TASK 1

**PROBLEM STATEMENT 1:**

Try to understand the dataset of Social\_Network\_Ads.csv and try to find the best suitable ML algorithm and write the code in python for algorithm from scratch and try to achieve the below output plot.

**OBJECTIVE:**

The given social network data consists of four dependent x variables and one independent variables.

Main objective is to analysis the data and compare with 2 or more algorithms and to find the best fit.

* Data preprocessing
* Featured selection
* Scaling
* Fitting models
* Predicting the result
* Visualisation
* Comparing the accuracy
* Interpretation

**DATA PREPROCESSING:**

In this data cleaning is done and the missing values are imputed . In this Mean imputation method is used as it is a continuous data.

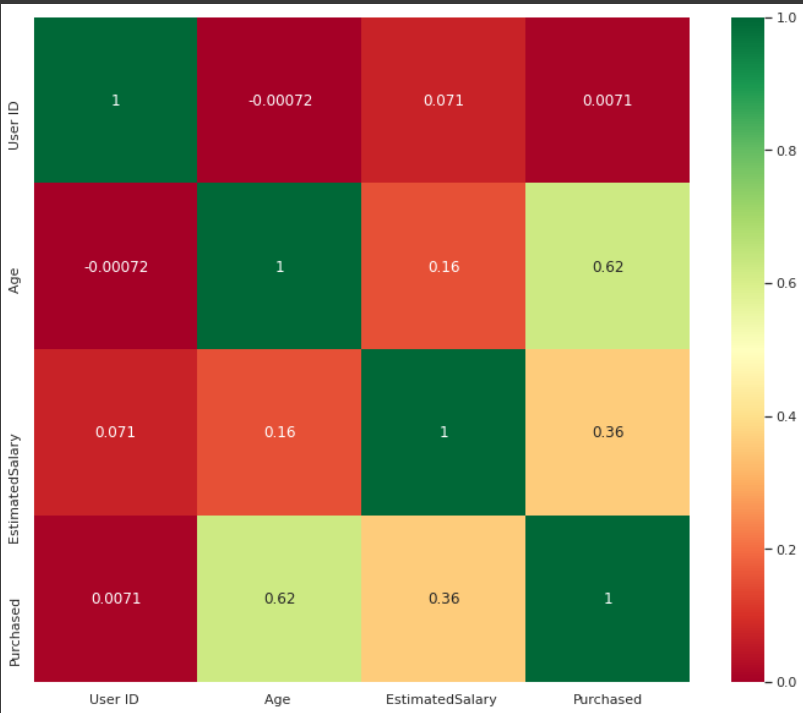
In this Mean imputation method the missing values under certain variables is replaced by the mean of the total values in the variable.

This is done so as to make the data efficient and it maintains the sample size and is also easy to use.

**FEATURE SELECTION:**

It is the process of reducing the number of parameters to maintain the optimality.In this I have used the method named correlation matrix. It is noting but the tables showing the relation between the n independent variables with one dependent variable.

From this we can able to identify which features are really necessary for our method. It reduces the attributes by retaining the information that is necessary.



**SCALING:**

This is a method used to normalize the features of the data. This is done because the higher values starts dominating . In this StandardScaler() , which removes the mean and scales of each features to unit variance. It results in a distribution with a standard deviation equals to 1.

[[ 0.58164944 -0.88670699]

[-0.60673761 1.46173768]

[-0.01254409 -0.5677824 ]

[-0.60673761 1.89663484]

[ 1.37390747 -1.40858358]

[ 1.47293972 0.99784738]

[ 0.08648817 -0.79972756]

[-0.01254409 -0.24885782]

[-0.21060859 -0.5677824 ]

[-0.21060859 -0.19087153]

[-0.30964085 -1.29261101]

[-0.30964085 -0.5677824 ]

[ 0.38358493 0.09905991]

[ 0.8787462 -0.59677555]

[ 2.06713324 -1.17663843]

[ 1.07681071 -0.13288524]

[ 0.68068169 1.78066227]

[-0.70576986 0.56295021]

[ 0.77971394 0.35999821]

[ 0.8787462 -0.53878926]

**FITTING THE MODEL:**

* **RANDOM FOREST:**

Random forests is a popular method for classification and regression that is operated by multitude of decision tree.

