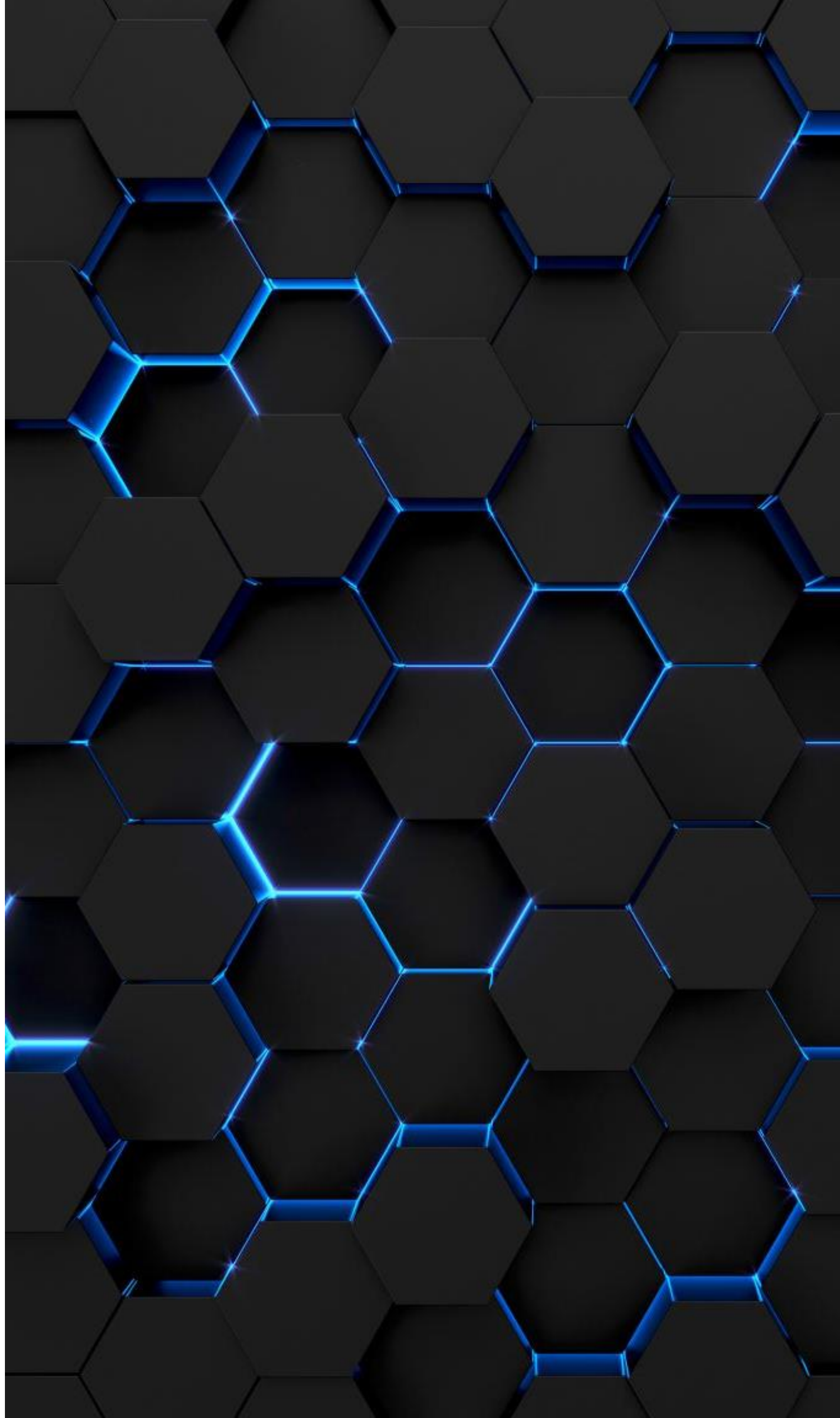
SUPERVISED LEARNING PROJECT

Presented by Mehzabeen Gheesah

May 29 DS



PROJECT GOAL

The ultimate goal of the project is to gain insights from the dataset and communicate these insights to stakeholders, using appropriate visualizations and metrics to make informed decisions based on the business questions asked.

