Project Title: Diabetes Dataset Analysis

Introduction:

The aim of this project report is to analyze the Diabetes dataset, which contains 768 rows and 9 columns (features). The dataset focuses on predicting the presence or absence of diabetes in individuals based on various health-related features. The target variable, "Outcome," indicates 1 for persons with diabetes and 0 for non-diabetic patients.

Dataset Overview:

Number of Rows: 768

Number of Columns (Features): 9

Feature Columns: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age

Target Variable: Outcome

```
df = pd.read_csv('diabetes.csv')
df.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

```
df.tail()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

```
df.shape
```

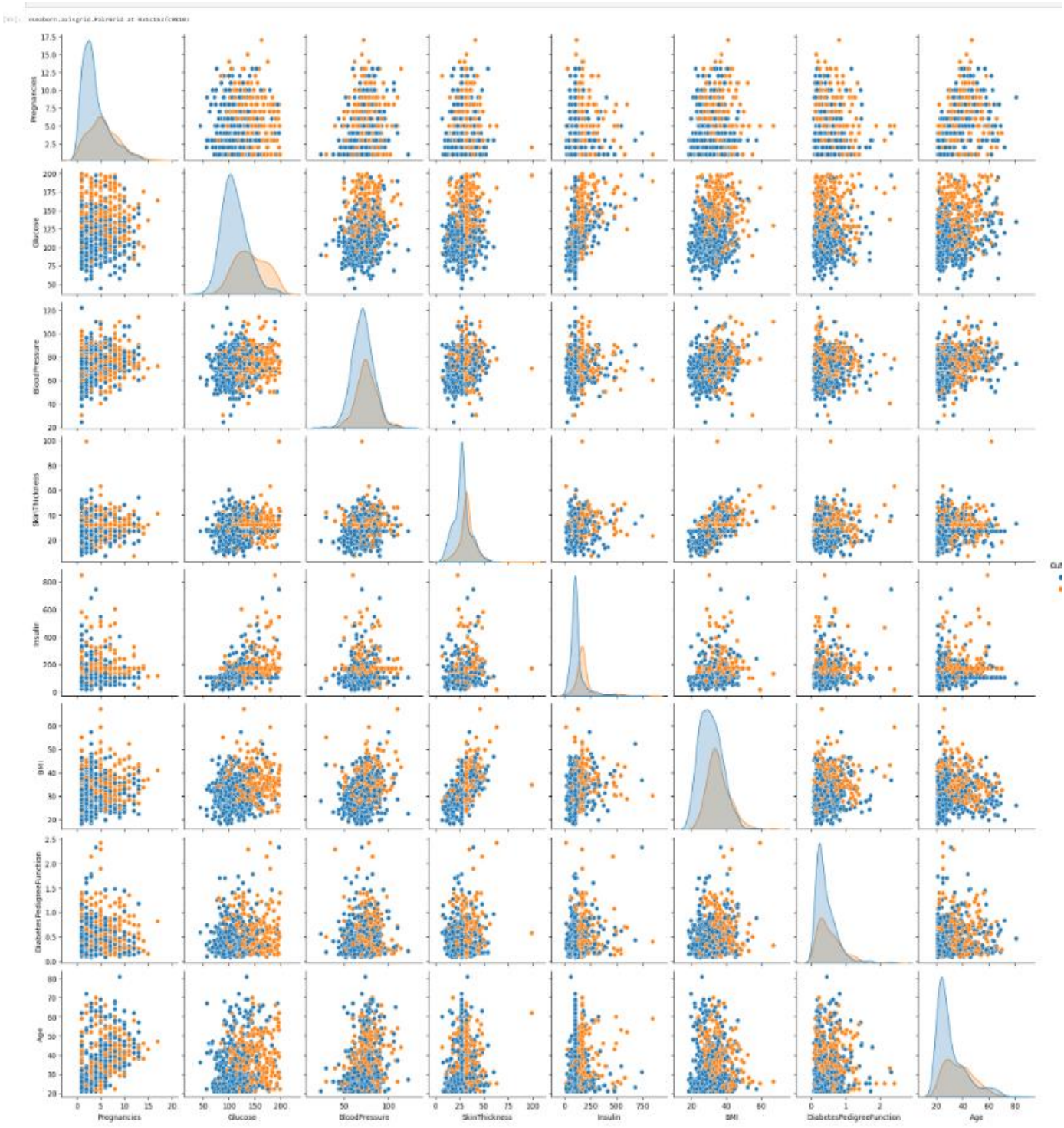
```
(768, 9)
```

Data Quality:

The dataset contains no null values, indicating that it is complete in terms of data availability.

	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	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	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

How the predictor variables relates to each other?



