




PREDICTING DIABETES



INTRODUCTION

- I worked with a diabetic dataset to develop a predictive model that can determine the likelihood of a patient developing diabetes.
- Diabetes is a chronic medical condition that affects millions of people worldwide, and early prediction can play a crucial role in its management and prevention.

- 
- The dataset contains various medical measurements for patients, including attributes like glucose level, blood pressure, BMI, age, and more.
 - Each record also includes a binary indicator of diabetes status (1 for diabetes, 0 for non-diabetes).

Data Collection and Preprocessing

- **Handling Missing Values:**

Checked for missing values and decided on strategies to handle them (e.g., mean imputation, dropping rows).

- **Feature Scaling:**

Scaled numerical features to ensure all features contribute equally to the model.

- **Encoding Categorical Variables:**

Encoded categorical variables if any (not applicable in this dataset).

- **Splitting the Data:**

Split the data into training and testing sets for model evaluation.

Tools and Libraries

Pandas: Data manipulation and analysis.

NumPy: Numerical computing.

Matplotlib: Data visualization.

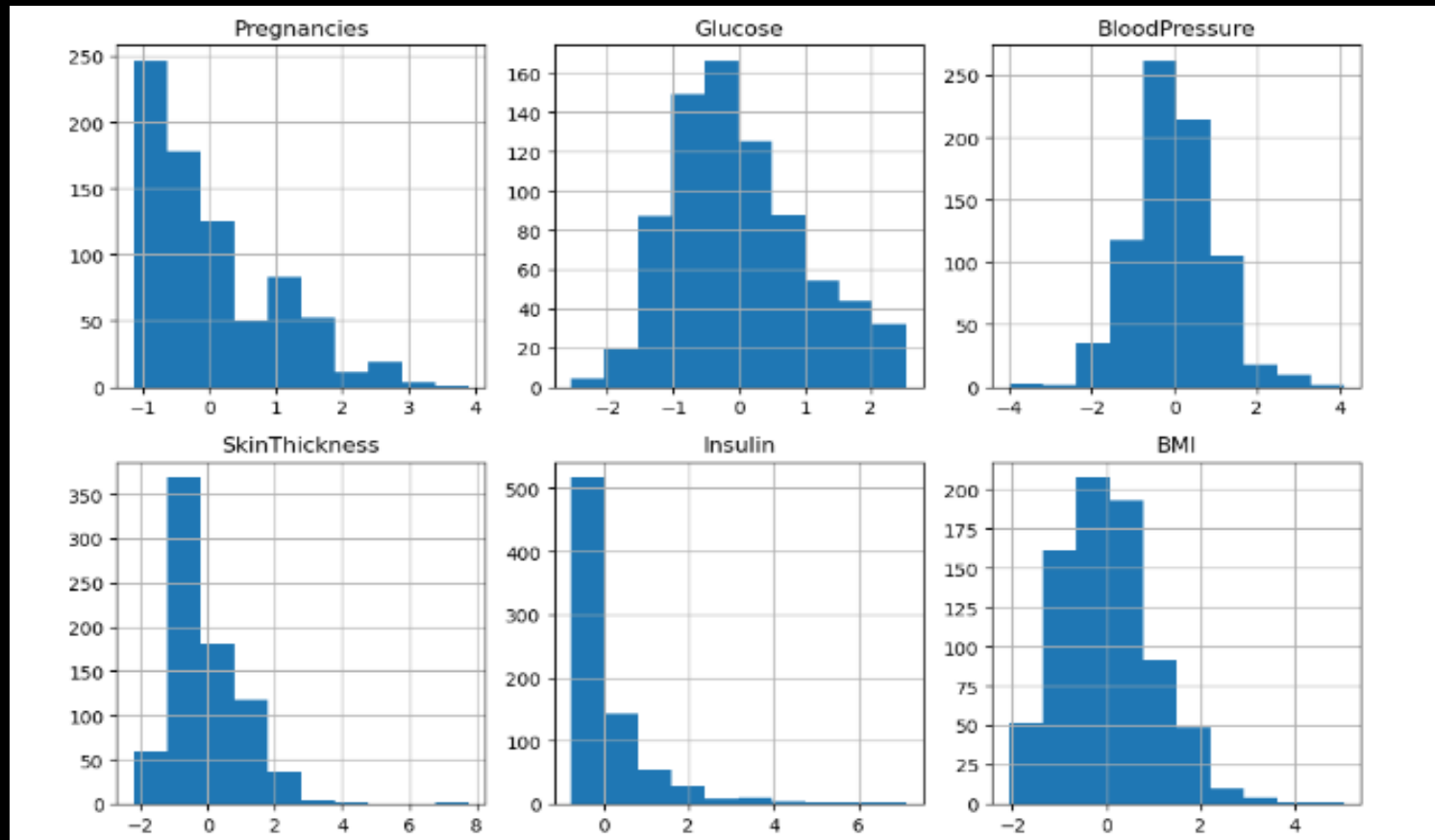
Seaborn: Statistical data visualization.

