Author:- *Vinay Rajesh Gor*

CWID:- *10446180*

Subject:- *EM-626 Applied AI/ML for Systems & Enterprises*

Instructor:- *Dr. Carlo Lipizzi*

Exploration of Data frame

Data Frame is observed to see the overall quantitative and qualitative analysis of data. This is done using the ***print(df)*** function.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
0	6	148	72	35	0	33.6
1	1	85	66	29	0	26.6
2	8	183	64	0	0	23.3
3	1	89	66	23	94	28.1
4	0	137	40	35	168	43.1
..	...	...	...	...	...	...
763	10	101	76	48	180	32.9
764	2	122	70	27	0	36.8
765	5	121	72	23	112	26.2
766	1	126	60	0	0	30.1
767	1	93	70	31	0	30.4

	DiabetesPedigreeFunction	Age	Outcome
0	0.627	50	1
1	0.351	31	0
2	0.672	32	1
3	0.167	21	0
4	2.288	33	1
..	...	...	...
763	0.171	63	0
764	0.340	27	0
765	0.245	30	0
766	0.349	47	1
767	0.315	23	0

[768 rows x 9 columns]

Fig. Quantitative and qualitative analysis of data

From initial exploratory analysis it can be noticed that all the values in the dataframe are quantitative ranging from integer to float. The outcome is a binary value which indicates if the person is a diabetic or not. Hence Outcome can be considered as Target variable (Dependent) whereas all other variables are considered as Independent variables.

Information Overview

The diabetes.csv file is imported in Jupyter Notebook and using the library pandas. For overview ***df.info()*** function is used and it gives the general information about the number of rows x columns, names of the columns, number of rows for that column, null values for each column and data type is given as output.

```
<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
```

Fig. Information Overview

It can be observed that there are *768 rows x 9 columns*. Some of the 9 attributes are *Pregnancies, Glucose, BloodPressure, BMI* etc. each having their number of counts. For every column there is a data type associated with it. For example BMI and DiabetesPedigreeFunction are float whereas all others are integer. There are no missing values or blank values.

Statistics

Statistics gives us an idea about many mathematical aspects of the data for each attribute such as *count, mean, standard deviation, minimum, maximum, and 1st, 2nd and 3rd quartiles* respectively. This is viewed using function ***df.describe()***

	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

Fig. Statistics Overview

It can be explored from the standard deviation that *Glucose, BloodPressure, Insulin, BMI and Age* are the attributes that will most likely have numeric outliers. This also indicates that our data contains all ages of people with all bio characteristics. Also mean, minimum and maximum gives us the idea of what are the limits of our data. For example the minimum age of the person is 21 years whereas the maximum is 81 years.

Correlation

Correlation is the study of numeric relation between variables. This forms an important aspect to determine how one factor is related to the other. Using the *library matplotlib.pyplot* and *seaborn correlation* can be analyzed using ***df.corr()*** and ***sb.heatmap()*** function.

