Blood Sugar Blueprint: Decision Tree Diagnostics

1. Dataset Overview:

- The data set contains information on factors influencing the diabetes amongst patients based on pregnancies, glucose, blood pressure, skin Thickness, Insulin, BMI, Age etc.

2. Data Preprocessing:

- Checking if there are any missing values.

```
   Pregnancies  Glucose  BloodPressure  ...  DiabetesPedigreeFunction  Age  Outcome
0            6      148             72  ...                     0.627   50        1
1            1       85             66  ...                     0.351   31        0
2            8      183             64  ...                     0.672   32        1
3            1       89             66  ...                     0.167   21        0
4            0      137             40  ...                     2.288   33        1

[5 rows x 9 columns]
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

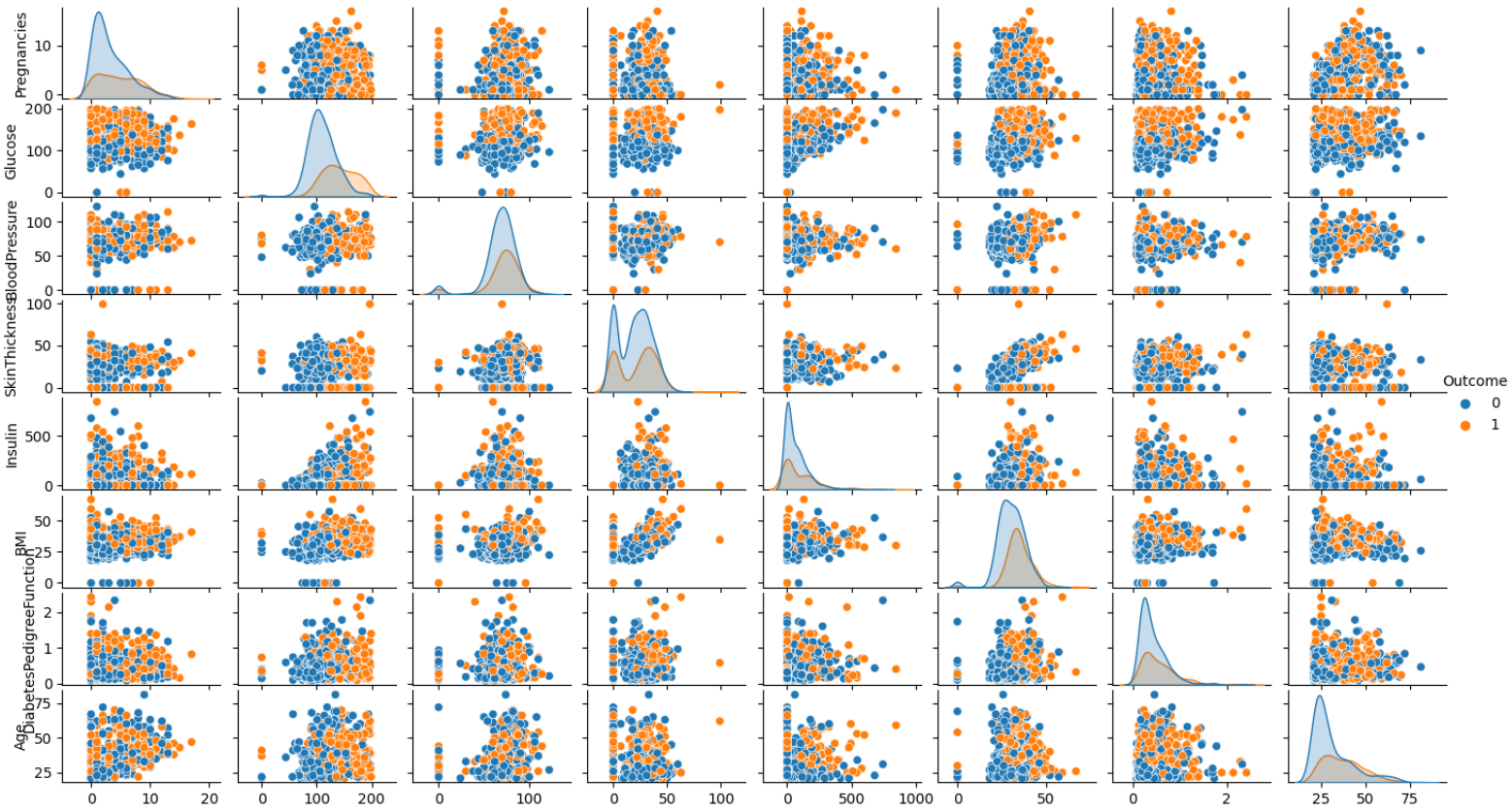
- Data statistics:

- There are 768 count of total data provided in the data set along with data outputs.

	Pregnancies	Glucose	...	Age	Outcome
count	768.000000	768.000000	...	768.000000	768.000000
mean	3.845052	120.894531	...	33.240885	0.348958
std	3.369578	31.972618	...	11.760232	0.476951
min	0.000000	0.000000	...	21.000000	0.000000
25%	1.000000	99.000000	...	24.000000	0.000000
50%	3.000000	117.000000	...	29.000000	0.000000
75%	6.000000	140.250000	...	41.000000	1.000000
max	17.000000	199.000000	...	81.000000	1.000000

- Data visualization:

- A pair plot is created using Seaborn to visualize relationship between different pairs of features. The plot is colored by the outcome variable, which indicates whether a patient has diabetes or not.



3. Model Building & training:

- **Data Splitting:**
  The data set is split into 2 sets training and testing sets using the train_test_split from sklearn. 80% of data isused for training & 20% is used for testing.