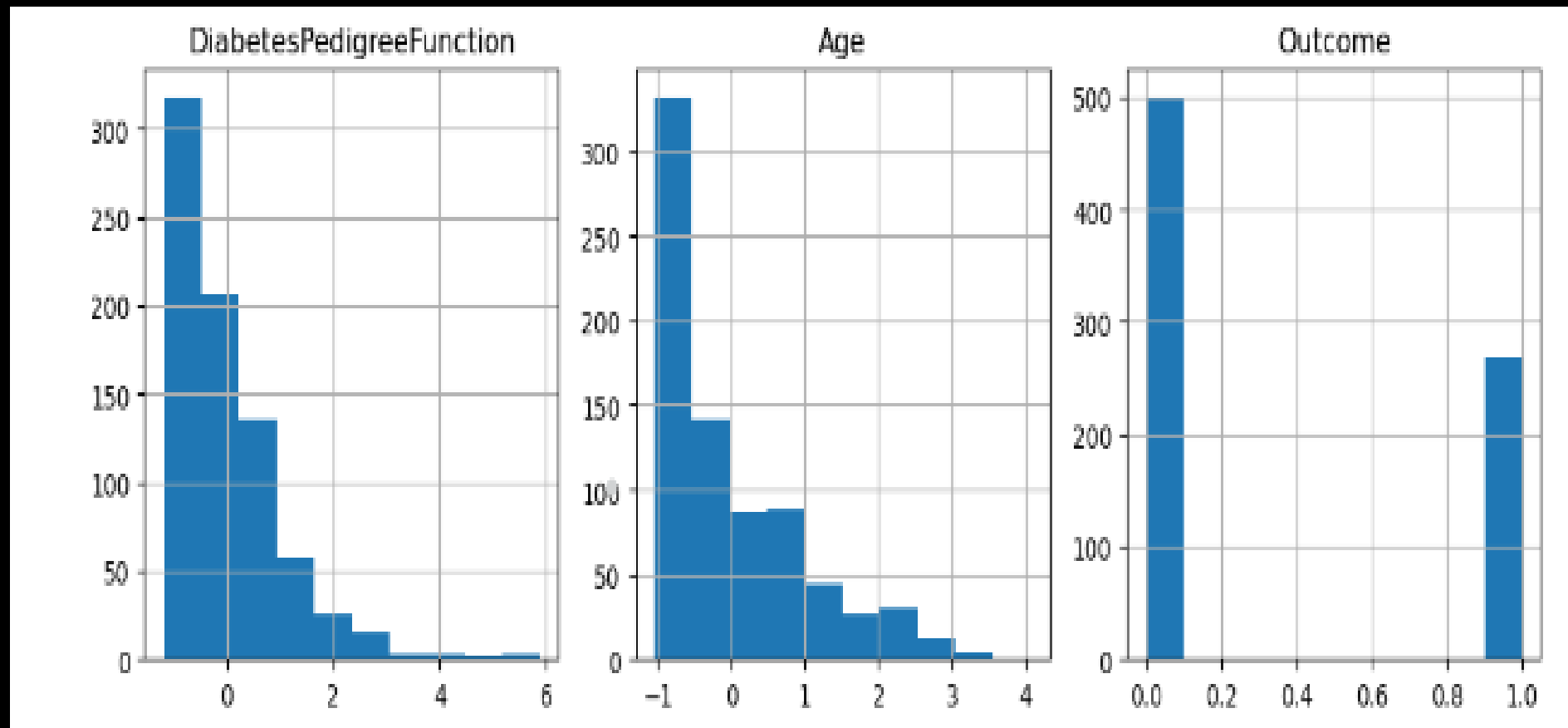
Scikit-learn: Machine learning library for model building and evaluation.