PROJECT STEPS

Exploratory Data Analysis

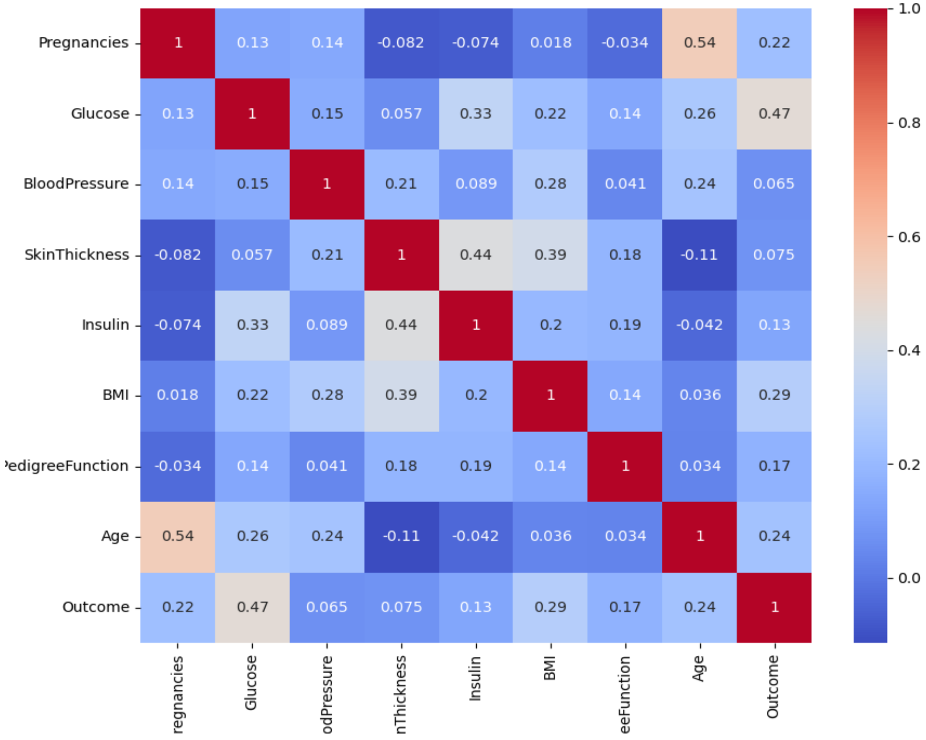
Preprocessing and Feature Engineering

Training a Machine Learning Model.

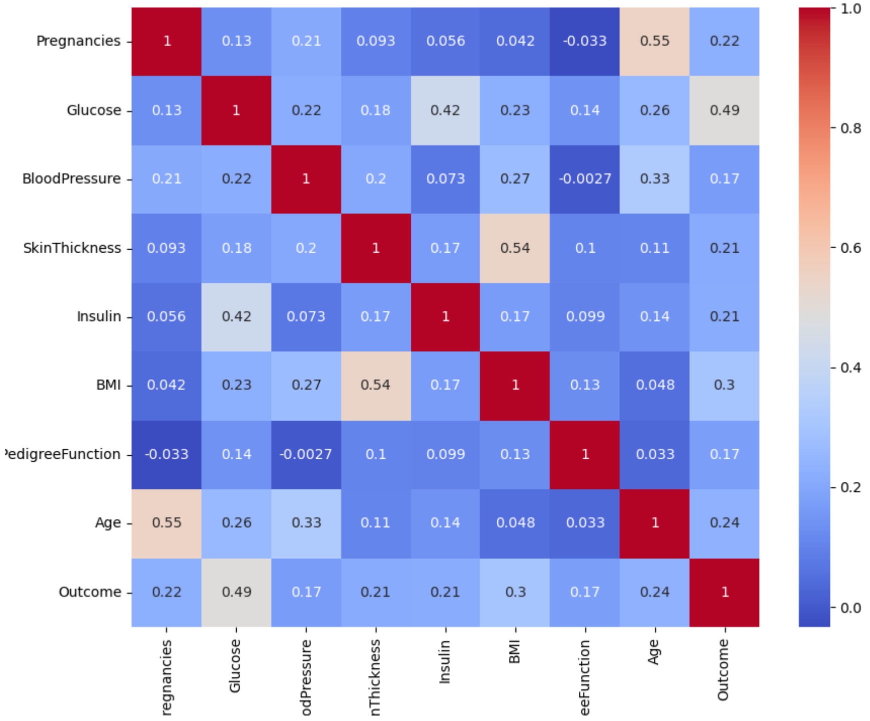
KEY FACTS

- Glucose, BMI , Age, Insulin, Skin Thickness and Pregnancies have a correlation with being diabetic
- Number of diabetics 268 and non-diabetics 500 – imbalanced dataset
- As Glucose, Age and BMI get higher, there is a tendency for people to become positive to diabetes
- Outliers detected in BMI, and Skin Thickness will be treated accordingly

