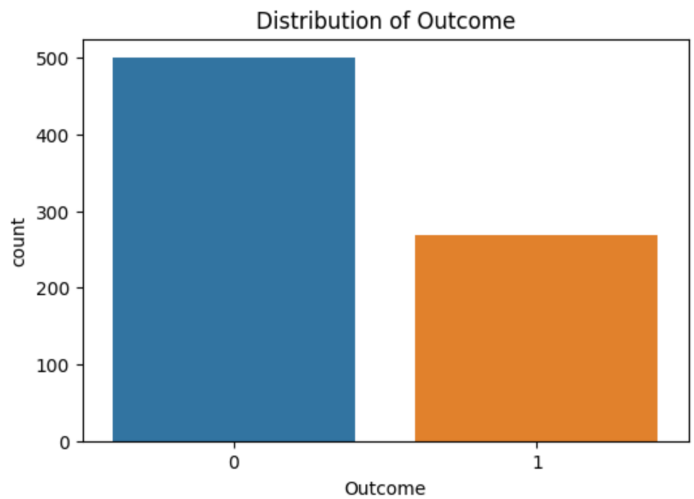
Supervised Learning Project
— based on the "Diabetes" dataset

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.

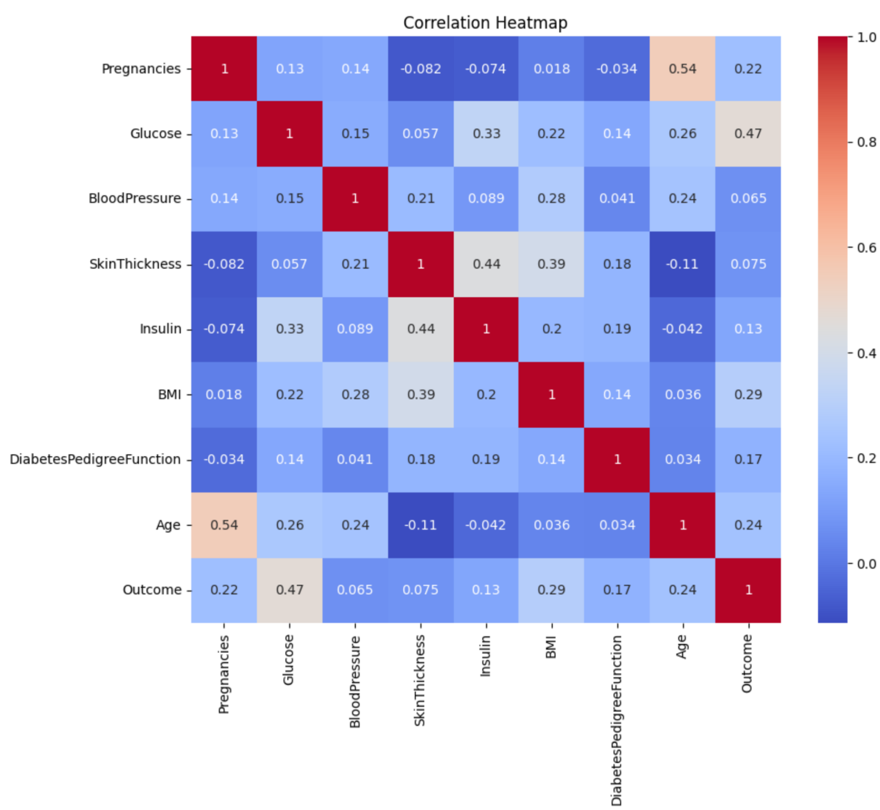
Visualization

2.1 Outcome Distribution



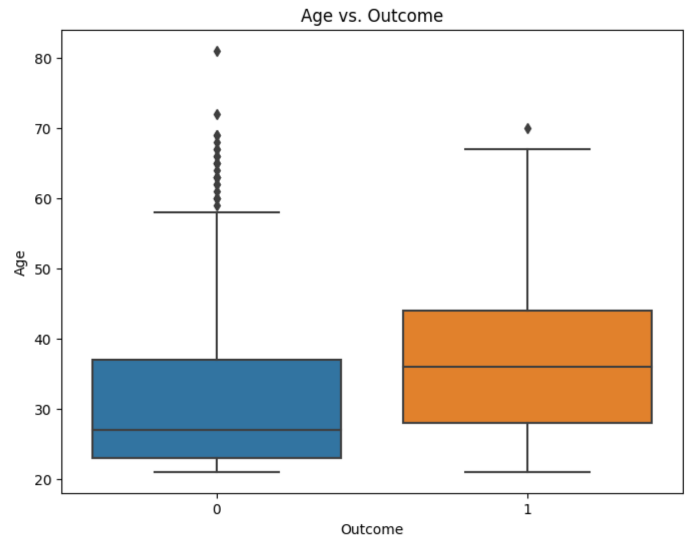
