SUPERVISED LEARNING PROJECT DIABETES DATASET

Lat Leger

Supervised Learning - Project

- Use supervised learning techniques to build a machine learning model that can predict whether a patient has diabetes or not, based on certain diagnostic measurements.
- The project involves three main parts:
    * Exploratory data analysis
    * Preprocessing and feature engineering
    * Training a machine learning model

Part I : EDA - Exploratory Data Analysis

- For this task, you are required to conduct an exploratory data analysis on the diabetes dataset.
- You have the freedom to choose the visualizations you want to use.

Q: Are there any missing values in the dataset?

A: No, there are no missing values. However, Glucose, BloodPressure, SkinThickness, Insulin, BMI have 0 values, which is not physically possible. I replaced the Zeros with the mean of each column.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Q: How are the predictor variables related to the outcome variable?

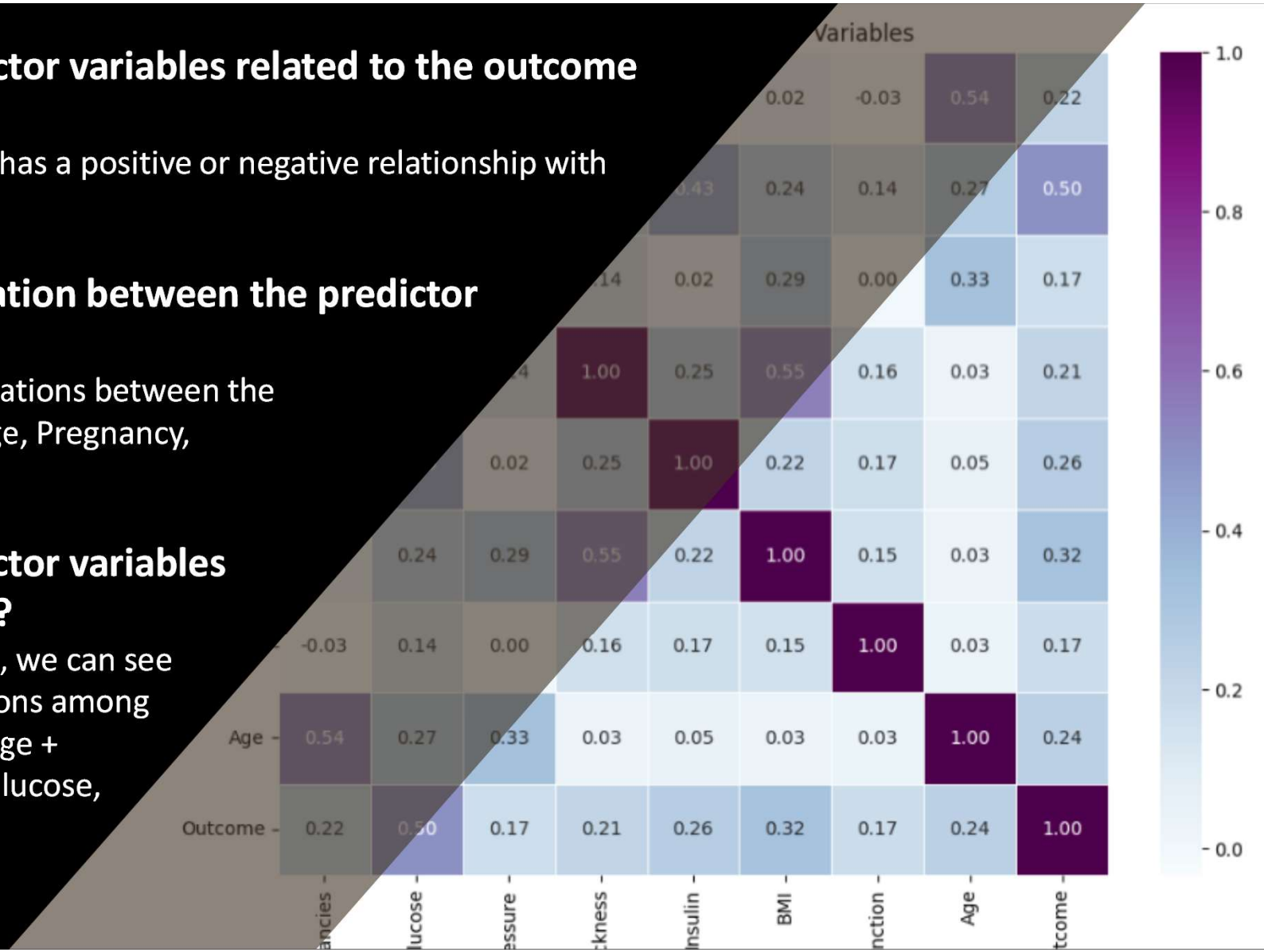
A: Each predictor variable has a positive or negative relationship with the outcome variable.

Q: What is the correlation between the predictor variables?

A: There are several correlations between the predictor variables, like Age, Pregnancy, Insulin, etc.

Q: How are the predictor variables related to each other?

Using a Correlation Matrix, we can see there are several correlations among predictors. For example: Age + Pregnancies, Outcome + Glucose, BMI + SkinThickness, and Insulin + Glucose.



Q: What is the distribution of each predictor variable?

