SYS-626

HOMEWORK 03

AJAY RANA

About Dataset

Diabetes is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased hunger. If left untreated, diabetes can cause many complications.

The dataset consists of several medical predictor variables and one target variable, Outcome. Predictor variables include the no. of pregnancies the patient has had, their BMI, Insulin level, age, Diabetes Pedigree Function, BP, Glucose and Skin Thickness.

There are a total of 8 Variables and 768 Non-Null data that present in each variable and there is no missing data.

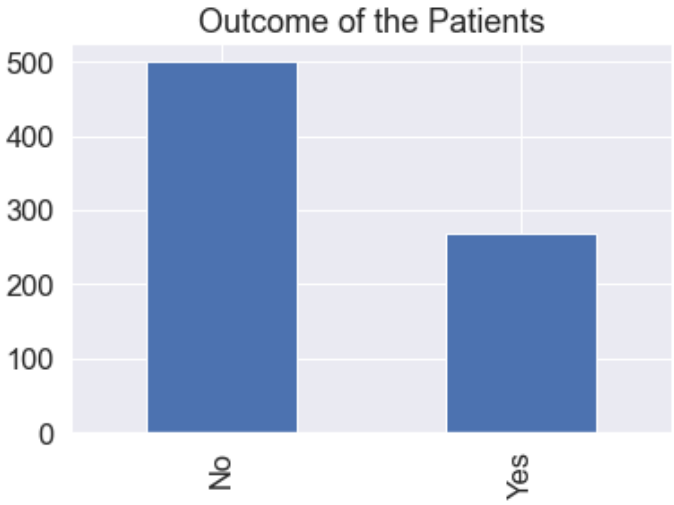
```
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #    Column                    Non-Null Count  Dtype
---   ------                    --------------  -----
 0    Pregnancies               768 non-null    int64
 1    Glucose                   768 non-null    int64
 2    BloodPressure             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
```

Data Exploration

Target Variable

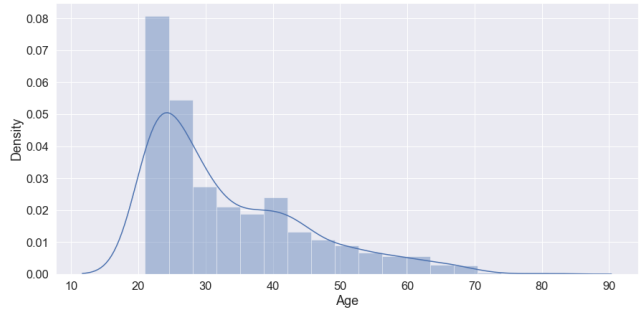
The target variable of the dataset is the Outcome. Where 0=No Diabetes and 1=diabetes.

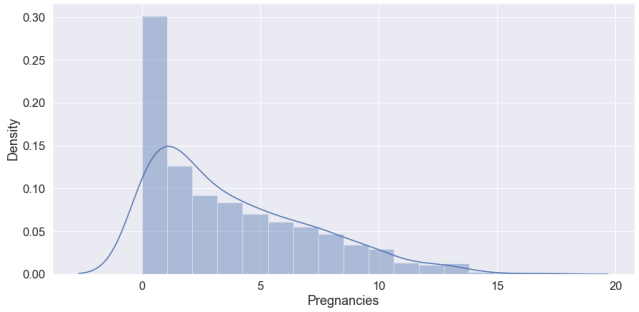
```
Outcome of Patients
 Outcome
0    500
1    268
```