Attention: imbalanced data

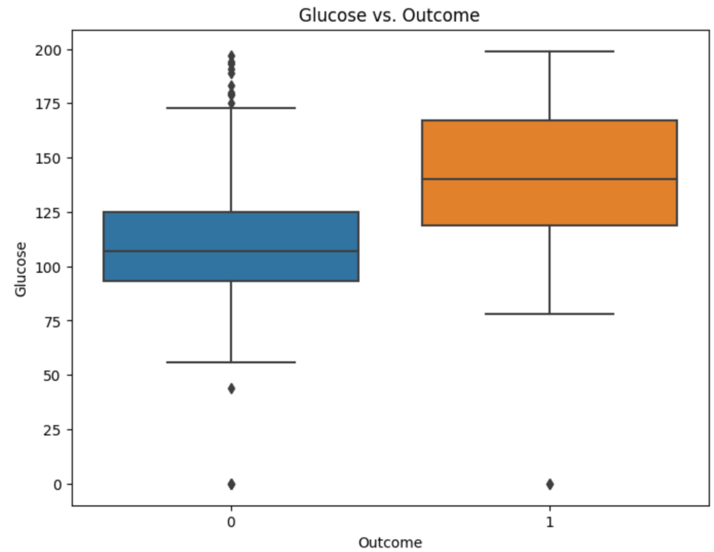
2.2 Correlation Heatmap

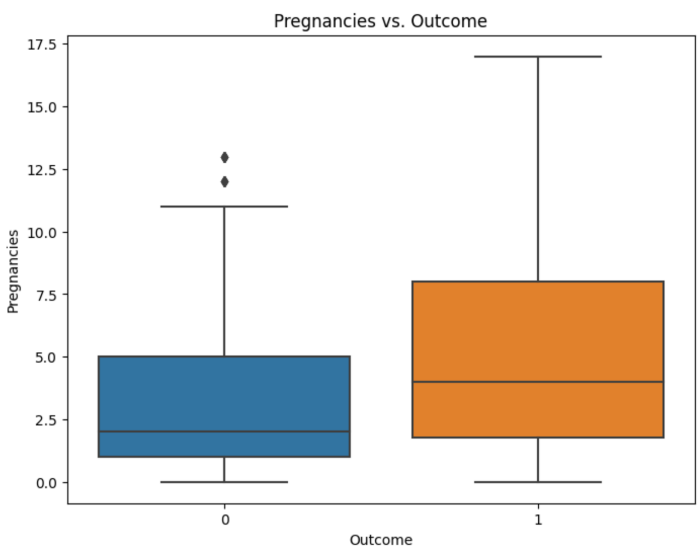


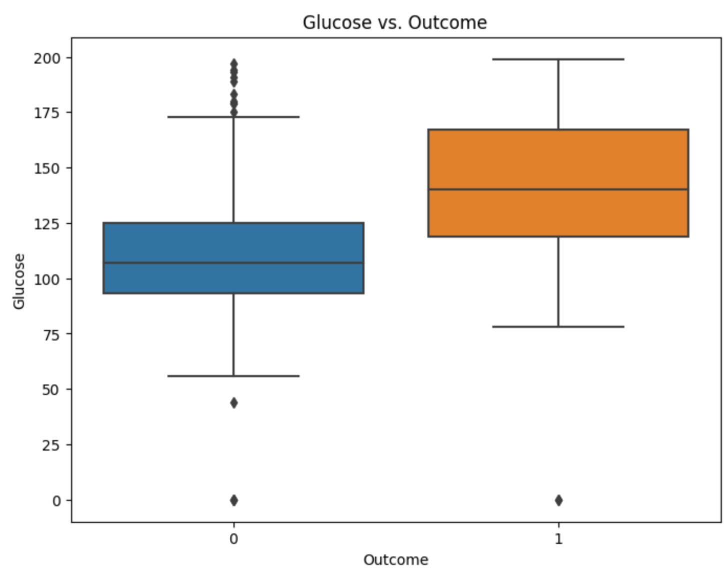
Glucose - Outcome: 0.47
BMI - Outcome: 0.29
Pregancies - Outcome: 0.22
Age - Outcome: 0.24