A: Using the describe() method, we can see the distribution of each predictor variable:

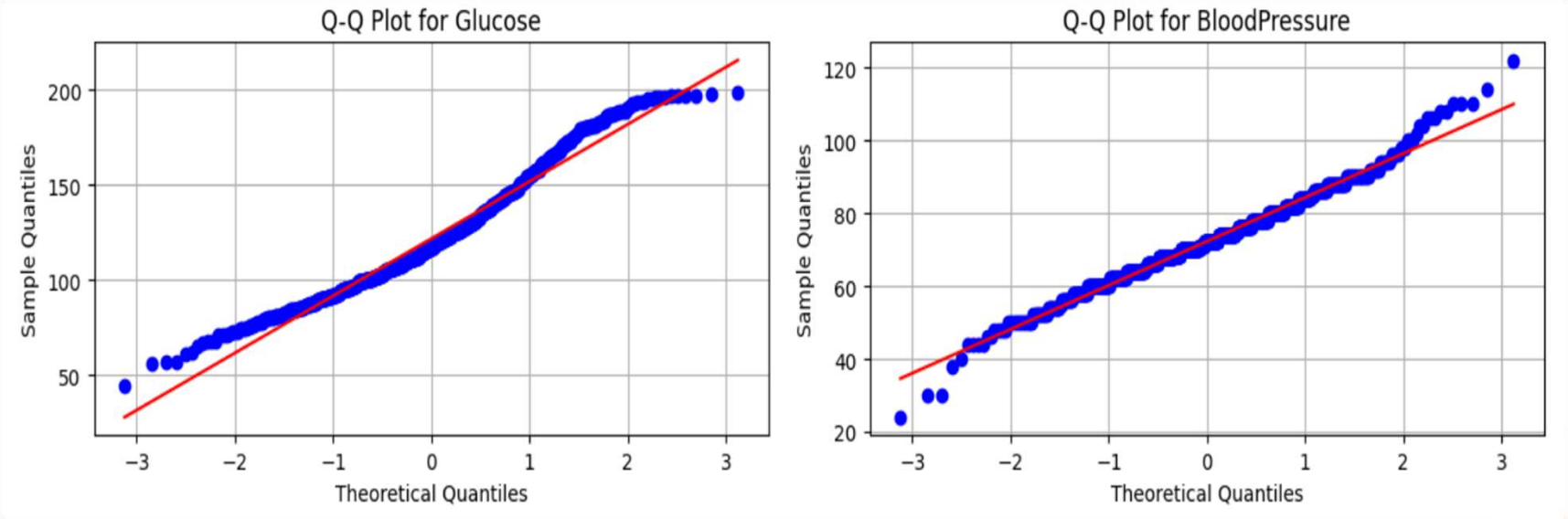
1. **Pregnancies:** Distribution: Appears to be positively skewed, as the mean is slightly greater than the median.
2. **Glucose:** Distribution: Approximately normally distributed, as the mean is close to the median.
3. **BloodPressure:** Distribution: Approximately normally distributed, as the mean is close to the median.
4. **SkinThickness:** Distribution: Appears to be positively skewed, as the mean is slightly greater than the median.
5. **Insulin:** Distribution: Appears to be positively skewed, as the mean is slightly greater than the median.
6. **BMI:** Distribution: Approximately normally distributed, as the mean is close to the median.
7. **DiabetesPedigreeFunction:** Distribution: Appears to be positively skewed, as the mean is slightly greater than the median.
8. **Age:** Distribution: Appears to be positively skewed, as the mean is slightly greater than the median.

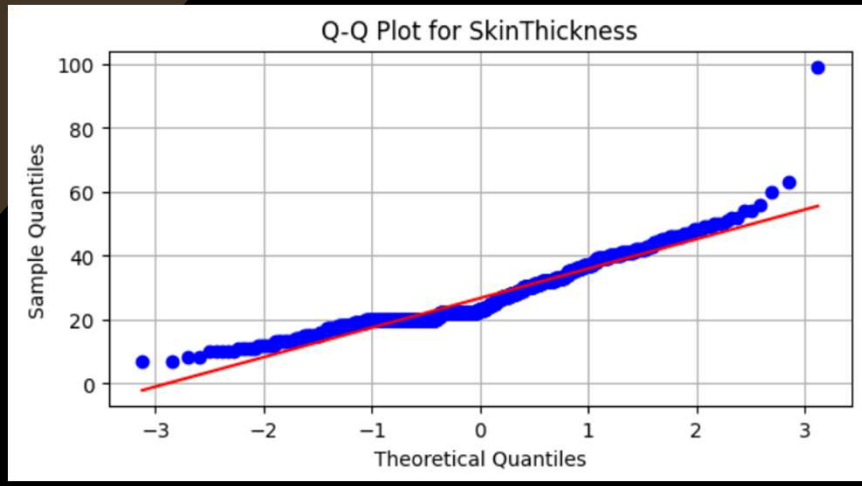
Using the describe() method:

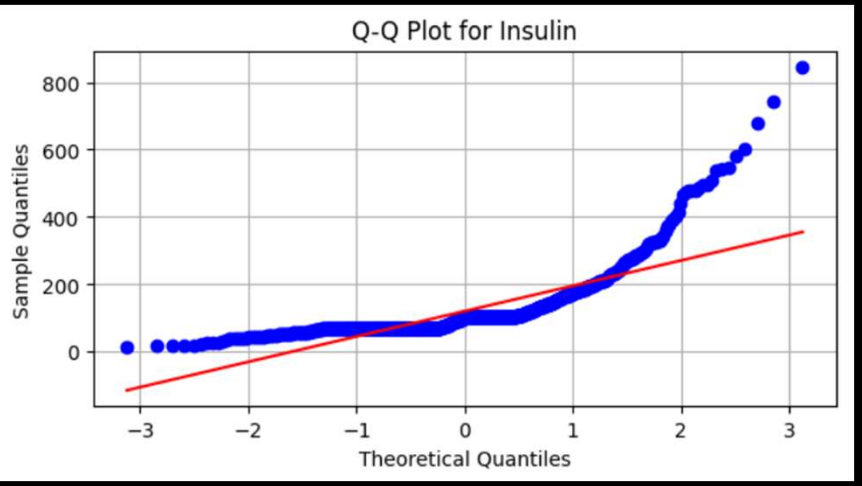
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	121.691999	72.267826	26.635083	118.967780	32.439222	0.471876	33.240885	0.348958
std	3.369578	30.461151	12.115948	9.636089	93.557899	6.880449	0.331329	11.760232	0.476951
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	19.664000	68.792000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	100.000000	32.050000	0.372500	29.000000	0.000000
75%	6.000000	141.000000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