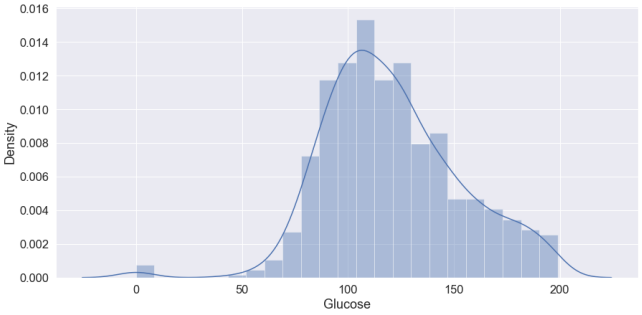


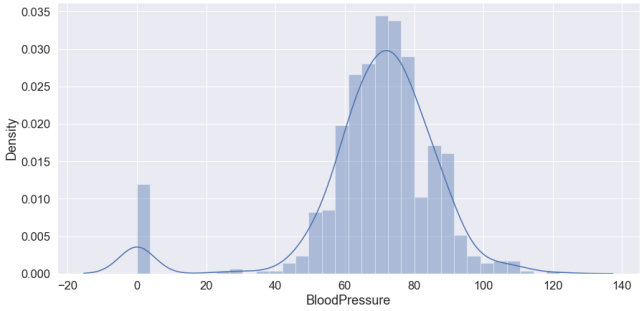
Bar & Histogram Plot

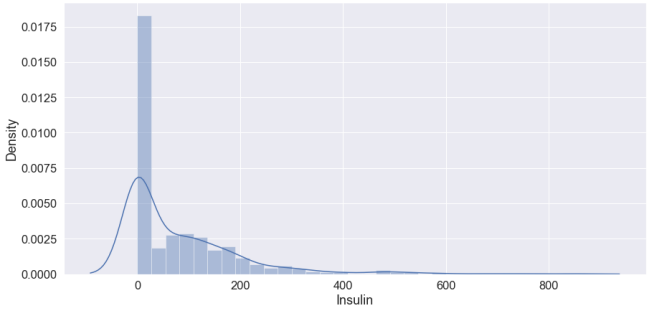
Below is the bar graph with a histogram which shows how each variable data is distributed.

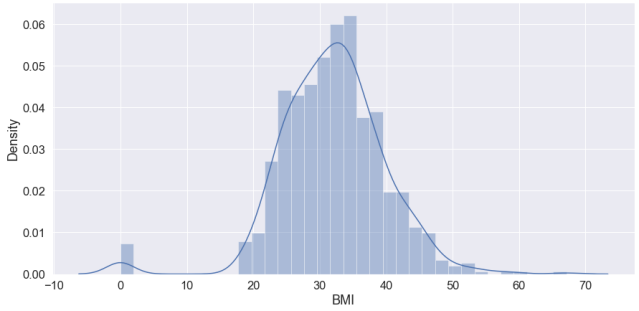


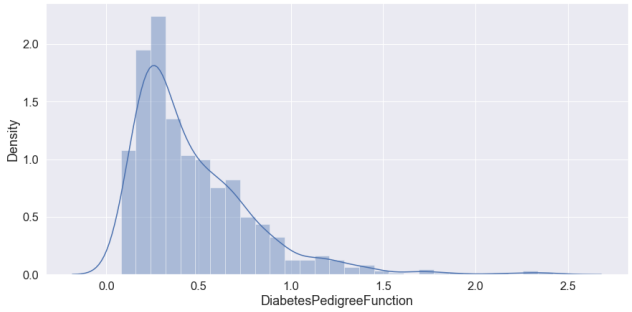


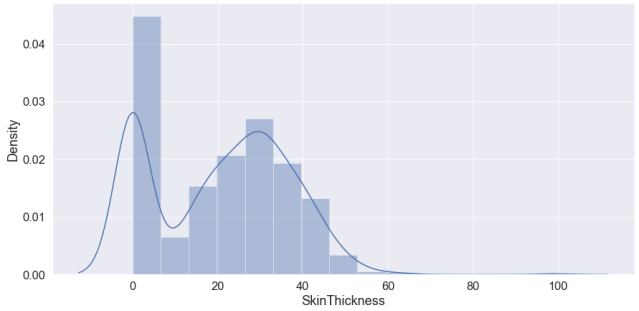












Describe data variables

```
Age of Patients
 Age
22    72
21    63
25    48
24    46
23    38
Name: Age, dtype: int64
No. of Pregnancies
 Pregnancies
1     135
0     111
2     103
3      75
4      68
Name: Pregnancies, dtype: int64
BP of Patients
 BloodPressure
70    57
74    52
68    45
78    45
72    44
Name: BloodPressure, dtype: int64
```

