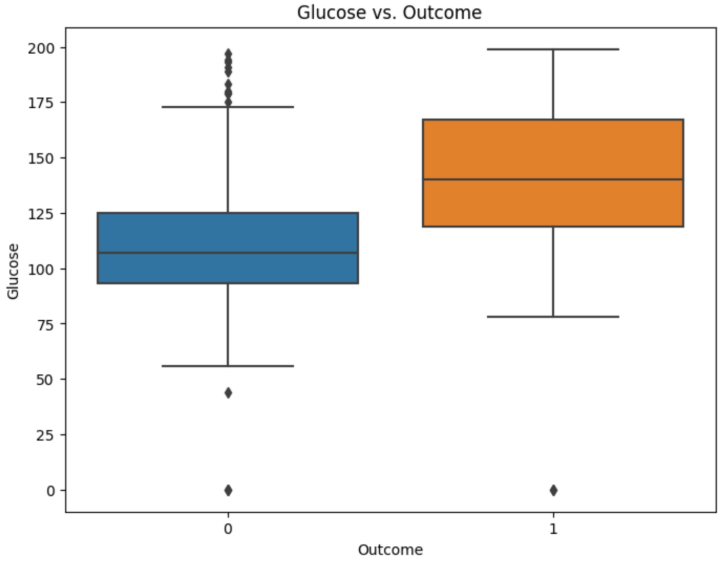
Glucose - Outcome: 0.47
BMI - Outcome: 0.29
Pregancies - Outcome: 0.22
Age - Outcome: 0.24