HEATMAP TO SHOW CORRELATION BETWEEN VARIABLES

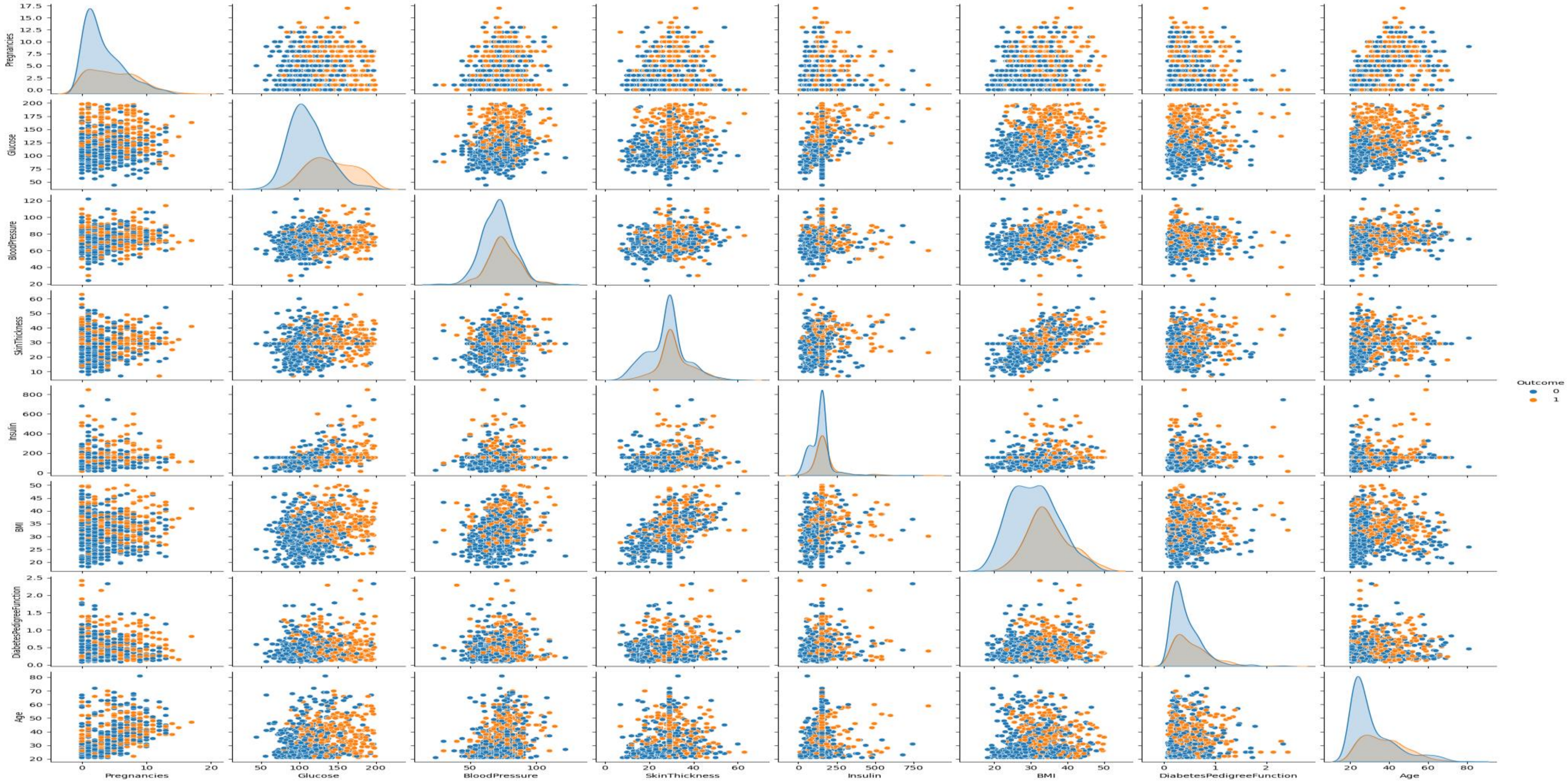


Before data cleaning



After data cleaning