- Most of the patients' age lies between 21-25
- Most No. of pregnancies lies in the range 0-2
- Avg Blood Pressure range is from 70-78 (in mm Hg).

```
Insulin level of Patients
 Insulin
0       374
105      11
130       9
140       9
120       8
Name: Insulin, dtype: int64
BMI of Patients
 BMI
32.0    13
31.2    12
31.6    12
0.0     11
32.4    10
Name: BMI, dtype: int64
Glucose level of Patients
 Glucose
99      17
100     17
106     14
111     14
125     14
```

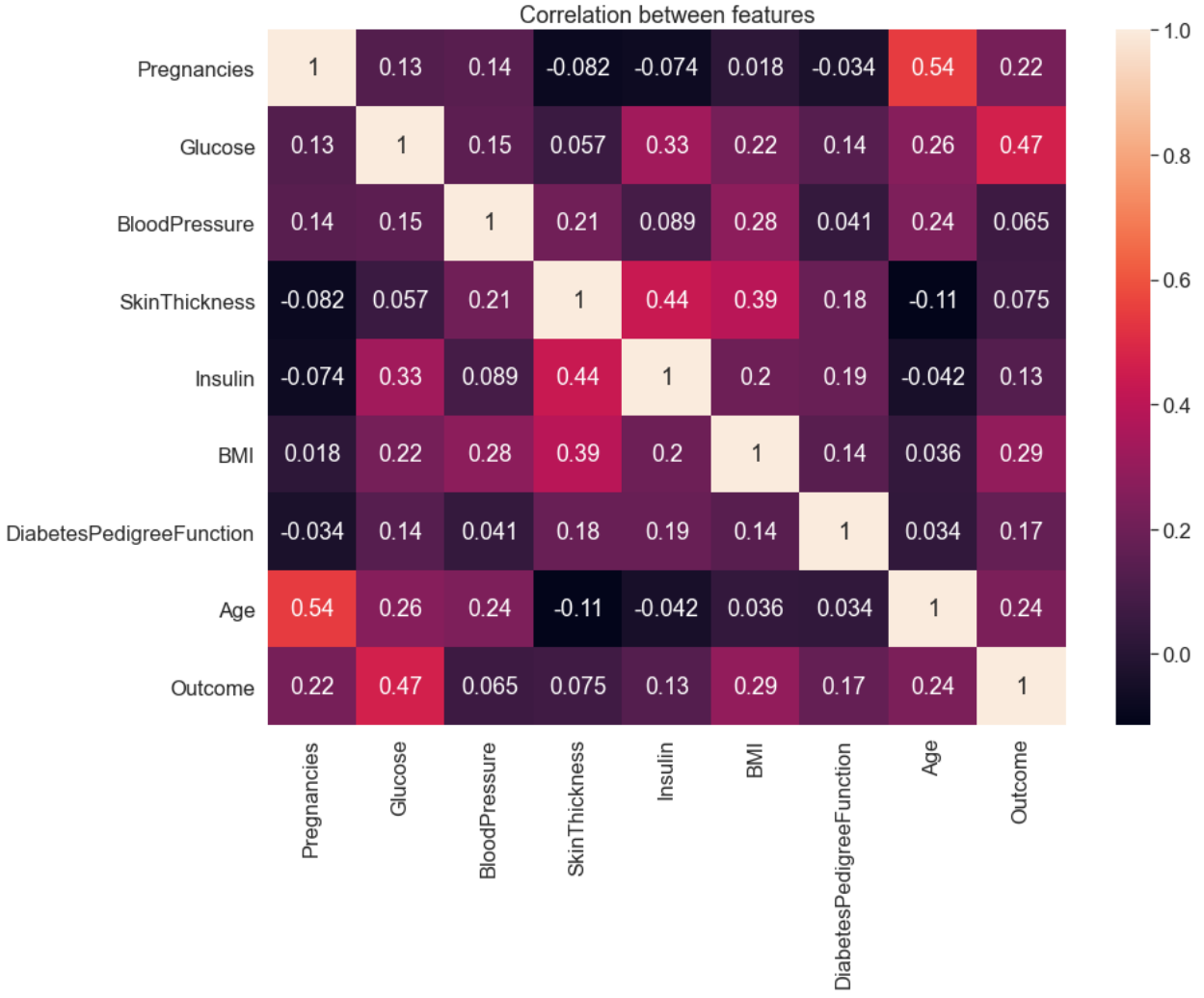
- Most Patients have Insulin Levels 0 (mu U/ml)
- Average BMI of Most Patients is 31-32(kg/m^2)
- Glucose level range of Patients is 99-125

```
Skin Thickness Patients
 SkinThickness
0      227
32      31
30      27
27      23
23      22
Name: SkinThickness, dtype: int64
Diabetes Pedigree Function of Patients
 DiabetesPedigreeFunction
0.254    6
0.258    6
0.207    5
0.238    5
0.259    5
```