It is more accurate than building and computing individual trees and it is flexible ,and easy to use

* **NAIVE BAYES:**

It uses a similar method to predict the probability of different class based on various attributes. Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem.

* **SVM:**

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems.

* **DECISION TREE:**

A decision tree is a flowchart-like structure in which each internal node represents

a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each

branch represents the outcome of the test, and each leaf node represents a class

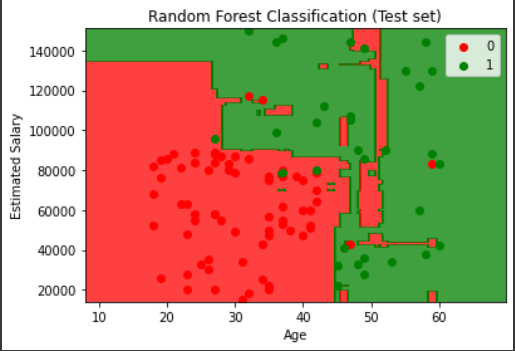
label.

**PREDICTING THE RESULT:**

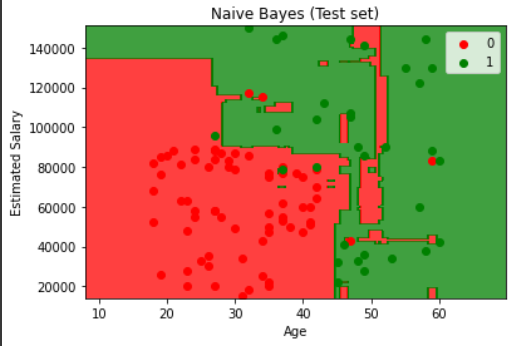
By using these models the target variable purchased was predicted based on the input variables .

**VISUALIZATION:**

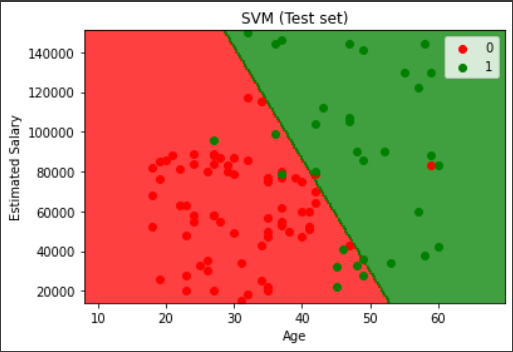
**RANDOM FOREST:**



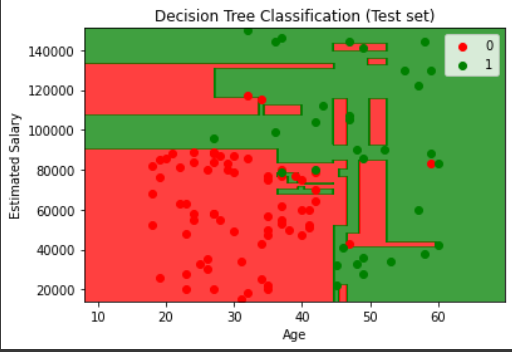
**NAIVE BAYES:**



**SVM:**



**DECISION TREE:**



**COMPARING THE ACCURACY:**

**RANDOM FOREST :**

Accuracy 0.91

**NAÏVE BAYES:**

Accuracy 0.9

**SVM:**

Accuracy 0.9

**DECISION TREE:**

Accuracy 0.9

**INTERPRETATION:**

When compared to other classification models these four show best results with accuracy more than 90% .But when the output plot of these four is compared random forest shows greater accuracy like the plot given in the problem statement.

**TASK 2**

**PROBLEM STATEMENT:**

Mandatory Predict which patient has diabetes from Diabetes Database.csv and try to understand the dataset attributes and try to figure out type ML model suits and build from scratch.