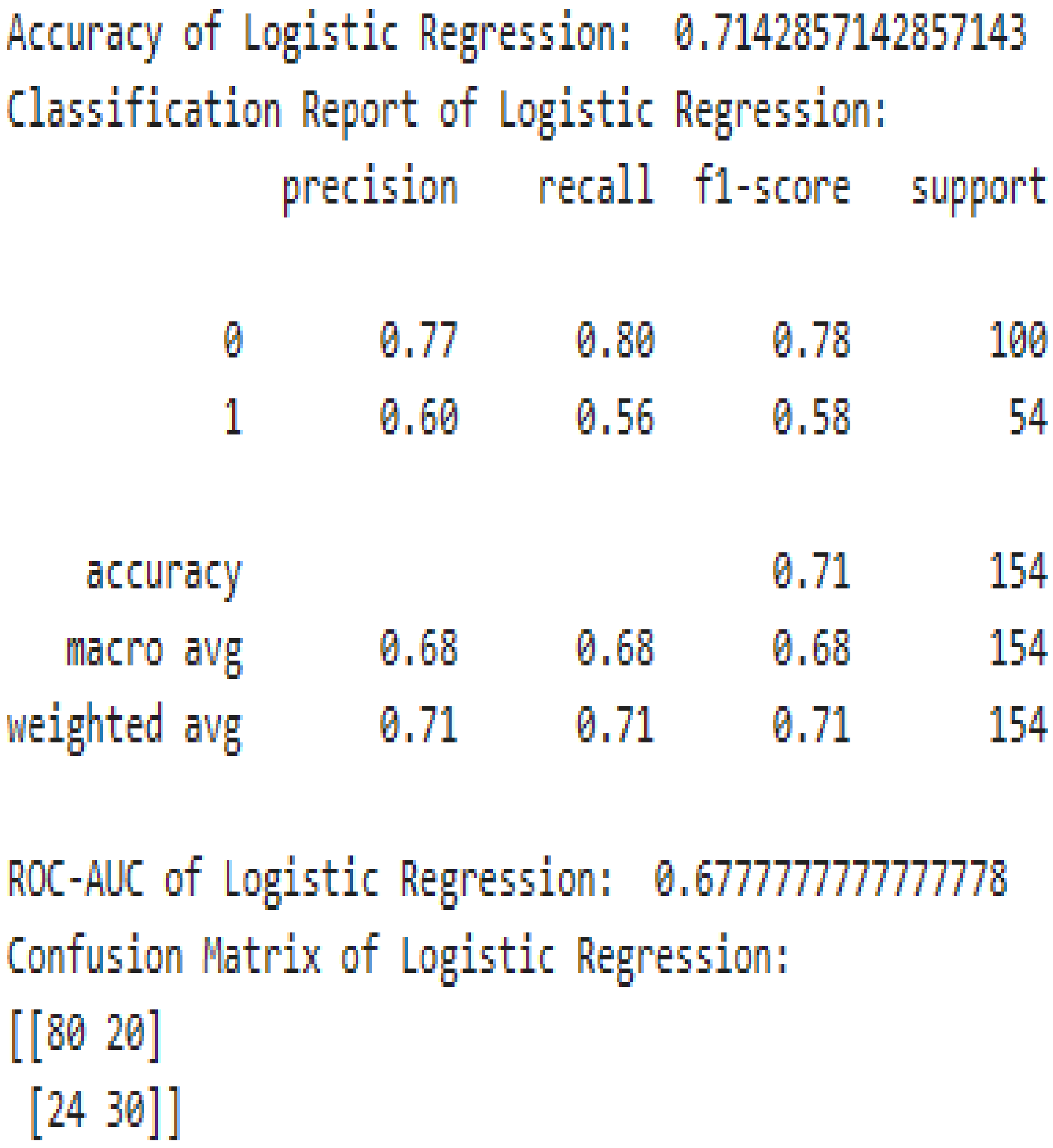
THE CORRELATIONS BETWEEN PREDICTOR VARIABLES, THE DATA DISTRIBUTION, AND POTENTIAL RELATIONSHIPS WITH THE TARGET VARIABLE