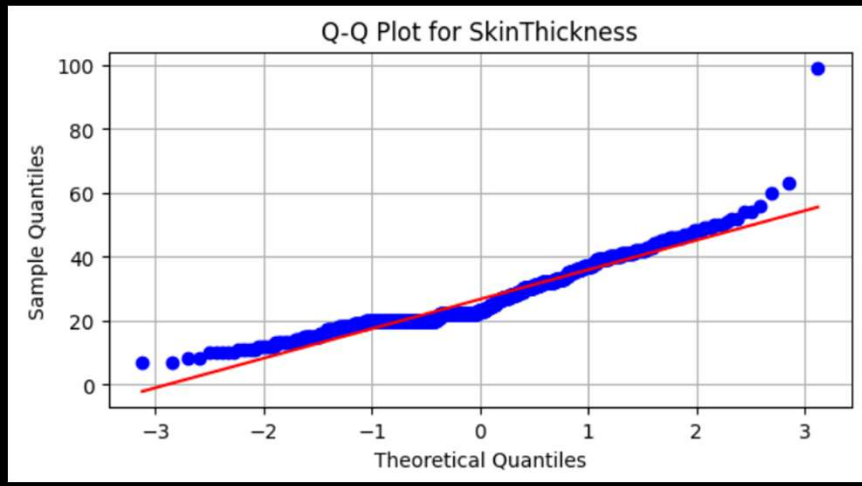
Q: Are there any outliers in the predictor variables?

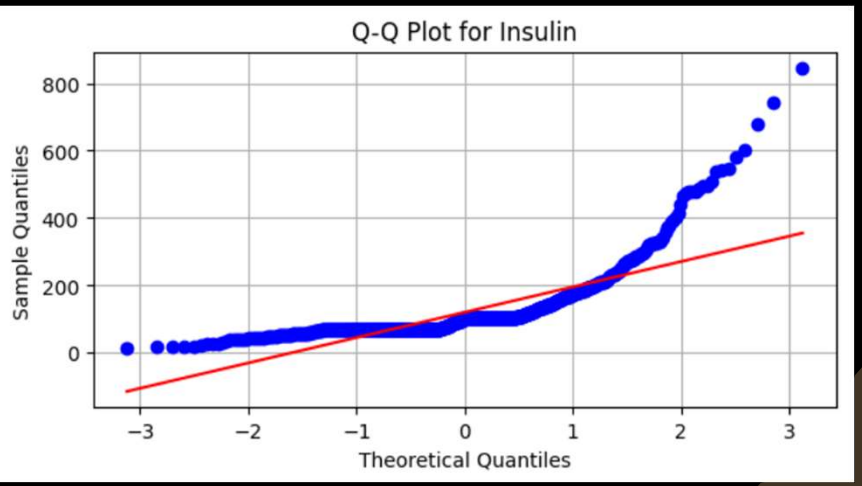
A: Yes, there are outliers. Using Q-Q Plots, we can Insulin has the most outliers.




Q-Q Plot for SkinThickness
Sample Quantiles
Theoretical Quantiles
100
80
60
40
20
0
−3
−2
−1
0
1
2
3


Q-Q Plot for Insulin
Sample Quantiles
Theoretical Quantiles
800
600
400
200
0
−3
−2
−1
0
1
2
3