- Skin Thickness of most of the patients is 0 (mm)

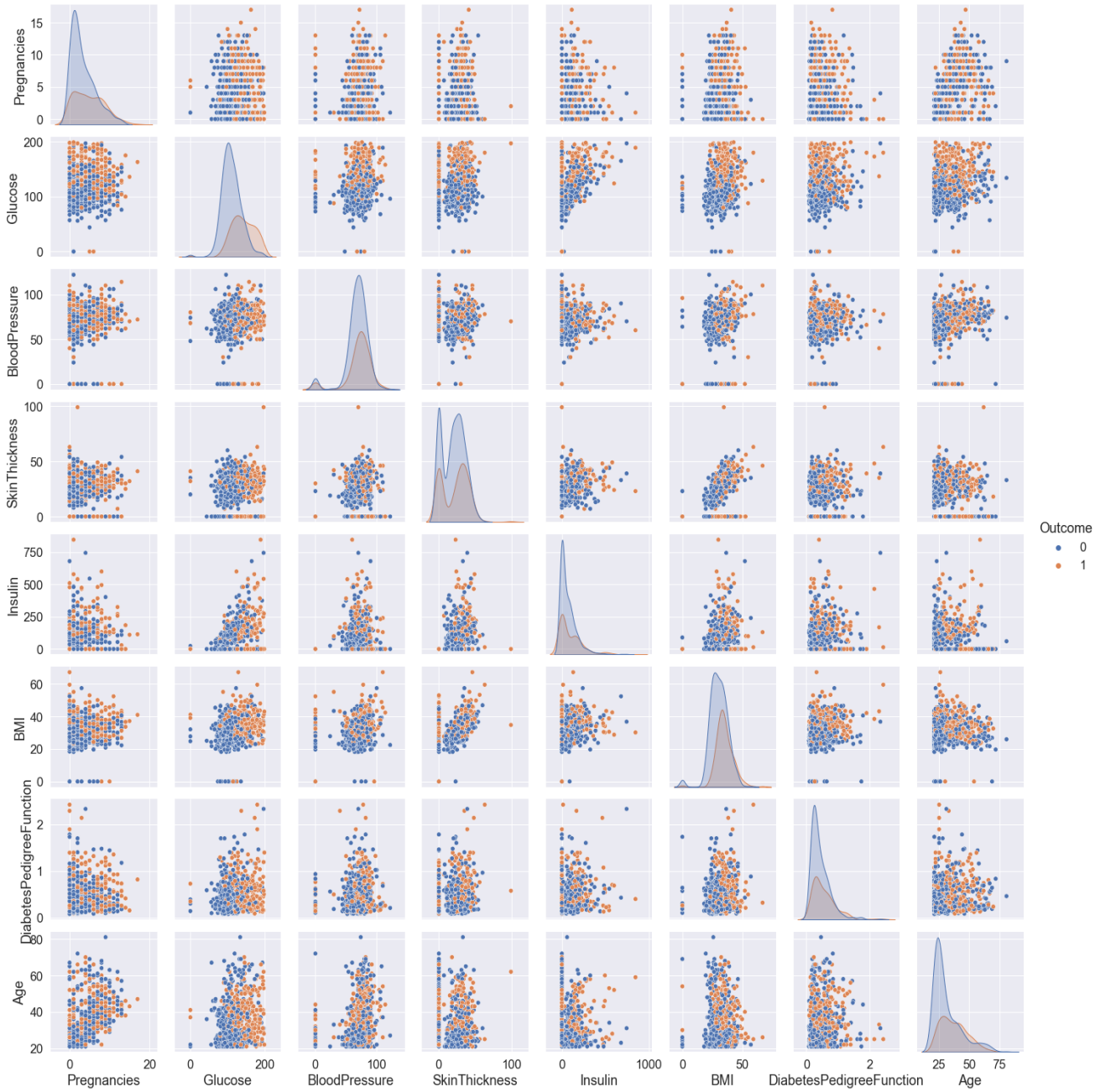
Correlation Matrix

Correlation Matrix tells us how two variables are related to each other. By looking at the correlation matrix we can see that Outcome that person will have diabetes or not is mostly correlated to Glucose, BMI, Age, No. of Pregnancies and lightly correlated to Diabetes Pedigree Function and Insulin, Other variables have little impact on it. But some variables are indirectly related through other variables for e.g BMI is highly positively correlated with Skin Thickness whereas Skin thickness is very lightly correlated to Outcome.



Pair Plot

Below is the Pair plot of all the variables. Pair plot helps to explain a relationship between two variables or to form the most separated clusters. Below we have used the target variable Outcome for colour encoding. Scatter Plot shows us the correlation between the two variables. The closer the data points come when plotted to make a straight line, the higher the correlation between the two variables, or the stronger the correlation.