MODELS USED

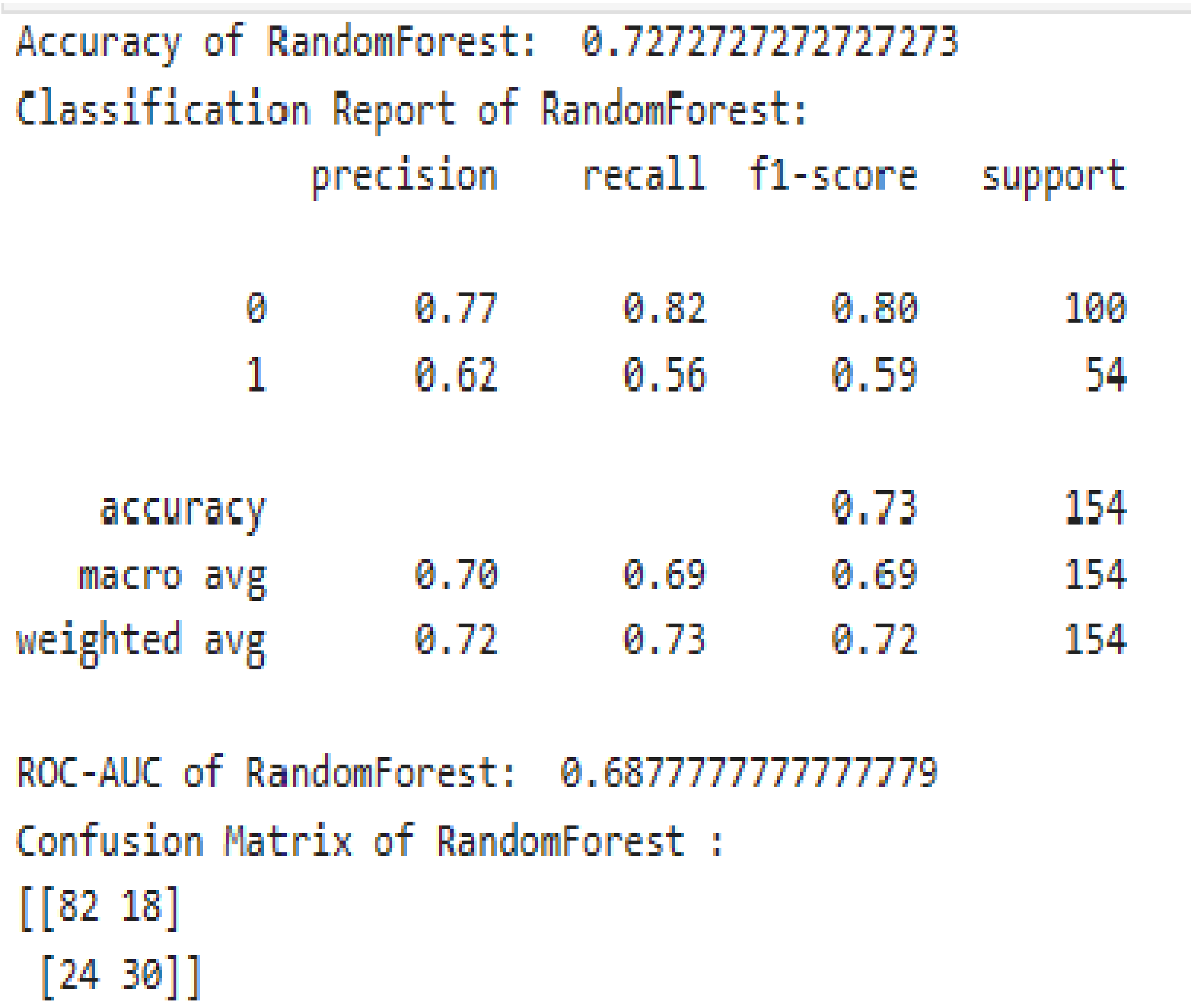
- Logistic Regression Model
- RandomForest Classifier
- XGBoost Model