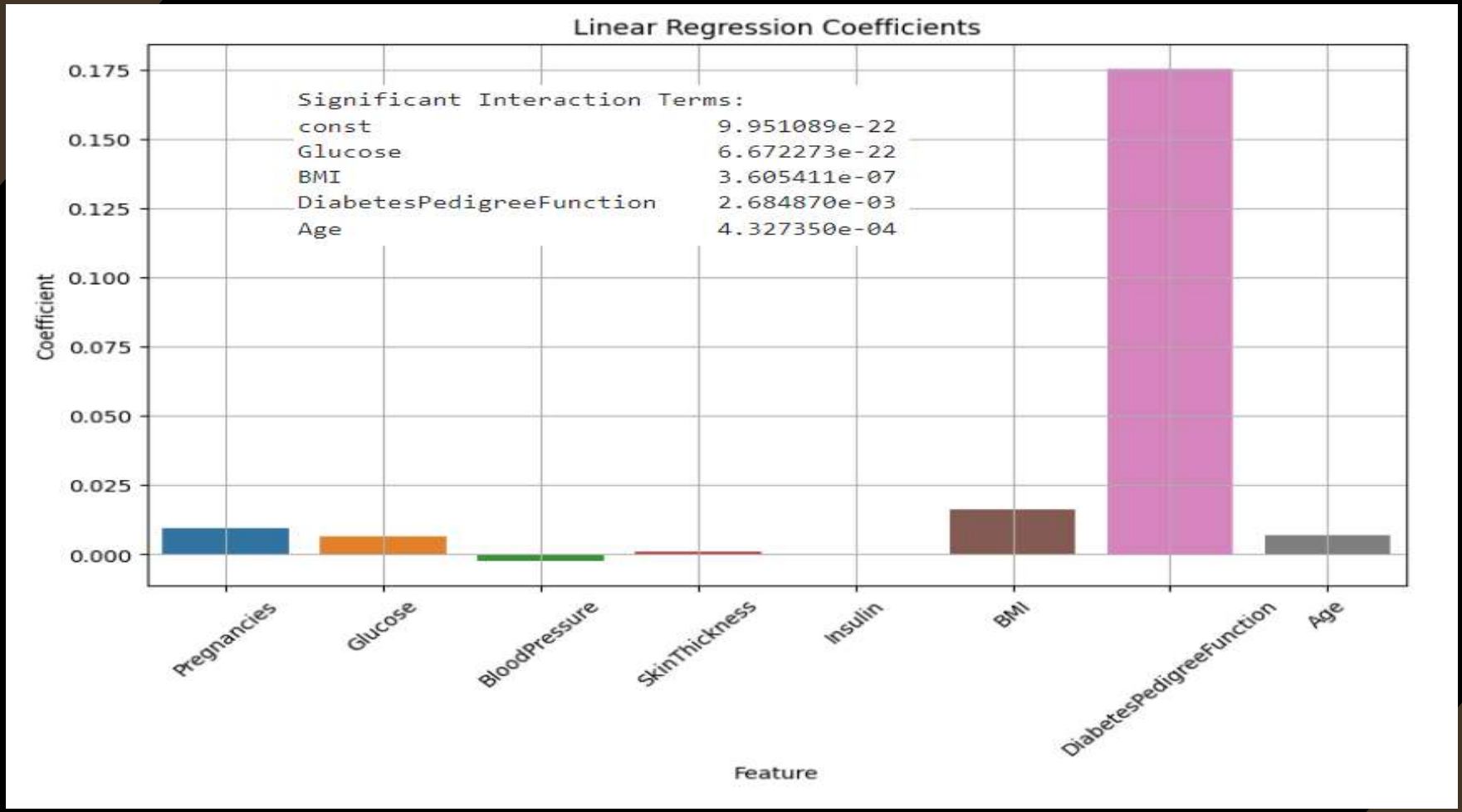

Q-Q Plot for SkinThickness
Sample Quantiles
Theoretical Quantiles
100
80
60
40
20
0
−3
−2
−1
0
1
2
3


Q-Q Plot for Insulin
Sample Quantiles
Theoretical Quantiles
800
600
400
200
0
−3
−2
−1
0
1
2
3

Q: Is there any interaction effect between the predictor variables?

A: Using a Multiple Linear Regression model to analyze the relationship between multiple predictor variables and the target variable 'Outcome’.

These results indicate that the interaction between these predictor variables and possibly other predictors in the model significantly affects the outcome variable. Therefore, considering these interactions is crucial for interpreting the relationship between the predictors and the target variable accurately.


Linear Regression Coefficients
Significant Interaction Terms:
const 9.951089e-22
Glucose 6.672273e-22
BMI 3.605411e-07
DiabetesPedigreeFunction 2.684870e-03
Age 4.327350e-04
Coefficient
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Feature

Q: HOW DOES THE DISTRIBUTION OF THE PREDICTOR VARIABLES DIFFER FOR INDIVIDUALS WITH DIABETES AND WITHOUT DIABETES?

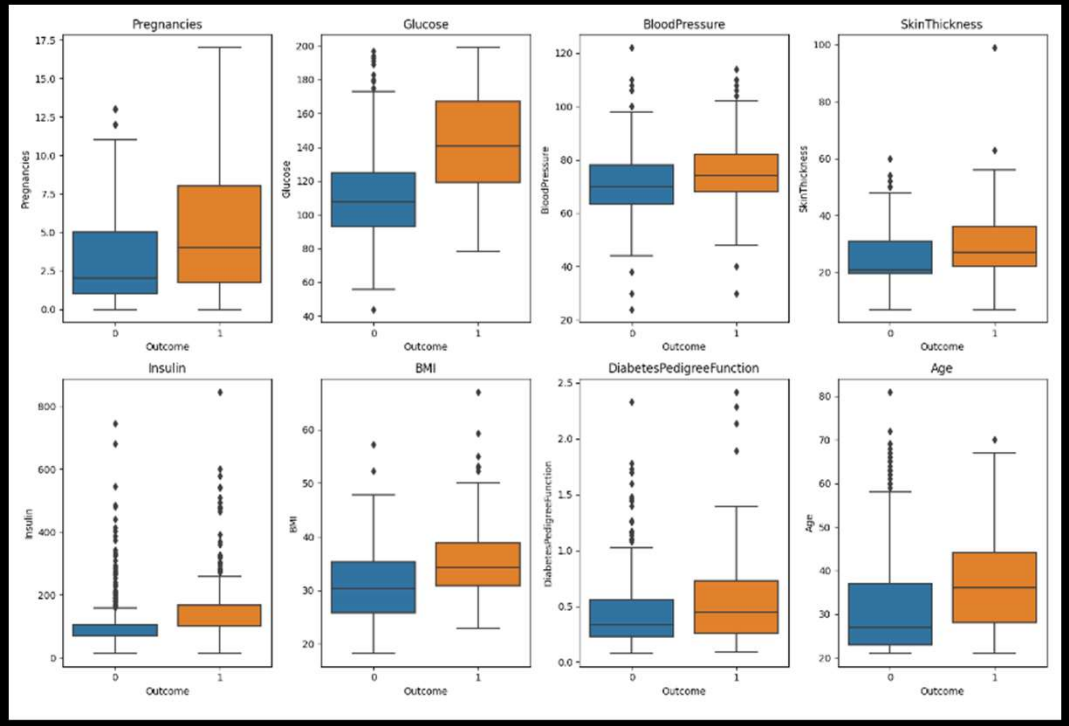
A: BASED ON THE BOX PLOTS, WE CAN SEE THAT THERE IS MORE DATA VARIABILITY IN THE DATA FOR INDIVIDUALS WITH DIABETES.