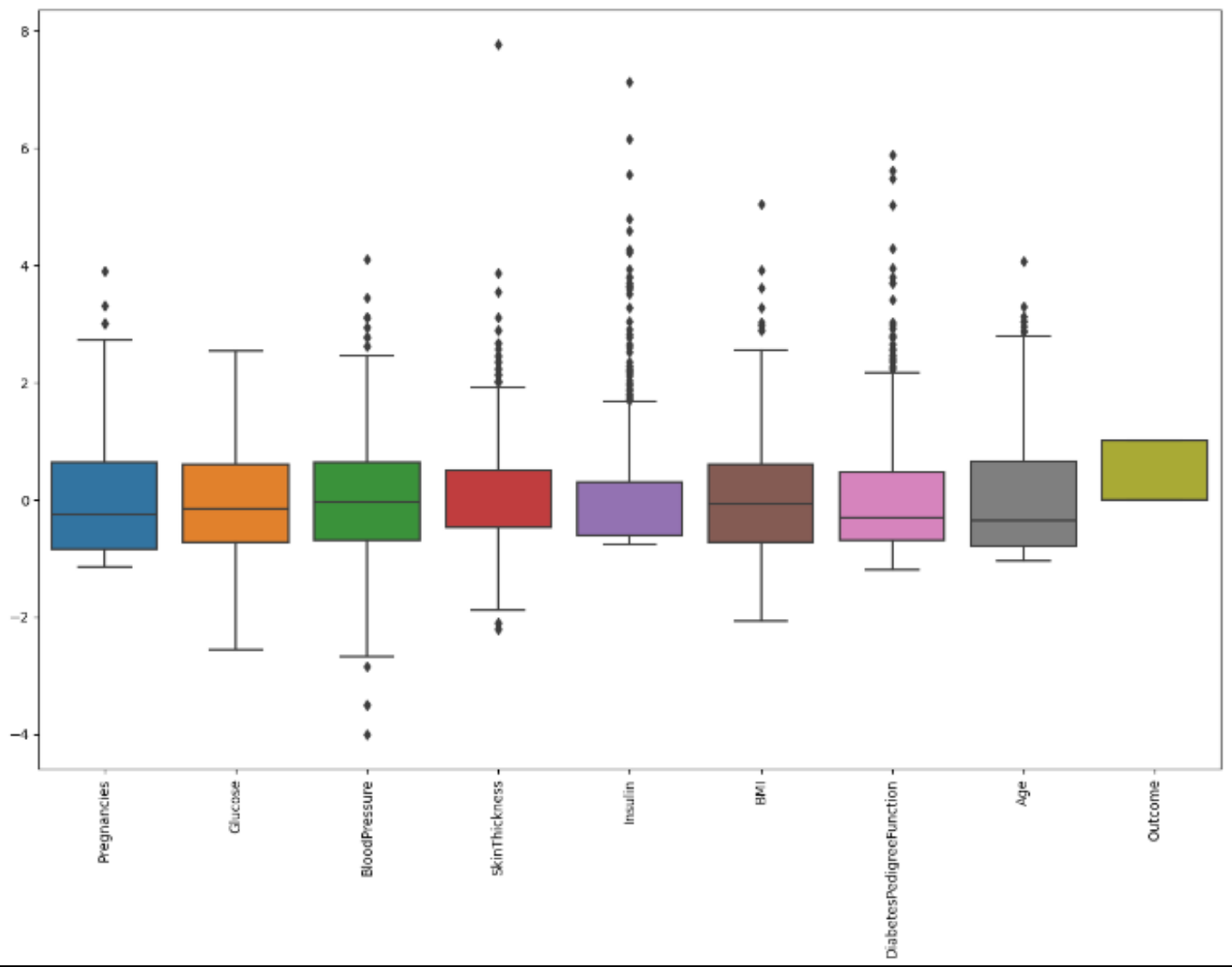
Jupyter Notebook: Interactive computing environment

Exploratory Data Analysis

Describe any significant patterns or insights gained during EDA By using various Histograms, box plots, pair plots.