- **Decision Tree:**
  Using decision tree classifier from sklearn. It initialized a random state for reproductibility. It used CART algorithm to construct decision tree for classification tasks.
- The classifier is trained on training portion that is determined from data set i.e 80%.

4. Model Evaluation:

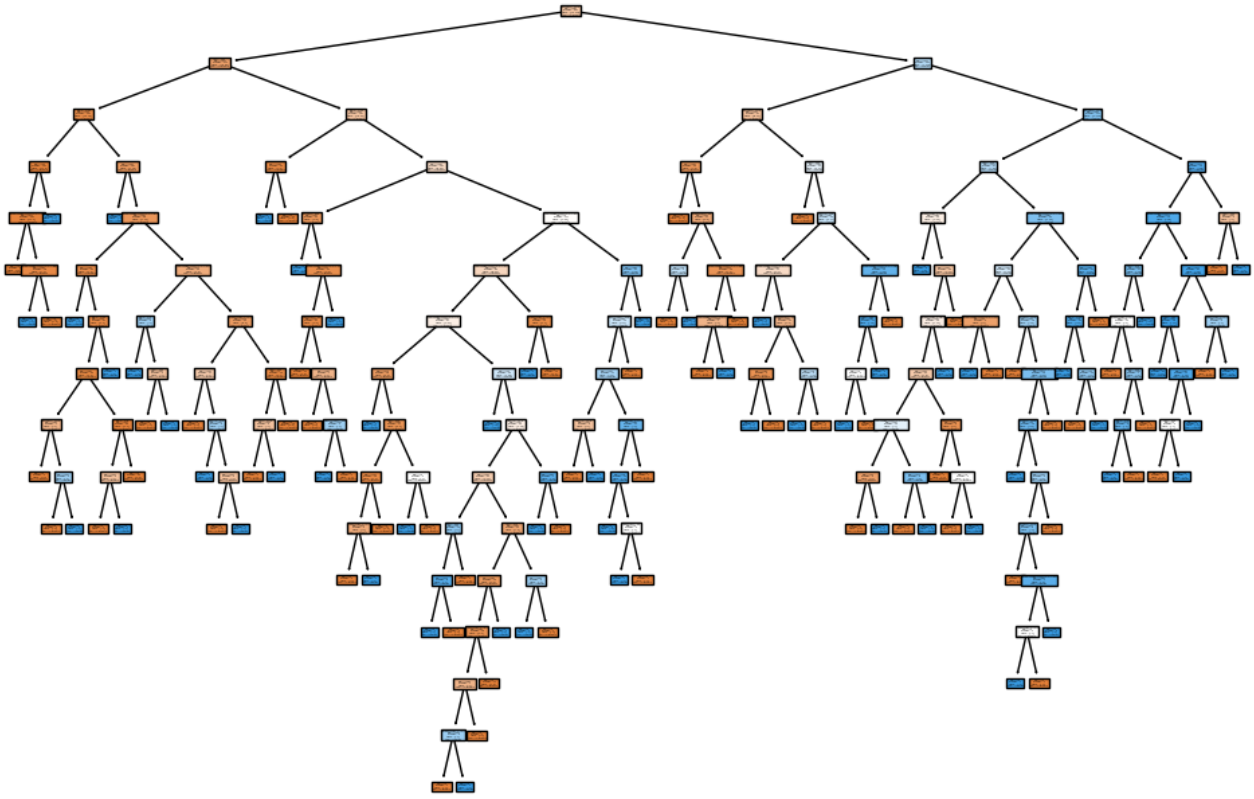
- The model performance is evaluated with the following metrics:
  Accuracy: The accuracy of the model on the test set is calculated using accuracy_score. It represents the percentage of correct predictions.
  Confusion matrix: A confusion matrix is created to show the number of true positives, true negatives, false positives, and false negatives.
  Classification reports: The classification report provides precision, recall, F1-score, and support for each class.

```
[8 rows x 9 columns]
Accuracy: 0.7758620689655172
Confusion Matrix:
 [[61 15]
 [11 29]]
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.80      0.82        76
           1       0.66      0.72      0.69        40

    accuracy                           0.78       116
   macro avg       0.75      0.76      0.76       116
weighted avg       0.78      0.78      0.78       116
```

5. Decision Tree Visualization:



6. Conclusion:

In conclusion, the Decision Tree Classifier has been built, trained, and evaluated for the diabetes dataset. The model achieved a certain level of accuracy in predicting diabetes cases, as shown in the evaluation metrics. The decision tree visualization provides a clear picture of how the model makes decisions based on the features. Further analysis and optimization can be performed to fine-tune the model for better predictive performance, but this code serves as a comprehensive starting point for working with the diabetes dataset.