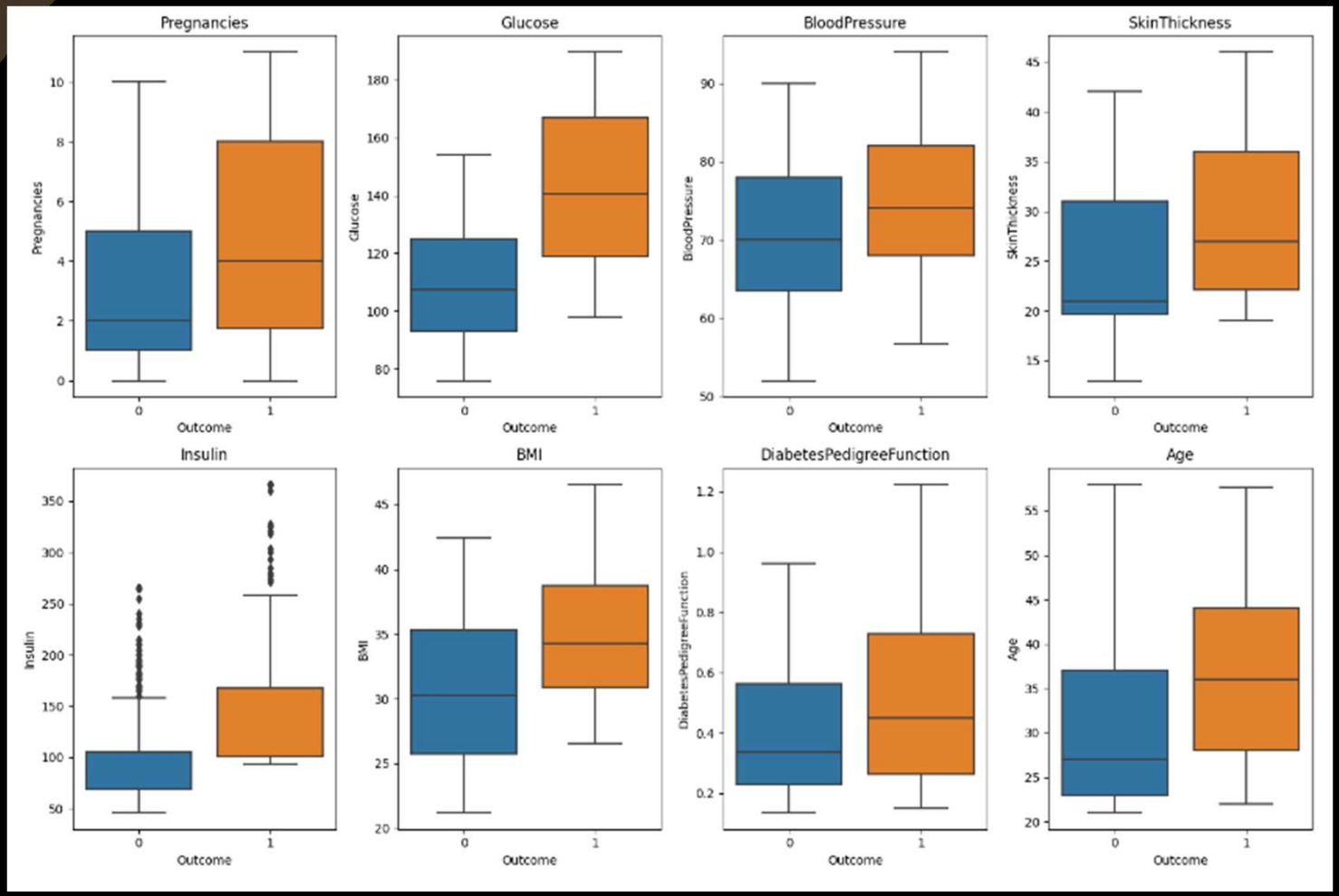
Part II : Preprocessing & Feature Engineering

You need to perform preprocessing on the given dataset. Please consider the following tasks and carry out the necessary steps accordingly.