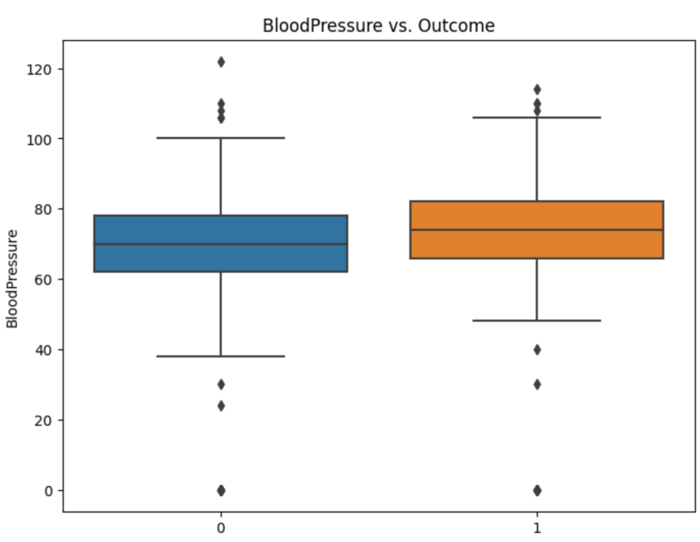
2.3 Boxplot for outliers

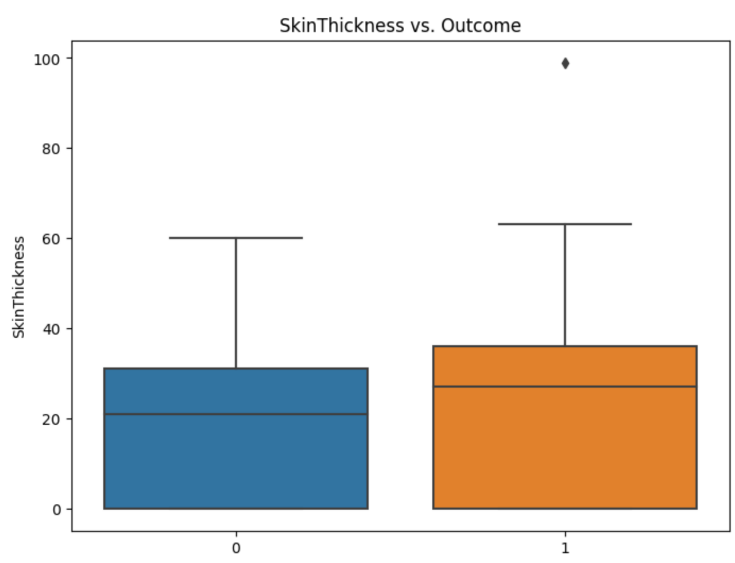


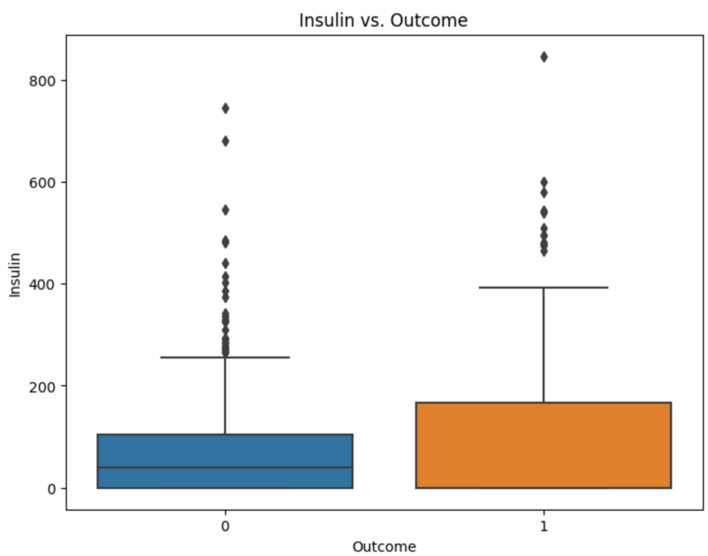


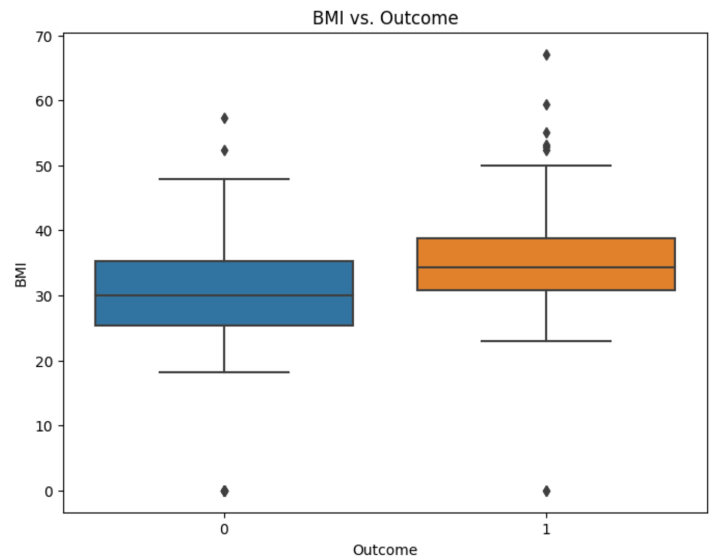


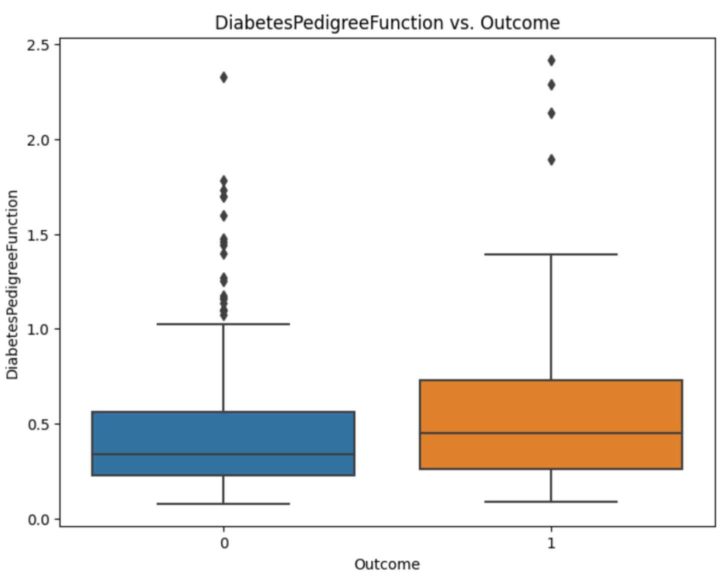












Part2: Preprocessing & Feature Engineering

1. Handling missing values
   Filled the mising values with medium of each column.
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 Base model result

```
Type 1 Error (False Positive Rate):
Logistic Regression: 0.2222222222222222
Random Forest: 0.1717171717171717

Type 2 Error (False Negative Rate):
Logistic Regression: 0.32727272727272727
Random Forest: 0.2909090909090909

Classification Report logreg:               precision    recall  f1-score   support

           0       0.81      0.78      0.79        99
           1       0.63      0.67      0.65        55

    accuracy                           0.74       154
   macro avg       0.72      0.73      0.72       154
weighted avg       0.75      0.74      0.74       154

Classification Report rf:               precision    recall  f1-score   support

           0       0.84      0.83      0.83        99
           1       0.70      0.71      0.70        55

    accuracy                           0.79       154
   macro avg       0.77      0.77      0.77       154
weighted avg       0.79      0.79      0.79       154
```

Random Forest has less Type1 and type2 errors and a higher F1 score.

3.2 Tuned model result

```
Type 1 Error (False Positive Rate):
Logistic Regression: 0.2222222222222222
Random Forest: 0.18181818181818182

Type 2 Error (False Negative Rate):
Logistic Regression: 0.34545454545454546