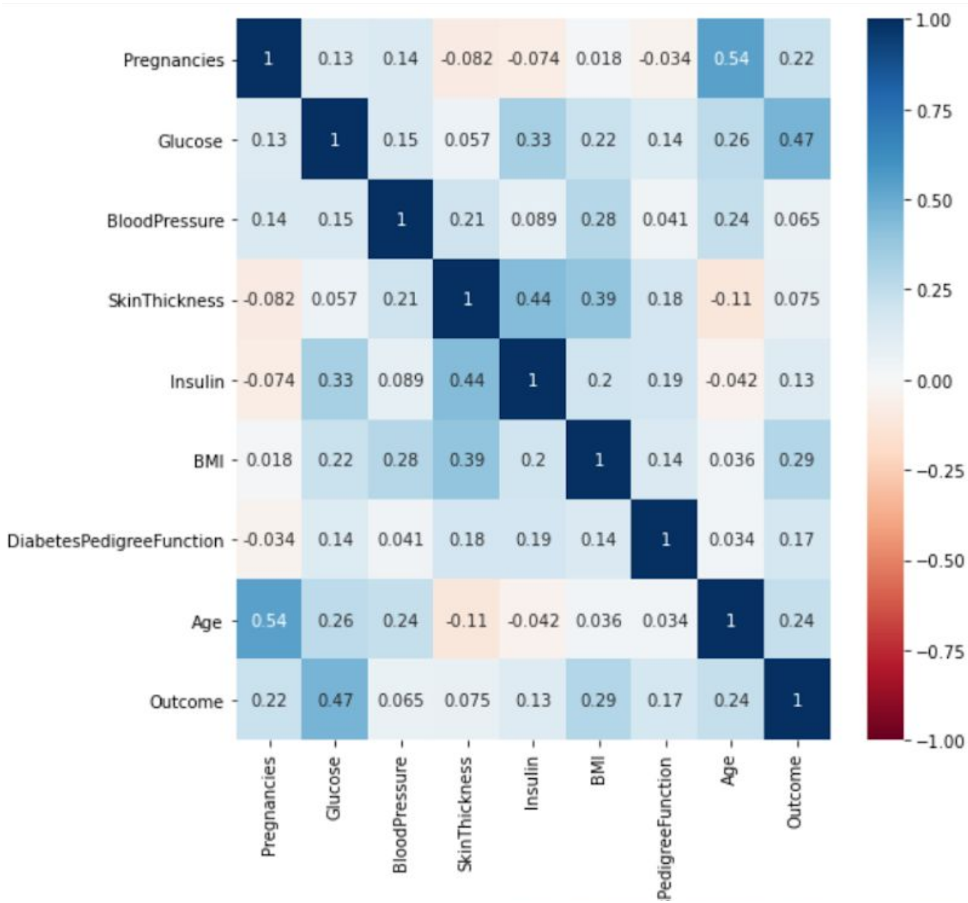


Fig. Correlation Matrix

It can be clearly visualized that the matrix consists of two colours red and blue. Red indicates the variables are negatively correlated to each other whereas Blue indicates the variables are positively correlated to each other.

The Correlation

Insulin, Glucose, SkinThickness, Age are moderately positively related to each other whereas SkinThickness, Insulin and pregnancies are very weakly negatively related to each other

Decision Tree

Complex Structure of Decision Tree

Decision Tree was created using libraries *sklearn* and *graphviz.* Initially a complex structure was obtained.