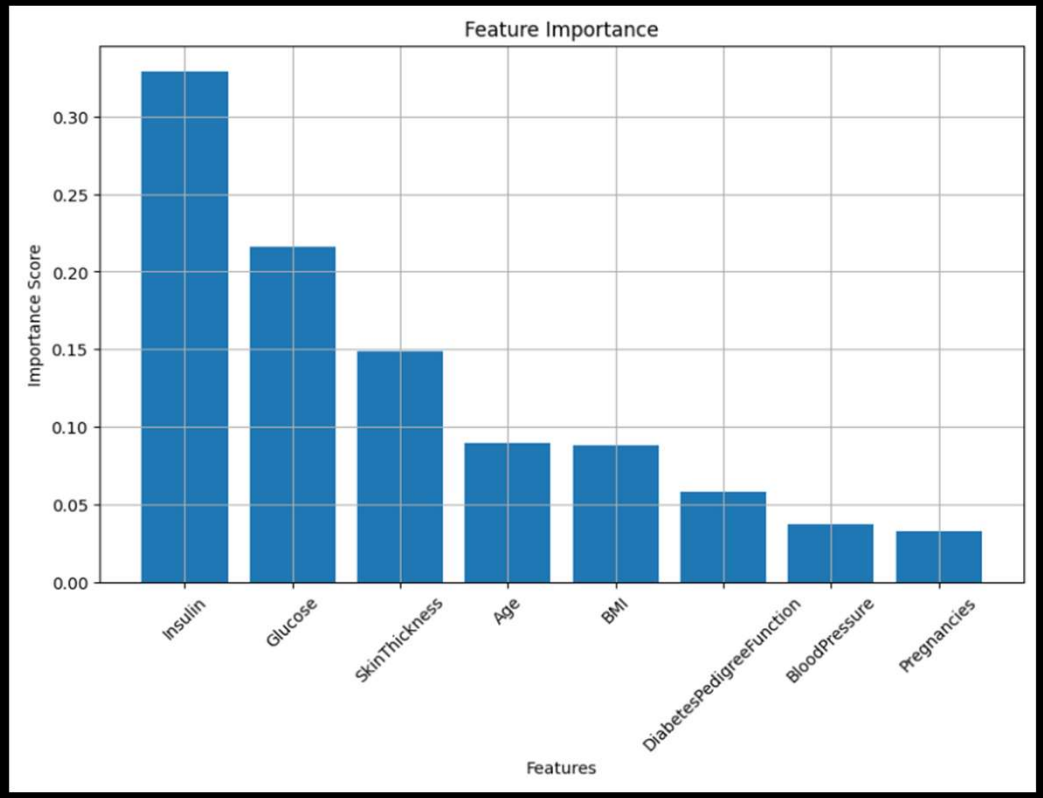
- **Handling missing values:** There are no null values, however, 0 values did not make sense in the results. I converted the 0 values to the mean in each feature split by diabetes vs non-diabetes.
- **Handling outliers:** I used box plots to look for outliers, then used the IQR (Interquartile Range) method to perform outlier imputation.
- **Scaling and normalization:** I did not use scaling and normalization, since I am using a Random Forest Module.
- **Feature Engineering:** Feature Engineering is performed with Random Forest Module.
- **Handling imbalanced data:** I did not find this data set to be imbalanced.

Handling outliers: I used box plots to look for outliers, then used the IQR (Interquartile Range) method to perform outlier imputation.




Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome

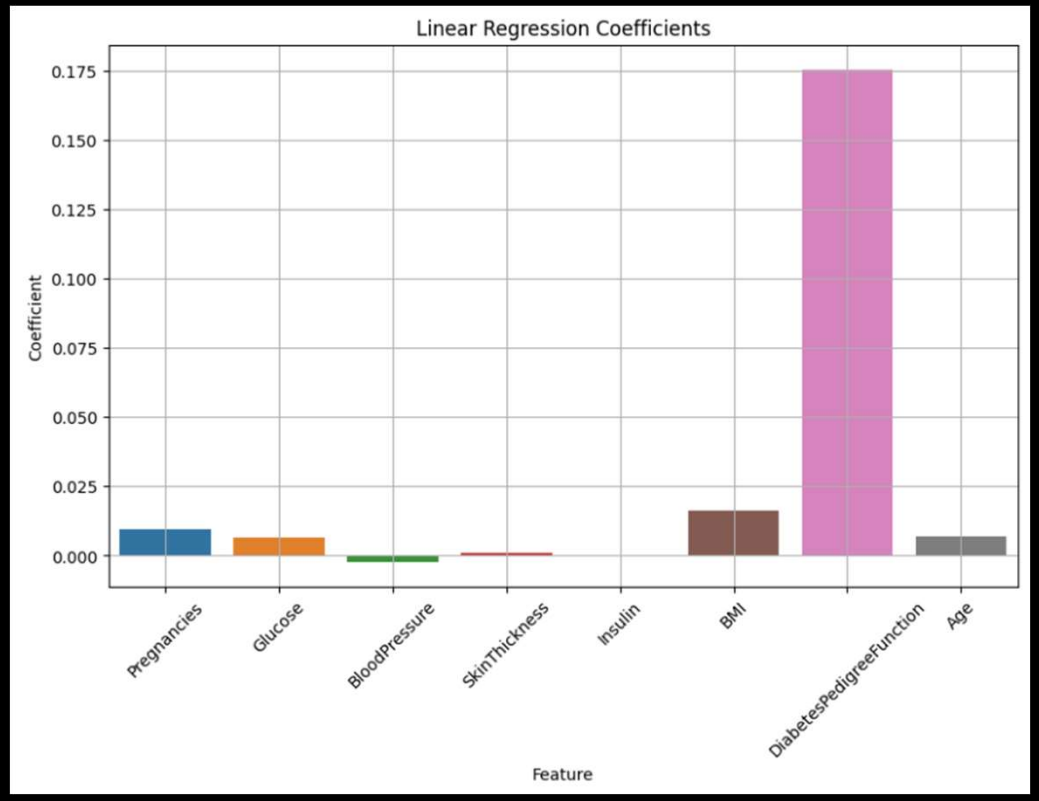
Feature Engineering: Feature Engineering is performed with Random Forest Module.



Part III : Training ML Model

- For this task, you are required to build a machine-learning model to predict the outcome variable. This will be a binary classification task, as the target variable is binary. You should select at least two models, one of which should be an ensemble model, and compare their performance.
- **Train the models:** Train the selected models on the training set.
- **Model evaluation:** Evaluate the trained models on the testing set using appropriate evaluation metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC.
- **Model comparison:** Compare the performance of the selected models and choose the best-performing model based on the evaluation metrics. You can also perform additional analysis, such as model tuning and cross-validation, to improve the model's performance.