Random Forest: 0.2909090909090909

Classification Report logreg:               precision    recall  f1-score   support

           0       0.80      0.78      0.79        99
           1       0.62      0.65      0.64        55

    accuracy                           0.73       154
   macro avg       0.71      0.72      0.71       154
weighted avg       0.74      0.73      0.74       154

Classification Report rf:               precision    recall  f1-score   support

           0       0.84      0.82      0.83        99
           1       0.68      0.71      0.70        55

    accuracy                           0.78       154
   macro avg       0.76      0.76      0.76       154
weighted avg       0.78      0.78      0.78       154
```

Random Forest has less Type1 and type2 errors and a higher F1 score.

Part4: Conclusion

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

1. Logistic Regression and Random Forest were developed as predictive models for diabetes outcome. The base Random Forest model shows a better F1 score with less Type1 and Type2 errors. I tried to tune both the model and the Fi scores and Type1, Type2 error are worse than the original model, so the original Random Forest model is the best model for predicting diabetes in my analysis.
2. Based on the correlation heatmap, Glucose is the most siginificant predictor of disbetes outcome. Also, age, BMI and pregnancy play important roles.
3. Proper preprocessing steps, including feature scaling, one-hot-encoding significantly improved the model's performance in this case.
4. The dataset shows an imbalanced distribution, with a higher number of non-diabetic cases compared to diabetic cases.