LOGISTIC REGRESSION REPORT



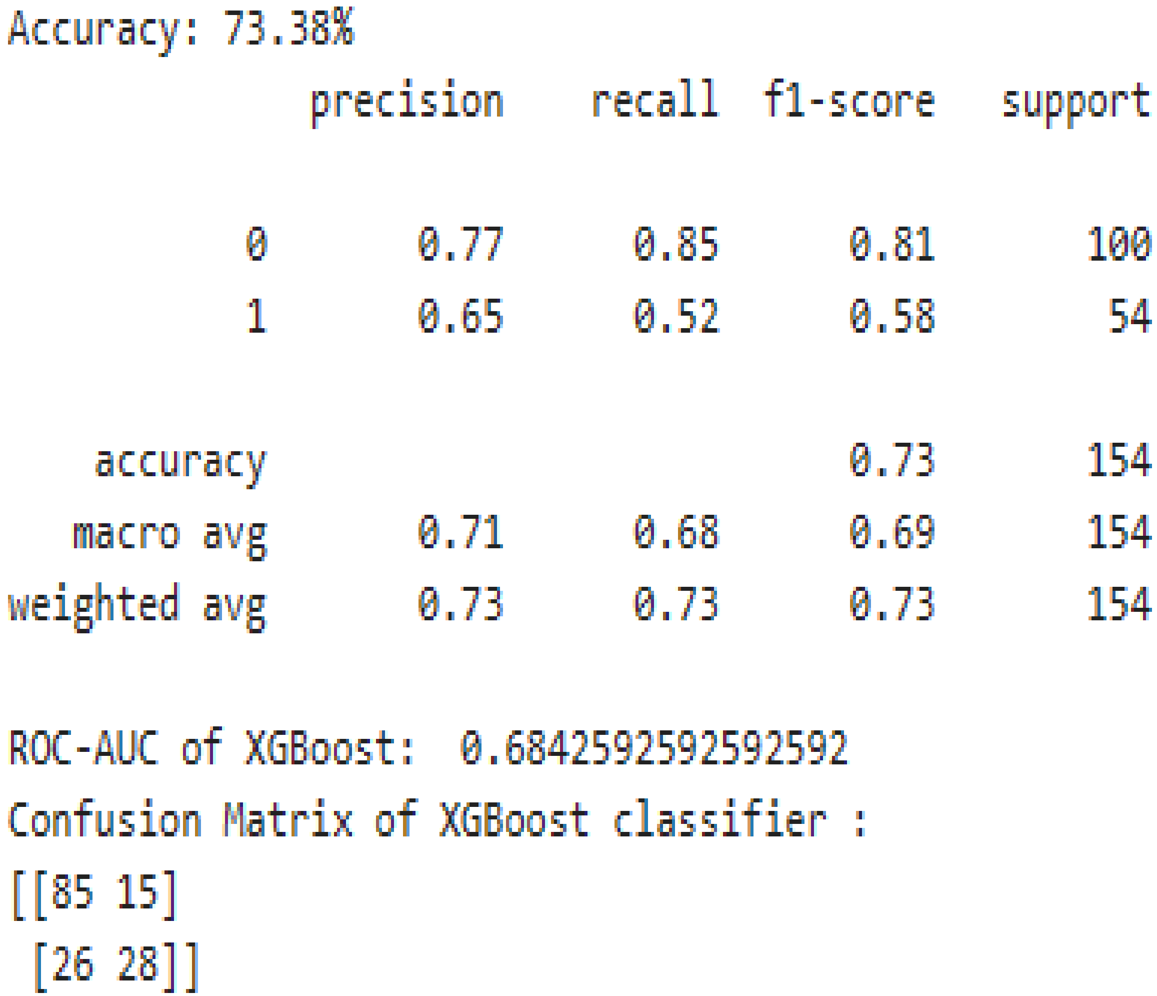
```
Accuracy of Logistic Regression:  0.7142857142857143
Classification Report of Logistic Regression:
              precision    recall  f1-score   support

           0       0.77      0.80      0.78       100
           1       0.60      0.56      0.58        54

    accuracy                           0.71       154
   macro avg       0.68      0.68      0.68       154
weighted avg       0.71      0.71      0.71       154

ROC-AUC of Logistic Regression:  0.6777777777777778
Confusion Matrix of Logistic Regression:
[[80 20]
 [24 30]]
```