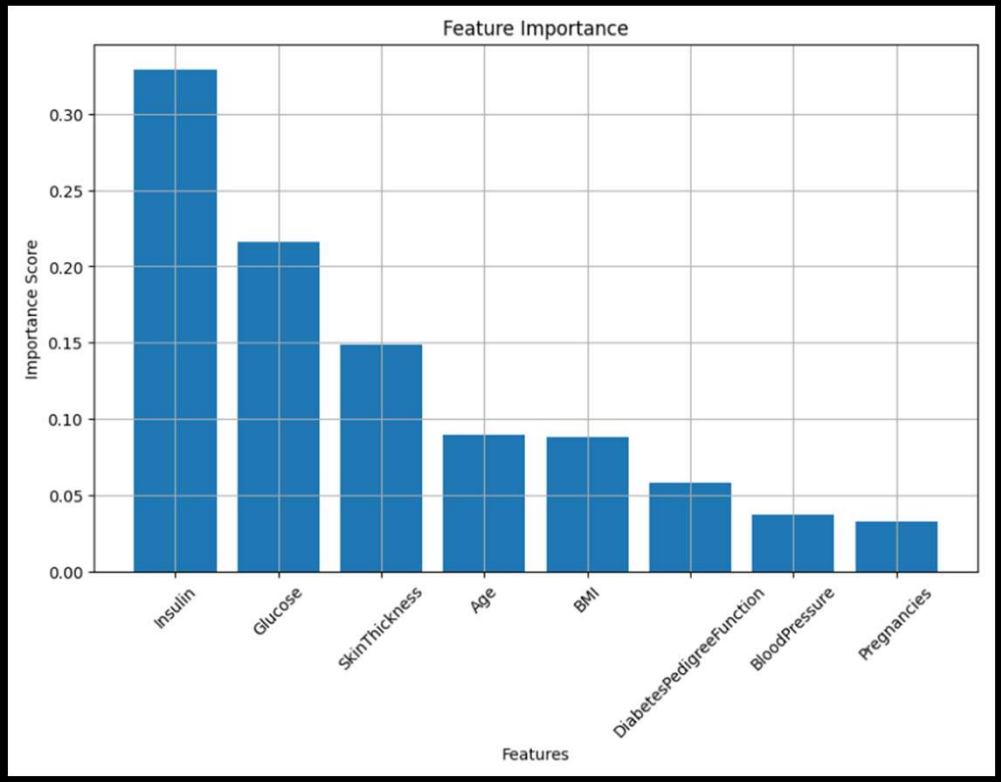
Multiple Linear Regression model with additional functions for data splitting, plotting coefficients, checking for interaction effects, cross-validation, and grid search



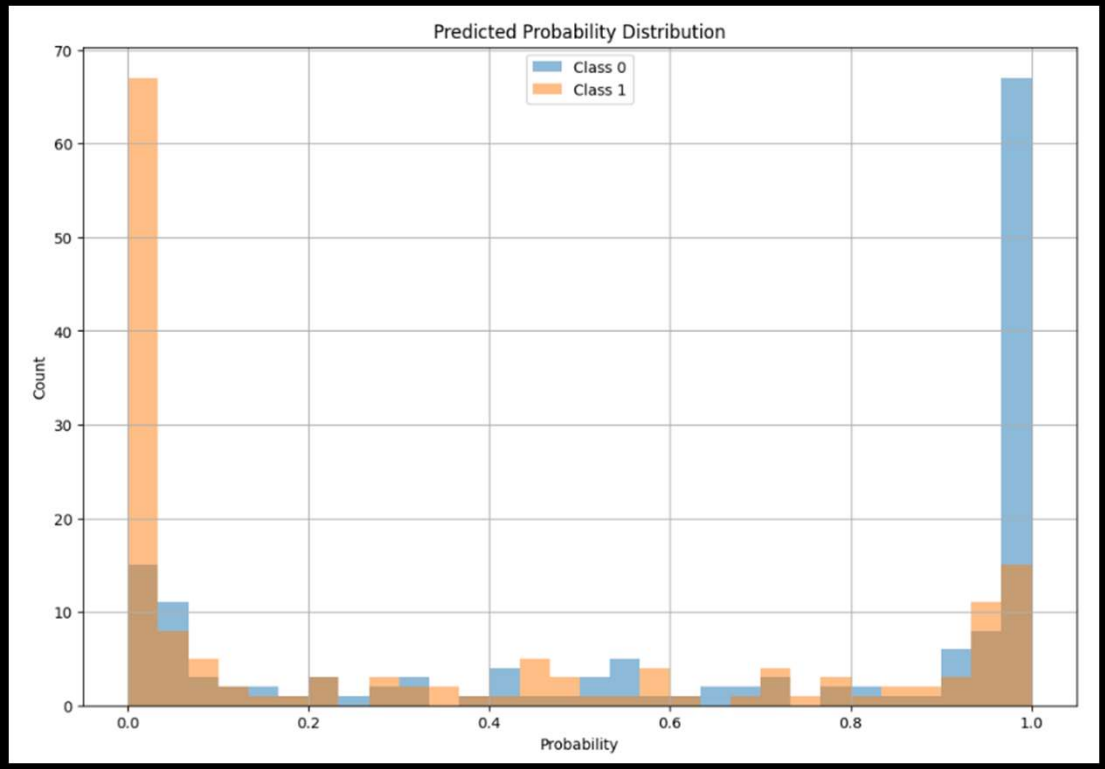
Multiple Linear Regression model with additional functions for data splitting, plotting coefficients, checking for interaction effects, cross-validation, and grid search

```
There are significant interaction effects in the model.
Significant Interaction Terms:
const                          9.951089e-22
Glucose                        6.672273e-22
BMI                            3.605411e-07
DiabetesPedigreeFunction       2.684870e-03
Age                            4.327350e-04
dtype: float64

Cross-validation Mean MSE: 0.14573083776799728
Cross-validation Std MSE: 0.01863136771814718

Best Hyperparameters: {'copy_X': True, 'fit_intercept': True}
Best Model MSE: 0.14593320807907761
```