Key Insights:

Average BMI for individuals with diabetes: 35.40

Average BMI for individuals without diabetes: 30.85

Average glucose level for individuals with diabetes: 142.30

Average glucose level for individuals without diabetes: 110.62

Model and Performance:

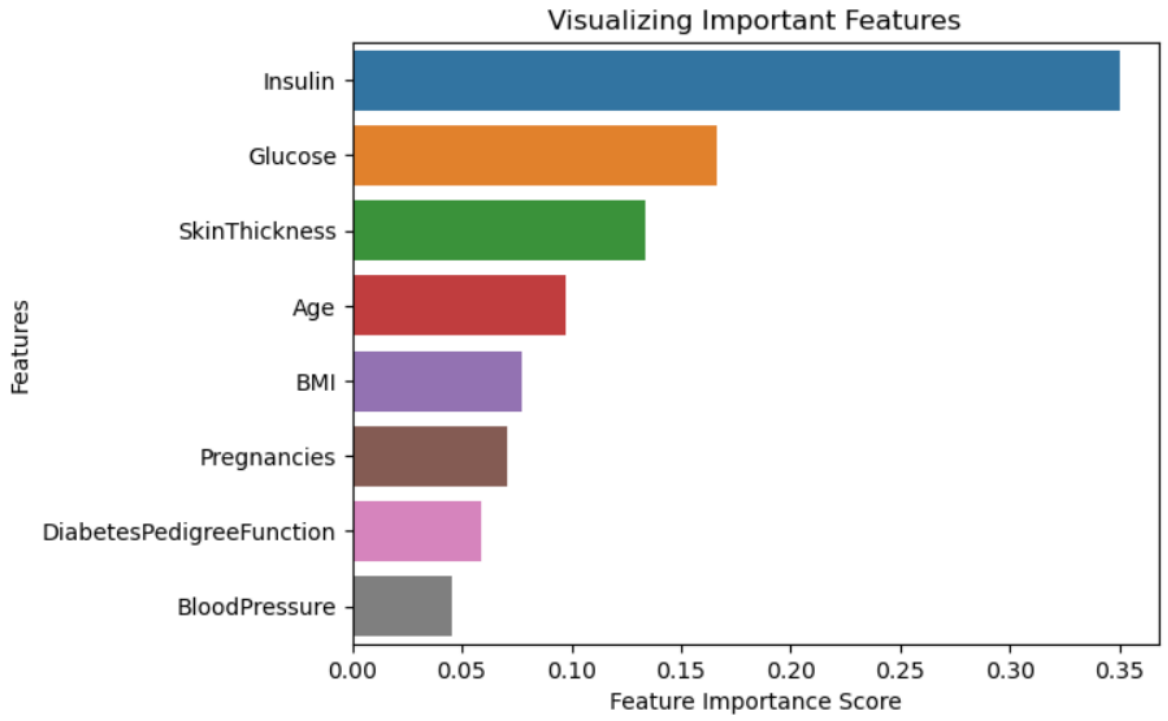
Visualize feature scores of the features

```
# view the feature scores

feature_scores = pd.Series(clf.feature_importances_, index=X_train.columns).sort_values(ascending=False)

feature_scores
```

```
Insulin                   0.350163
Glucose                   0.166065
SkinThickness             0.133898
Age                       0.097578
BMI                       0.077492
Pregnancies               0.070701
DiabetesPedigreeFunction  0.058932
BloodPressure             0.045171
dtype: float64
```



What is the average age of the individuals in the dataset? ¶

[86]:
```
avg_age = df_processed['Age'].mean()

print("Average Age:", avg_age)
```

```
Average Age: 33.240885416666664
```

What is the average BMI for individuals with diabetes and without diabetes?

[87]:
```
# Calculate the average glucose level for individuals with diabetes
avg_BMI_diabetes = df_processed[df_processed['Outcome'] == 1]['BMI'].mean()

# Calculate the average glucose level for individuals without diabetes
avg_BMI_no_diabetes = df_processed[df_processed['Outcome'] == 0]['BMI'].mean()

print("Average BMI for individuals with diabetes:", avg_BMI_diabetes)
print("Average BMI for individuals without diabetes:", avg_BMI_no_diabetes)
```

```
Average BMI for individuals with diabetes: 35.39850746268657
Average BMI for individuals without diabetes: 30.846
```

What is the average glucose level for individuals with diabetes and without diabetes?

[88]:
```
# Calculate the average glucose level for individuals with diabetes
avg_glucose_diabetes = df_processed[df_processed['Outcome'] == 1]['Glucose'].mean()

# Calculate the average glucose level for individuals without diabetes
avg_glucose_no_diabetes = df_processed[df_processed['Outcome'] == 0]['Glucose'].mean()

print("Average glucose level for individuals with diabetes:", avg_glucose_diabetes)
print("Average glucose level for individuals without diabetes:", avg_glucose_no_diabetes)
```

```
Average glucose level for individuals with diabetes: 142.30223880597015
Average glucose level for individuals without diabetes: 110.622
```

Model Used: Random Forest Classifier

Accuracy: 87%

Random Forest Classifier model with parameter n_estimators=100

```
# instantiate the classifier with n_estimators = 100

rfc_100 = RandomForestClassifier(n_estimators=100, random_state=0)


# fit the model to the training set

rfc_100.fit(X_train, y_train)


# Predict on the test set results

y_pred_100 = rfc_100.predict(X_test)


# Check accuracy score

print('Model accuracy score with 100 decision-trees : {0:0.4f}'. format(accuracy_score(y_test, y_pred_100)))
```

```
Model accuracy score with 100 decision-trees : 0.8701
```

Conclusion:

In this project, we analyzed the Diabetes dataset and gained insights into key features related to diabetes. The average BMI and glucose levels were higher in individuals with diabetes compared to those without diabetes. The Random Forest Classifier model was employed to predict diabetes presence, achieving an accuracy rate of 87%.

These findings highlight the importance of BMI and glucose levels as significant factors in diagnosing and understanding diabetes. The results obtained can be utilized for further research, medical interventions, and personalized patient care in diabetes management.

However, it is essential to note that further analysis and exploration can be performed to gain a more comprehensive understanding of the dataset and refine the predictive models.