RANDOMFOREST ACCURACY REPORT



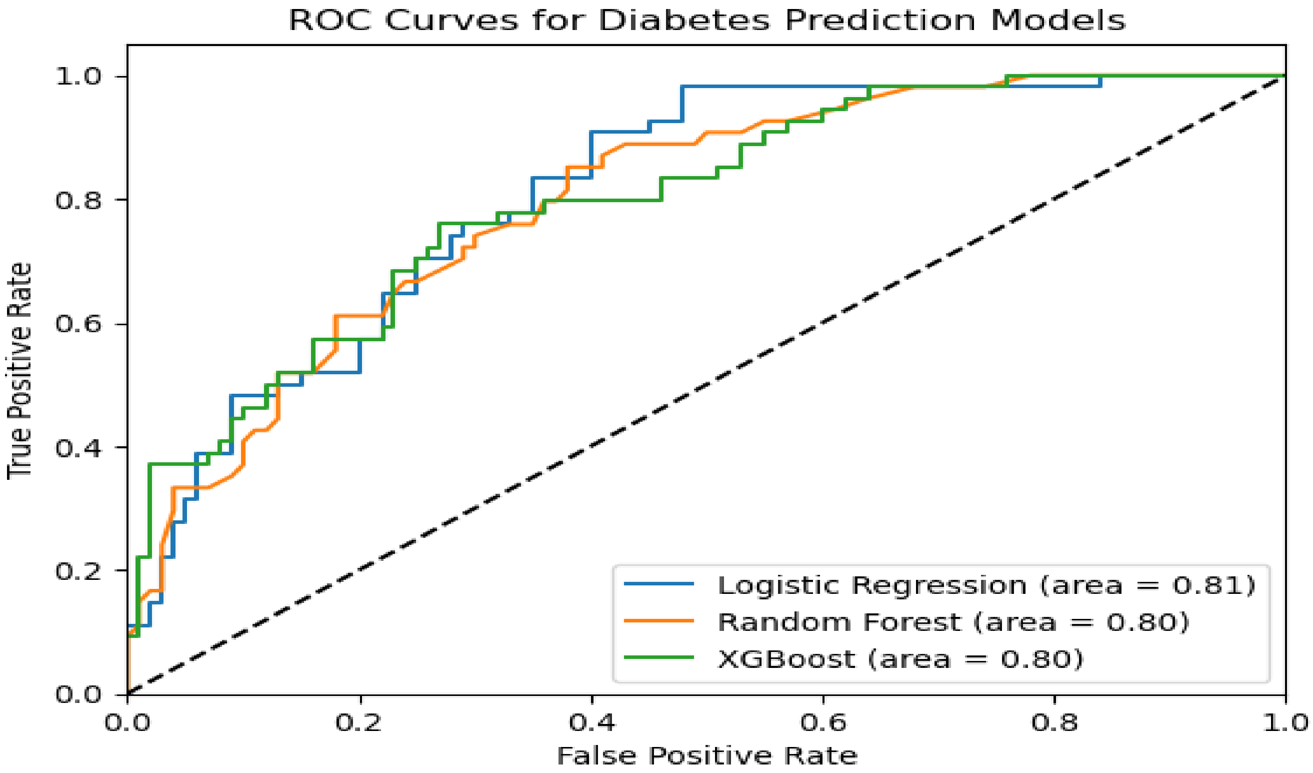
```
Accuracy of RandomForest:  0.7272727272727273
Classification Report of RandomForest:
              precision    recall  f1-score   support

           0       0.77      0.82      0.80       100
           1       0.62      0.56      0.59        54

    accuracy                           0.73       154
   macro avg       0.70      0.69      0.69       154
weighted avg       0.72      0.73      0.72       154

ROC-AUC of RandomForest:  0.6877777777777779
Confusion Matrix of RandomForest :
[[82 18]
 [24 30]]
```