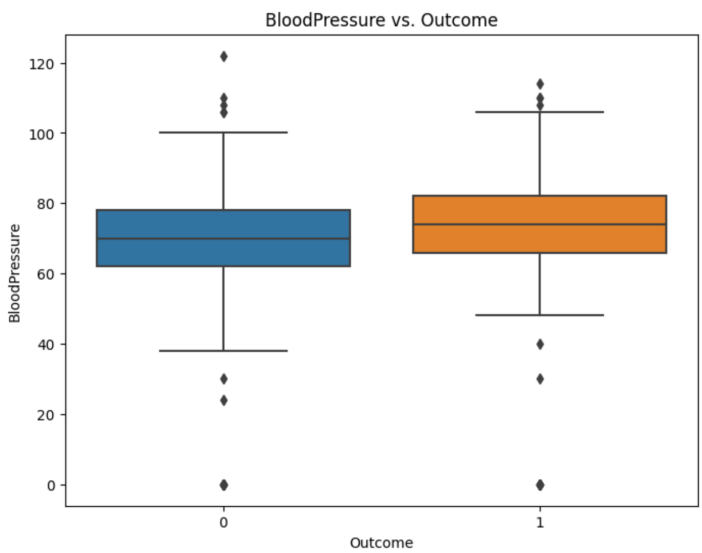
2.3 Boxplot for outliers

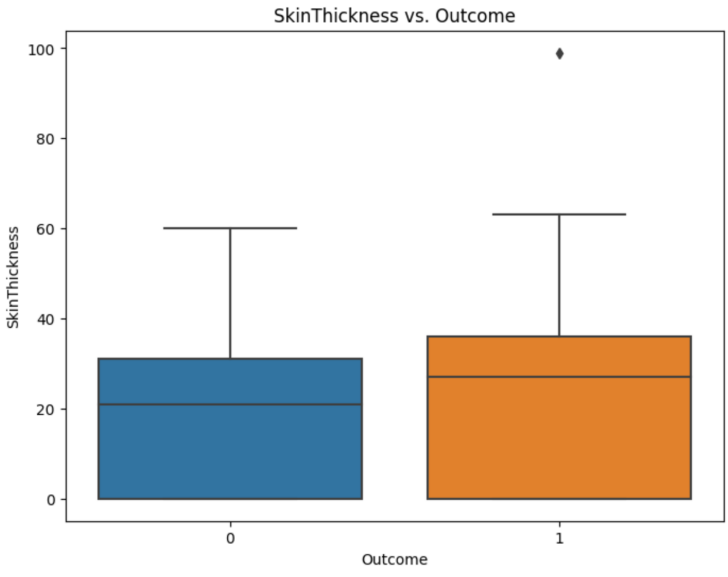


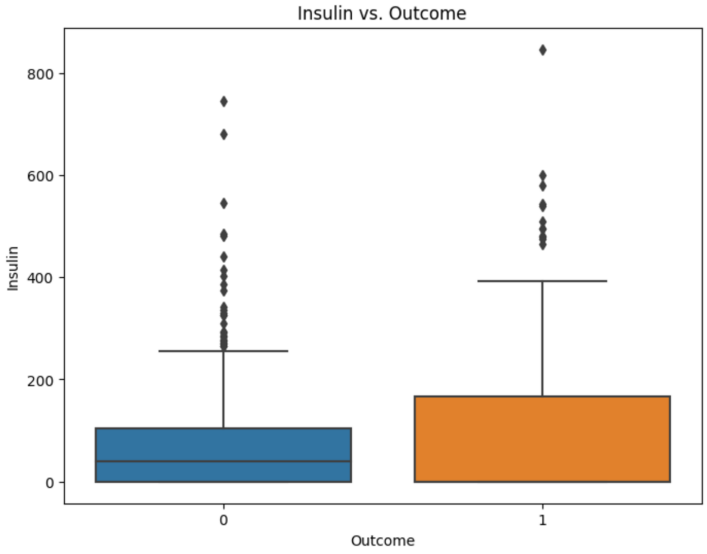


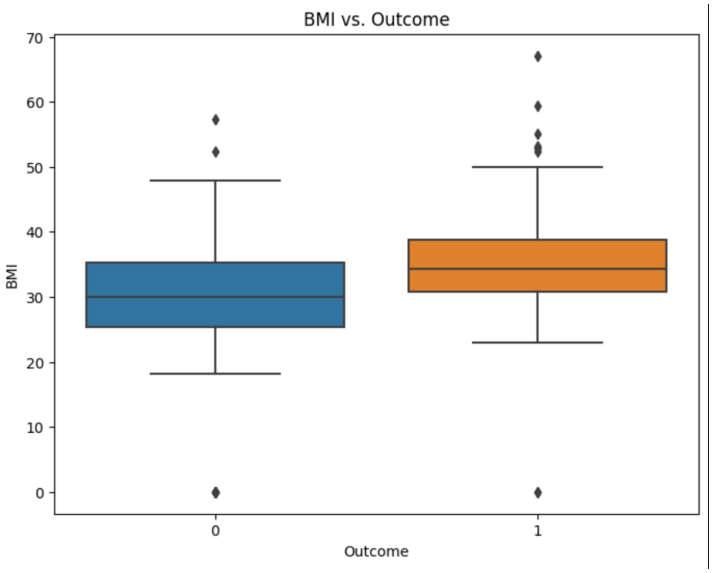


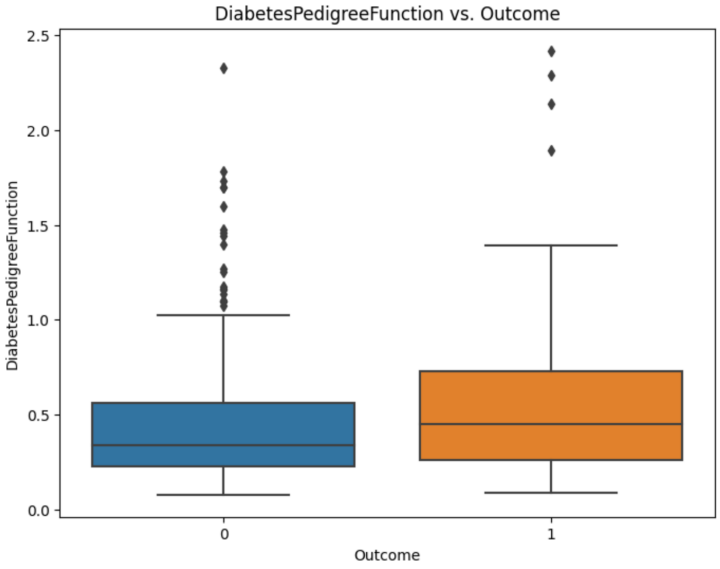












Part2: Preprocessing & Feature Engineering

1. Handling missing values
   No missing values in this dataset.
2. Handling outliers
   Using IQR-based approach.
3. Feature engineering
   Create a new feature named ‘BMI_Category’ with labels: Underweight, Normal, Overweight, Obese and applied label encoding.
4. Handling imbalanced data
   Apply random undersampling to balance the classes.

Part3: Train ML Model

Select **Logistic Regression** and **Random Forest** models.

3.1 Train and test model with all features

	Logistic Regression	Random Forest
accuracy	0.6962025316455697	0.7341772151898734
precision	0.5476190476190477	0.6060606060606061
recall	0.8214285714285714	0.7142857142857143
F-1 score	0.6571428571428571	0.6557377049180327
roc-auc	0.7244397759103641	0.7296918767507004

Random Forest WINs!!!

3.2 Remove BMI_Category feature

	Logistic Regression	Random Forest
accuracy	0.7341772151898734	0.7468354430379747
precision	0.5945945945945946	0.6111111111111112
recall	0.7857142857142857	0.7857142857142857
F-1 score	0.676923076923077	0.6875000000000001
roc-auc	0.745798319327731	0.7556022408963585

Random Forest WINs!!!

3.3 KFold cross-validation

	Logistic Regression	Random Forest
Fold 1	0.7722	0.7468
Fold 2	0.6835	0.7468
Fold 3	0.7722	0.7468
Fold 4	0.7468	0.7468
Fold 5	0.7564	0.7564
Mean Accuracy	0.7462	0.7488

Random Forest WINs!!!

Part4: Conclusion

From the machine learning models developed and the exploratory data analysis (EDA) conducted, there are my findings:

1. Based on the correlation heatmap, Glucose is the most siginificant predictor of disbetes outcome. Also, age, BMI and pregnancy play important roles.
2. Proper preprocessing steps, including feature scaling, one-hot-encoding significantly improved the model's performance in this case.
3. The accuracy incresed for both models after removing the new generated feature - BMI_Category_Encoded.
4. Logistic Regression and Random Forest were developed as predictive models for diabetes outcome. Random Forest has a batter performance on the test dataset.
5. The dataset shows an imbalanced distribution, with a higher number of non-diabetic cases compared to diabetic cases.