Random Forest Classifier with Feature Importance, Probability distribution, Precision, Recall, F1-score, ROC-AUC, Classification Report, Confusion Matrix, and ROC curve



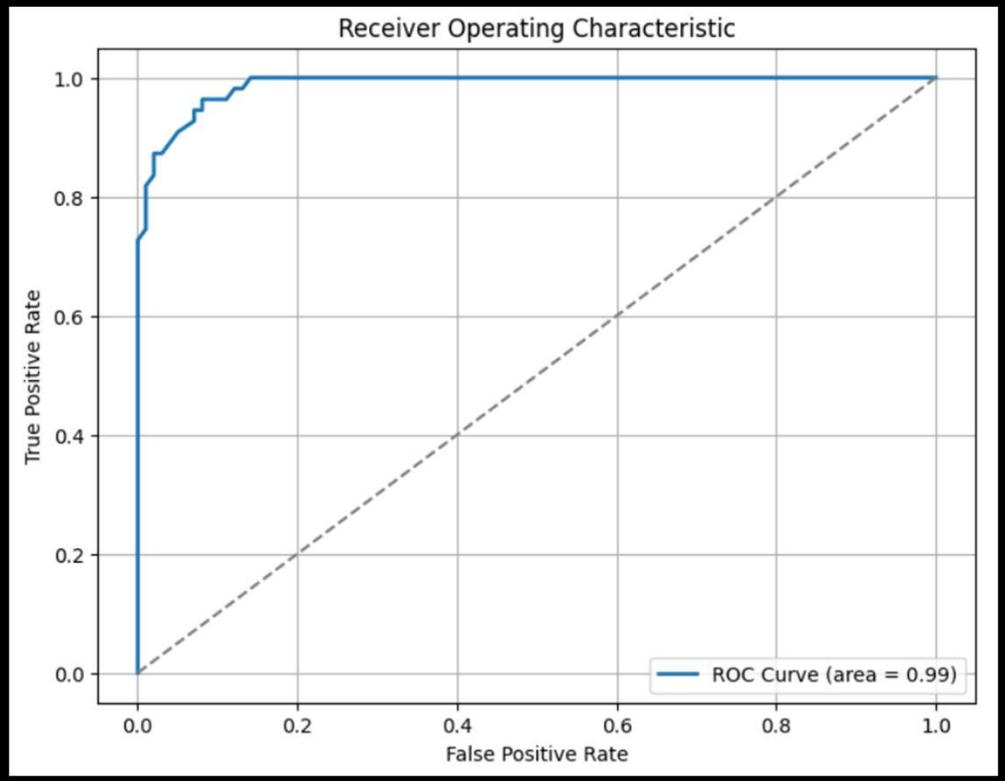
Random Forest Classifier with Feature Importance, Probability distribution, Precision, Recall, F1-score, ROC-AUC, Classification Report, Confusion Matrix, and ROC curve



Random Forest Classifier with Feature Importance, Probability distribution, Precision, Recall, F1-score, ROC-AUC, Classification Report, Confusion Matrix, and ROC curve

```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.98      0.96        99
           1       0.96      0.87      0.91        55

    accuracy                           0.94       154
   macro avg       0.95      0.93      0.93       154
weighted avg       0.94      0.94      0.94       154


Confusion Matrix:
[[97  2]
 [ 7 48]]
```

Random Forest Classifier with Feature Importance, Probability distribution, Precision, Recall, F1-score, ROC-AUC, Classification Report, Confusion Matrix, and ROC curve



Random Forest Classifier with Feature Importance, Probability distribution, Precision, Recall, F1-score, ROC-AUC, Classification Report, Confusion Matrix, and ROC curve

```
Random Forest Classifier:
Accuracy: 94.16%
Precision: 0.960
Recall: 0.873
F1-score: 0.914
ROC-AUC: 0.988
```

Training and Testing Results

```
Training Set Results:
Accuracy: 100.00%

Testing Set Results:
Accuracy: 94.16%

Possible Overfitting:
Overfitting: 5.84%
```

Part IV: Conclusion

- **Feature Importance:** The model identifies the top 5 features that have the most significant impact on predicting the outcome of diabetes. These features, in descending order of importance, are insulin, glucose, skin thickness, age, and BMI. These variables are crucial in predicting whether a person is likely to develop diabetes.
- **Interaction Effects:** The model reveals significant interaction effects between certain predictor variables. Specifically, there are statistically significant interaction terms between the constant term, glucose, BMI, diabetes pedigree function, and age. This indicates that the relationship between these variables is not independent and can influence the outcome differently based on their interactions.
- **Model Performance:** The Random Forest Classifier demonstrates strong performance with an accuracy of 94.16%. It achieves high precision (0.960) and recall (0.873) for the positive class (indicating diabetes). The F1-score, which balances precision and recall, is 0.914. The ROC-AUC score of 0.988 indicates that the model has excellent discrimination ability in distinguishing between the two classes.
- **Overfitting:** There is evidence of potential overfitting, as the model achieves 100% accuracy on the training set but slightly lower accuracy (94.16%) on the testing set. The difference between training and testing set accuracy (5.84%) suggests that the model may perform slightly worse on unseen data. It is essential to consider regularization techniques or further fine-tuning of hyperparameters to reduce overfitting and improve generalization to new data.