XGBOOST CLASSIFIER REPORT

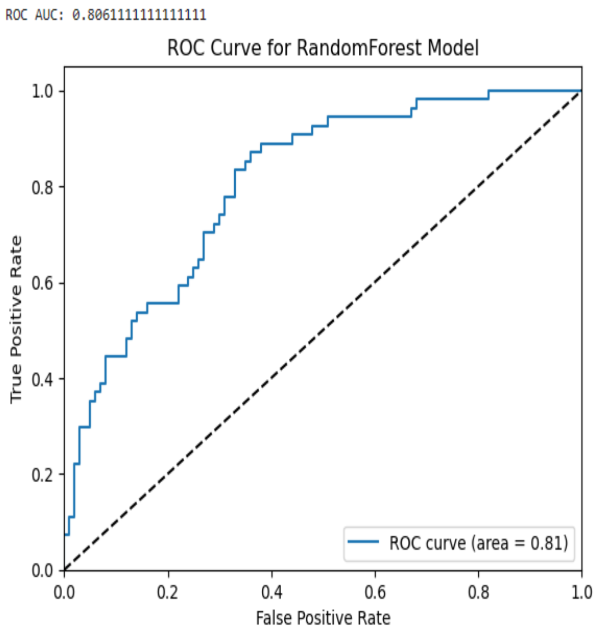


```
Accuracy: 73.38%
              precision    recall  f1-score   support

           0       0.77      0.85      0.81       100
           1       0.65      0.52      0.58        54

    accuracy                           0.73       154
   macro avg       0.71      0.68      0.69       154
weighted avg       0.73      0.73      0.73       154

ROC-AUC of XGBoost:  0.6842592592592592
Confusion Matrix of XGBoost classifier :
[[85 15]
 [26 28]]
```

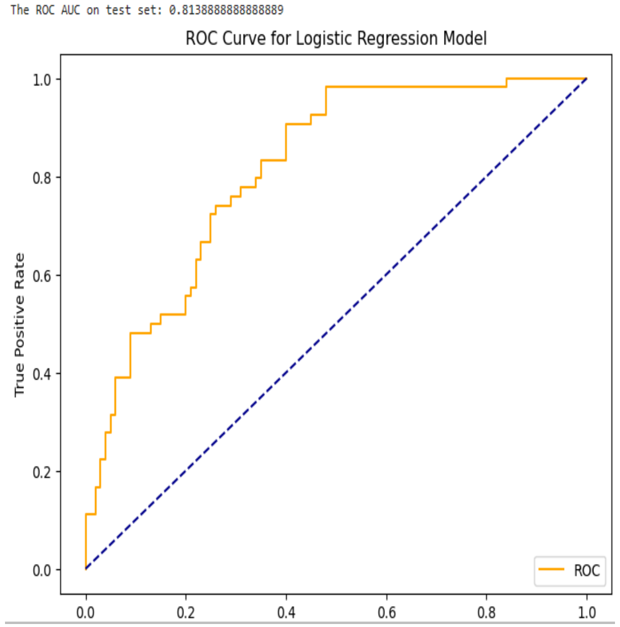
COMPARISON OF ROC CURVES FOR ALL 3 MODELS



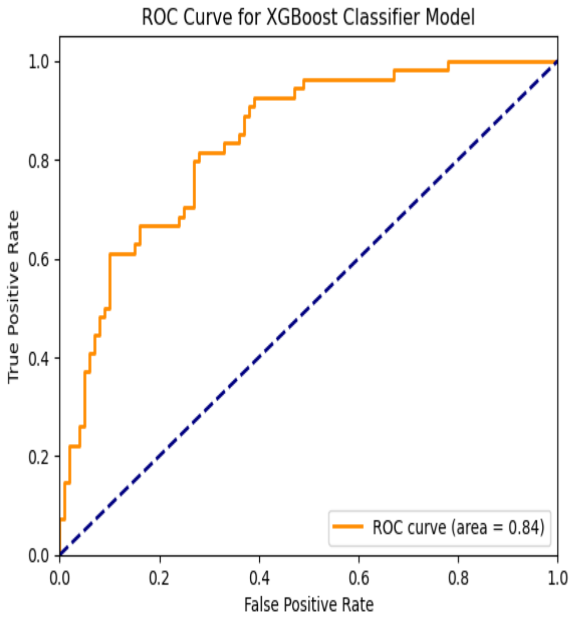
AFTER HYPERPARAMETER TUNING



```
ROC AUC: 0.8061111111111111
```



```
The ROC AUC on test set: 0.8138888888888889
```



CONCLUSION

In this analysis, three distinct models were applied: Logistic Regression, Random Forest, and XGBoost. Each model was selected based on its suitability for the binary classification task at hand, that is to predict whether a person is diabetic or non-diabetic based on the predictor variables. The Random Forest and XGBoost represent the ensemble category. Initial evaluations were based on the ROC AUC scores, and then the models were subjected to further refinement through hyperparameter tuning, where techniques such as GridSearchCV were employed. Post tuning, XGBoost emerged as the best model, reflecting a superior balance of bias and variance, and thereby offering the highest predictive accuracy for the dataset. This showcases the importance of exploring multiple models and tuning to optimize the performance of machine learning tasks.