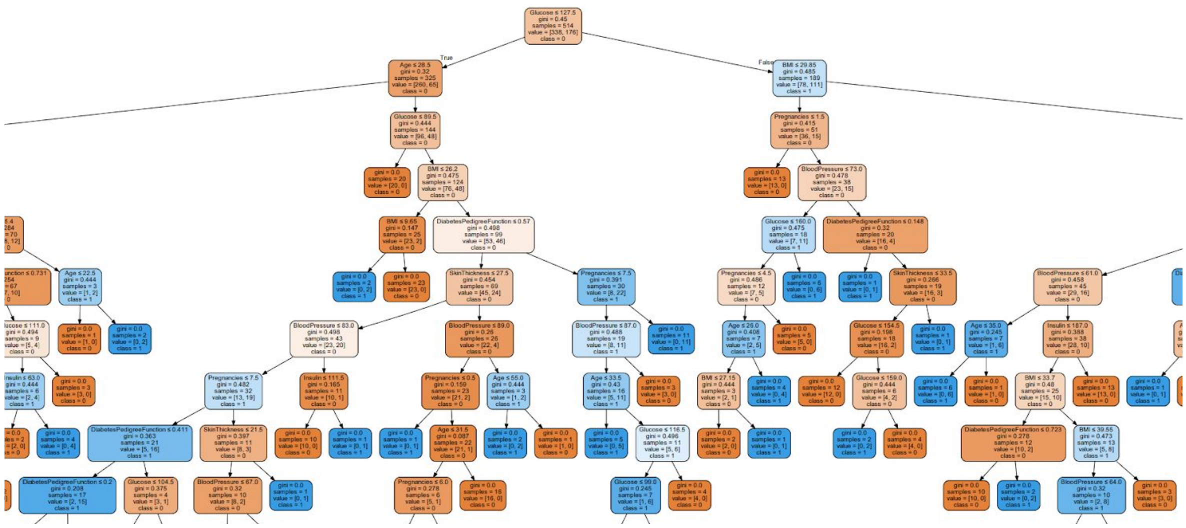


Fig. Complex structure of Decision tree

Simplified Decision Tree with max_depth = 5

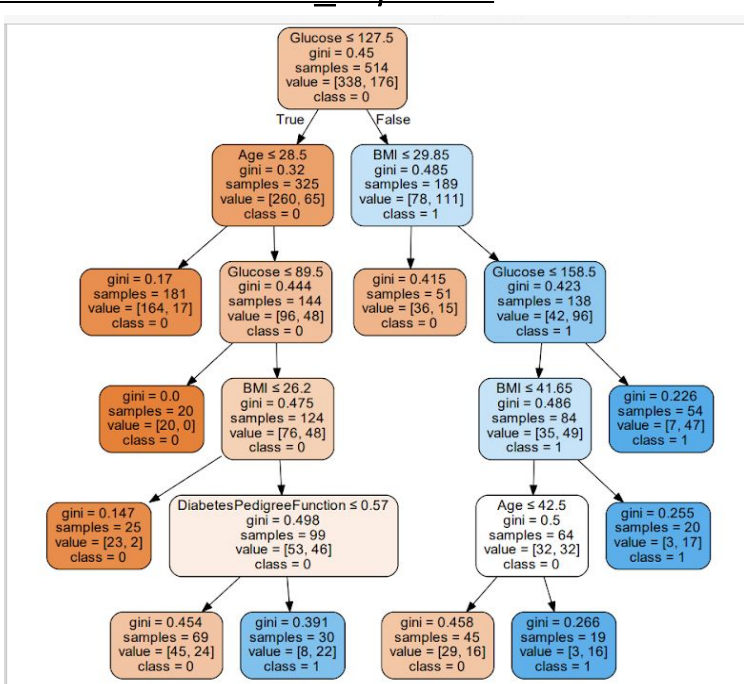


Fig. Decision tree with depth 5

Simplified Decision Tree with max_depth = 3

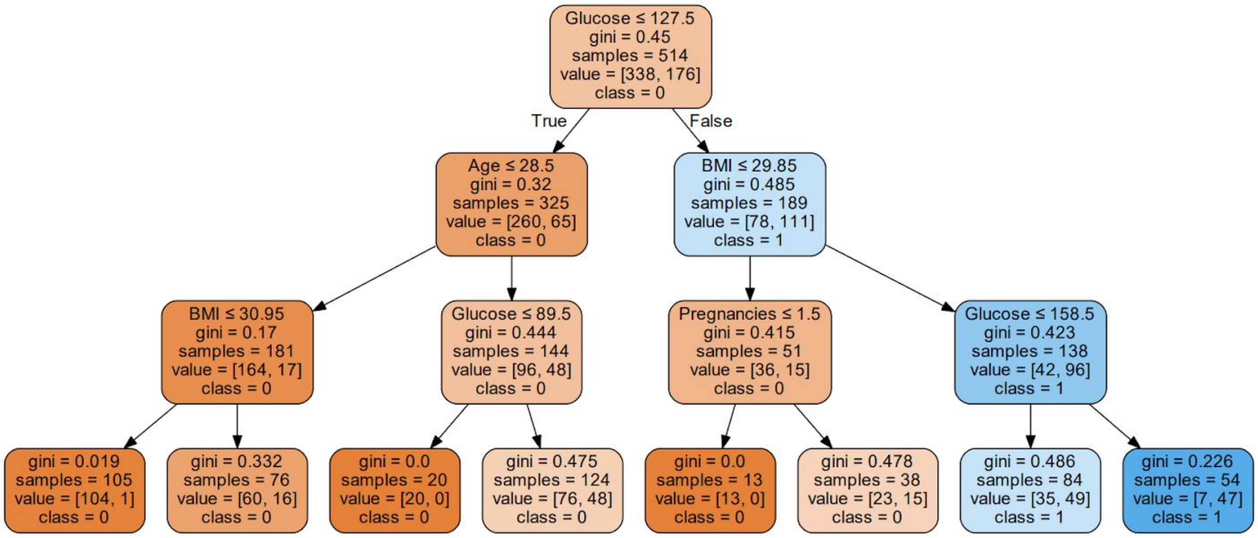


Fig. Decision tree with depth 5

Causes of condition

From all the above figures we can analyze using the Decision tree that the most important factor is Glucose which is the most pure factor with a gini index of 0.45. This logically makes sense if the person has more glucose levels in his/her body the patient will have diabetes.

The other two factors which are important are Age and BMI with a gini index of 0.32 and 0.485.

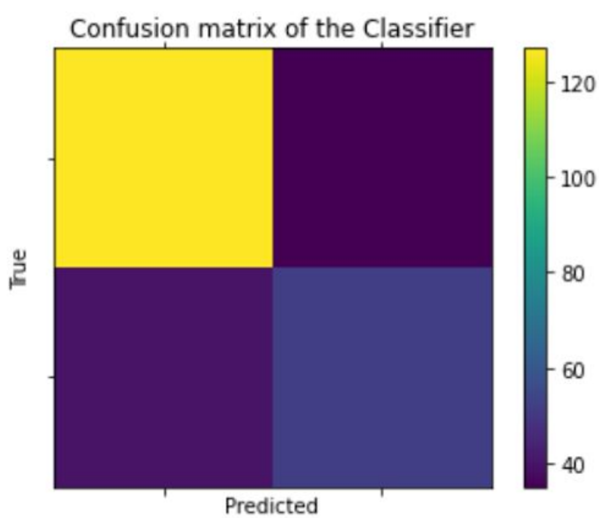
Finally other factors followed are DiabetesPedigreeFunction, Bloodpressure, Pregnancy and least is skinthickness.

Confusion Matrix

Confusion matrix helps us to analyze the performance of our classification model.

```
[[127  35]
 [ 40  52]]
```

254 Cases	Predicted No	Predicted Yes	
Actual No	127 (TN)	35 (FP)	**162**
Actual Yes	40 (FN)	52 (TP)	**92**
	167	**87**	



The table shows that out of 254 samples 87 were predicted to be diabetic and 167 were predicted to be non diabetic. Out of which in actual reality 92 were actually observed to be diabetic and 162 to be non diabetic.

True Negative - 127 cases
False Positive - 35 cases
False Negative - 40 cases
True Positive - 87 cases

Accuracy

Accuracy = (True Positive + True Negative) / Total
= (127+52)/254
= 70.47%

```
bc_tree.score(X_test, Y_test)
```

```
0.7047244094488189
```

The accuracy of our model is 70.47%