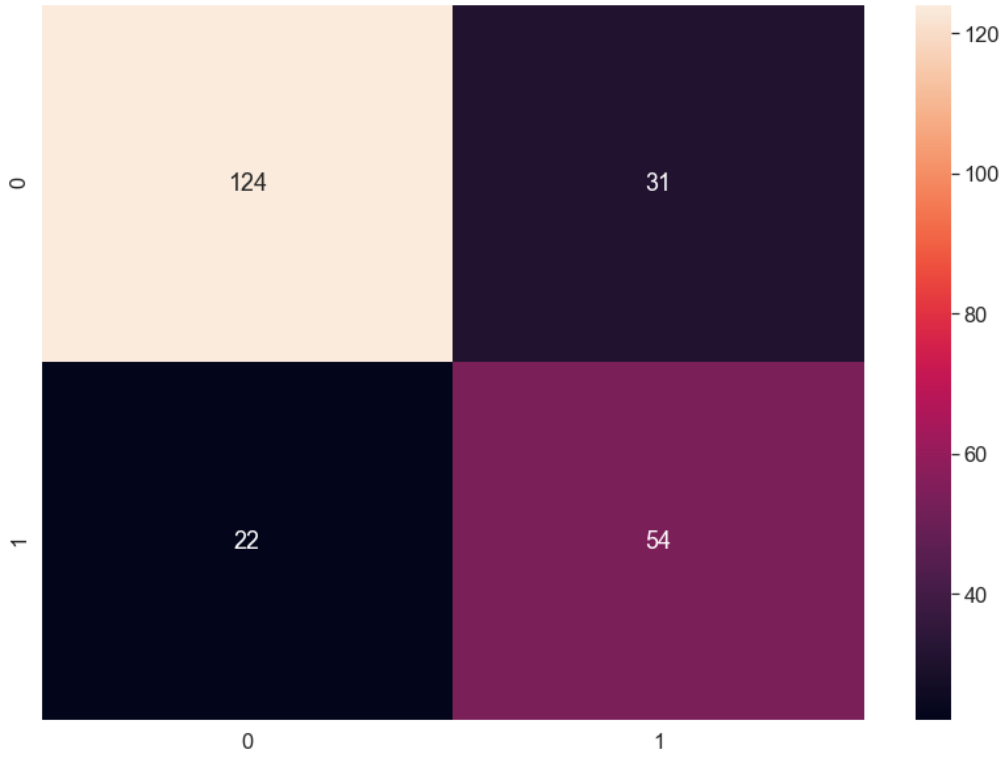


Data Analysis

Confusion/Error Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model or classifier on a set of test data for which the values are known.

n=231	Predicted: No	Predicted: Yes
Actual: No	124	31
Actual: Yes	22	54

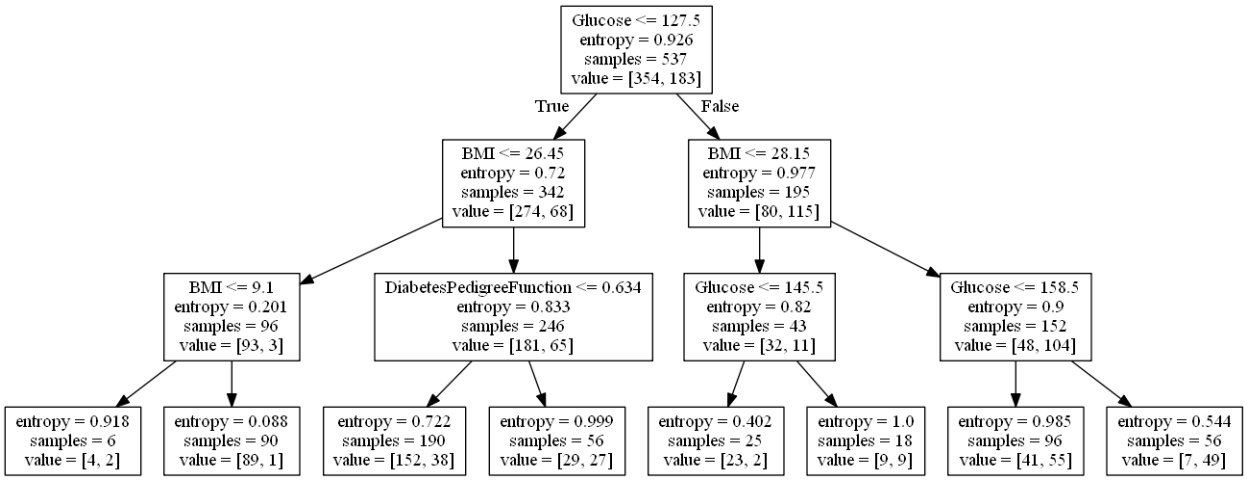


- The Classifier made a total of 231 predictions(e.g These no. of patients were tested for diabetes)
- Out of 231 cases, the classifier predicted “Yes” 85 times, and “No” 146 times.
- In reality, 76 patients in the sample have diabetes, and 155 patients do not.
- Accuracy=(124+54)/231=77.05% which is same as calculated using python

```
Accuracy: 0.7705627705627706
```

Decision Tree

We have plotted the Decision tree for dataset diabetes with a depth of 3. The decision tree depends on the variables Glucose, BMI and Diabetes Pedigree Function. At the top, we have Glucose as we know that it has the highest correlation value of 0.47 with the target variable Outcome. Then it is followed by BMI with the 2nd highest Correlation value of 0.29 and then Diabetes Pedigree Function.



Conclusion

After all the data analysis and by looking at the decision tree we can conclude that the Outcome that patients have diabetes or Not mostly depends on major factor such as Glucose level followed by BMI, Diabetes Pedigree Function, age and No. of pregnancies.