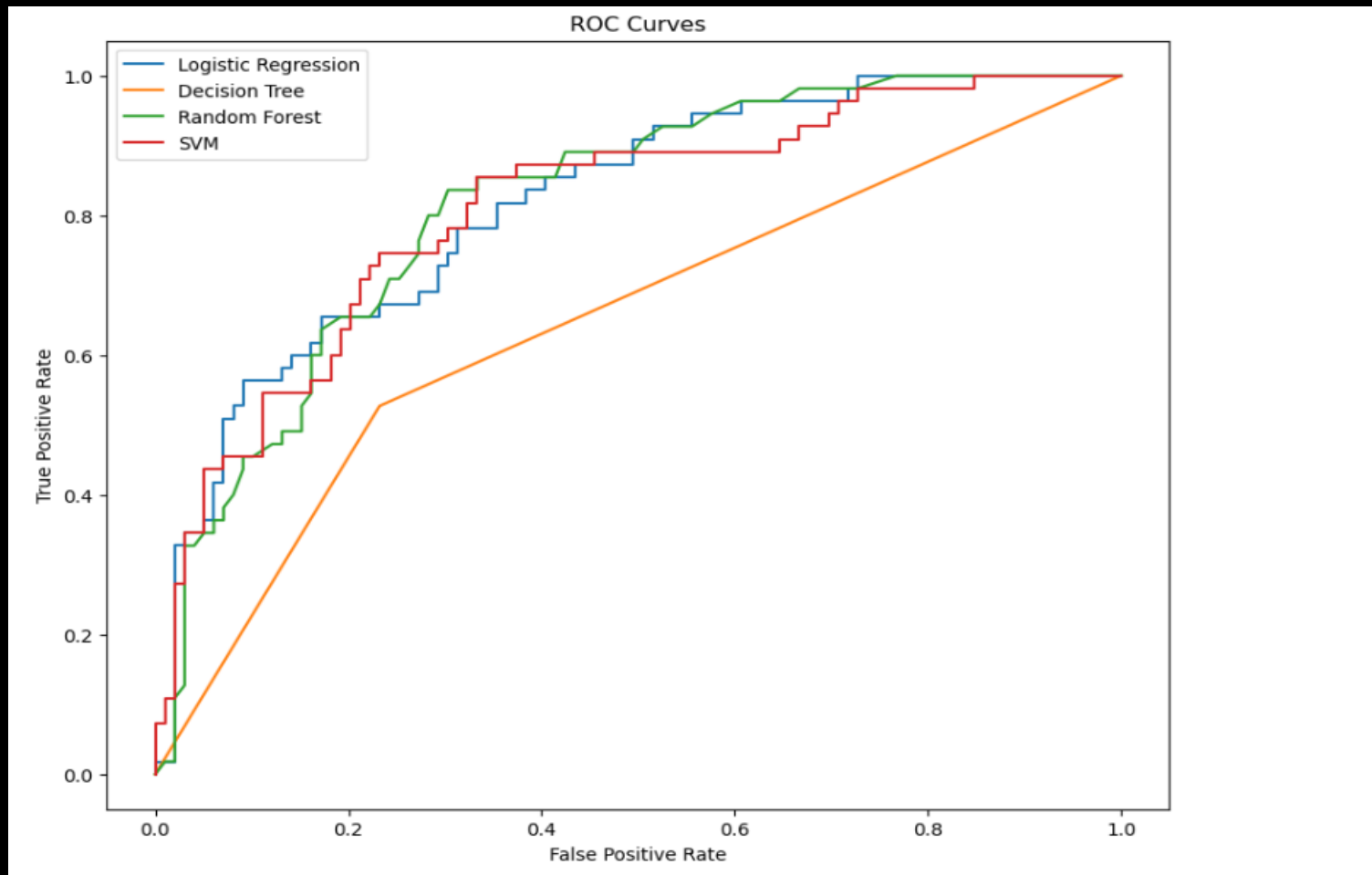


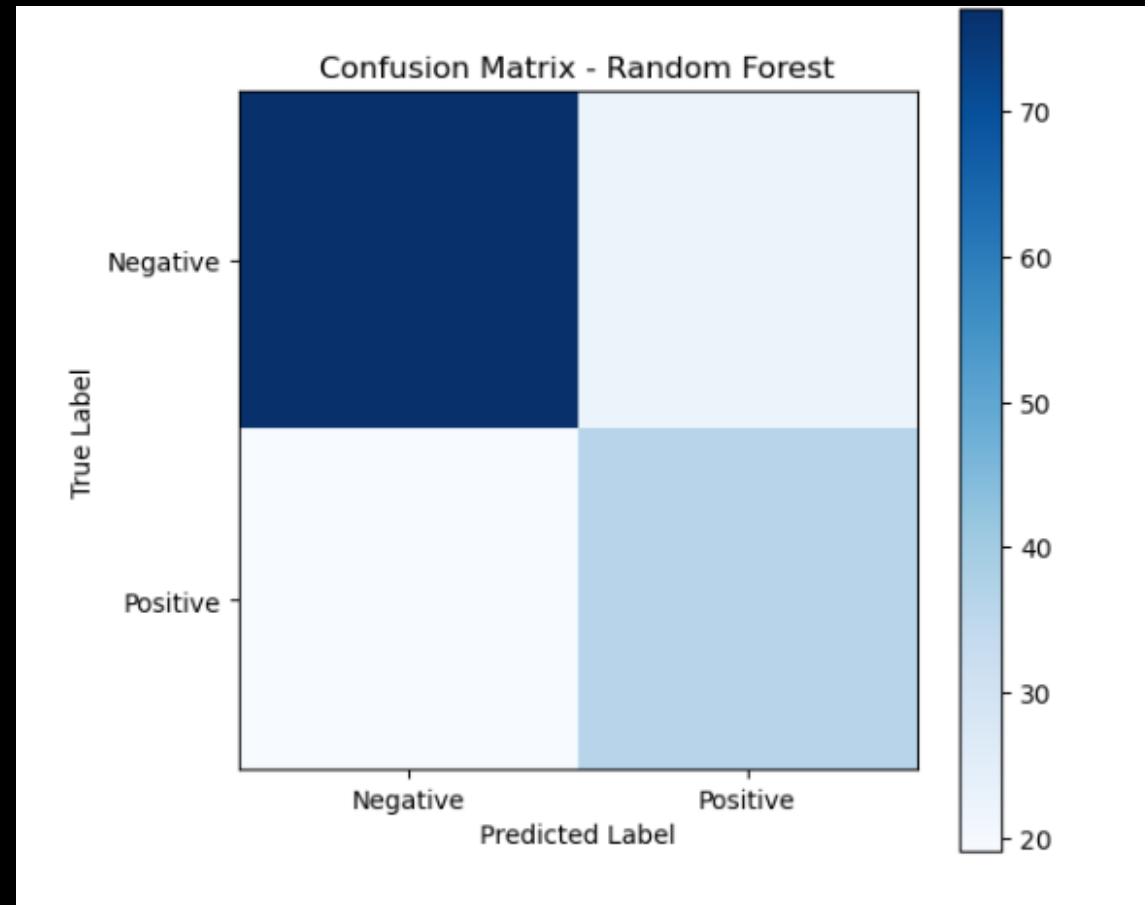
Model Building

Accuracy, precision, recall, F1-score of the model's performance

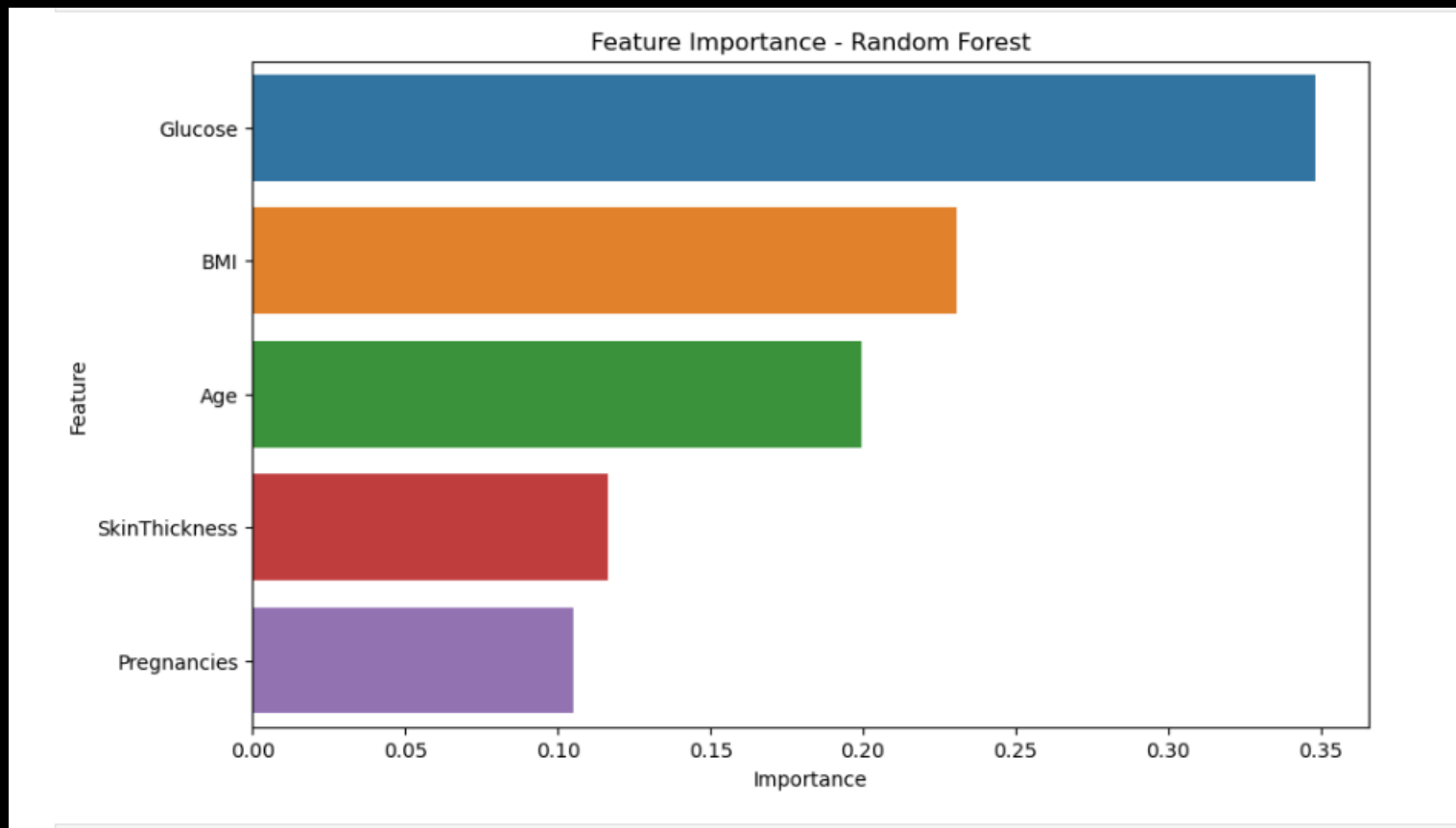
```
{ 'Logistic Regression': { 'Accuracy': 0.7597402597402597,  
  'Precision': 0.6730769230769231,  
  'Recall': 0.6363636363636364,  
  'F1 Score': 0.6542056074766355,  
  'ROC AUC': 0.8170798898071625 },  
  'Decision Tree': { 'Accuracy': 0.6818181818181818,  
    'Precision': 0.5576923076923077,  
    'Recall': 0.5272727272727272,  
    'F1 Score': 0.5420560747663552,  
    'ROC AUC': 0.6474747474747475 },  
  'Random Forest': { 'Accuracy': 0.7337662337662337,  
    'Precision': 0.6206896551724138,  
    'Recall': 0.6545454545454545,  
    'F1 Score': 0.6371681415929203,  
    'ROC AUC': 0.8165289256198347 },  
  'SVM': { 'Accuracy': 0.7337662337662337,  
    'Precision': 0.64,  
    'Recall': 0.5818181818181818,  
    'F1 Score': 0.6095238095238096,  
    'ROC AUC': 0.8104683195592286 } }
```

Model Performance Visualization





Feature Importance



Conclusion

The project's objective was to develop a predictive model for diabetes using various medical measurements. The Random Forest model, tested on the diabetes dataset, emerged as the best performer.

Regular monitoring of glucose levels is crucial as they are a strong diabetes indicator. Older age increases diabetes risk. BMI and pregnancies are also important factors in prediction and prevention strategies.