**OBJECTIVE:**

To analyze the Diabetes data and fits the best model which helps to predict the patients who has Diabetes .

The Diabetes data consists of 8 Dependent variables x and one Independent variables y.

* Data preprocessing
* Visualization
* Feature selection
* Scaling
* Fitting the model
* Confusion matrix
* Prediction
* Accuracy

**DATA PREPROCESSING:**

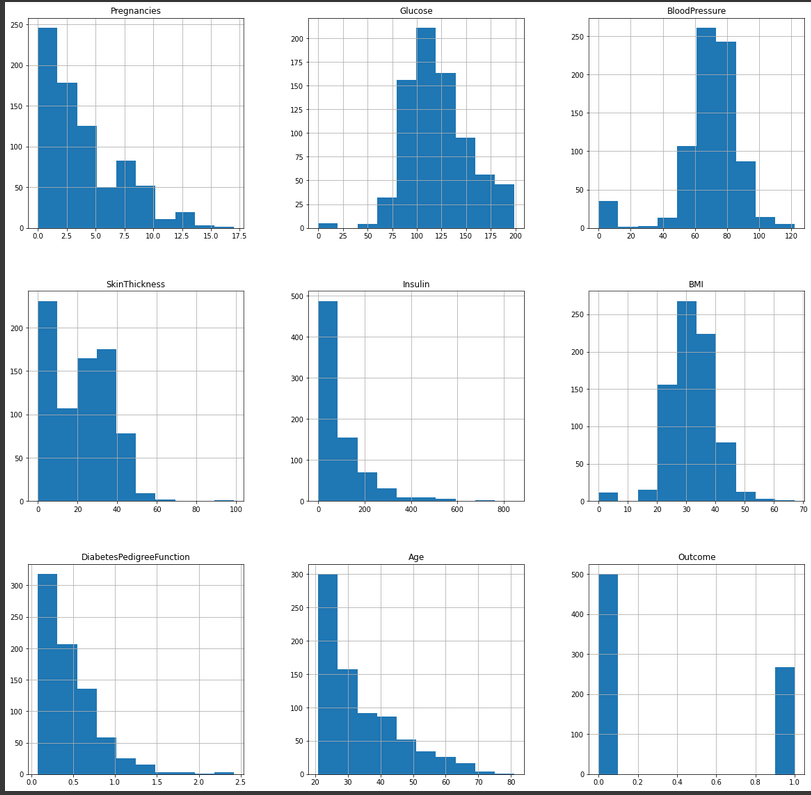
In this data cleaning is done and the missing values are imputed . In this Mean imputation method is used as it is a continuous data.

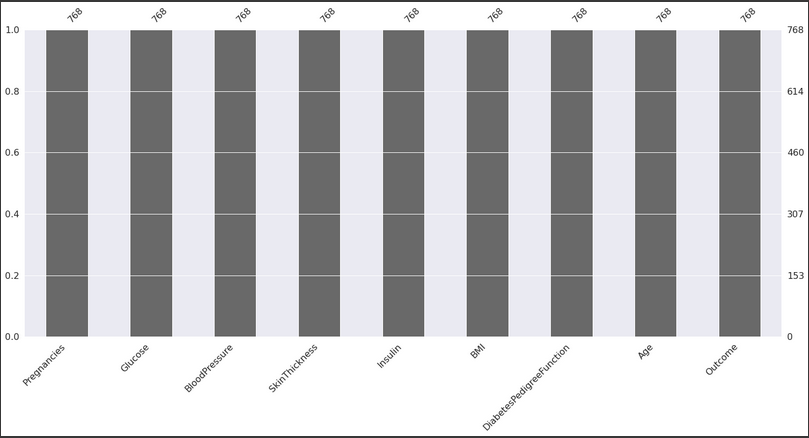
In this Mean imputation method the missing values under certain variables is replaced by the mean of the total values in the variable.

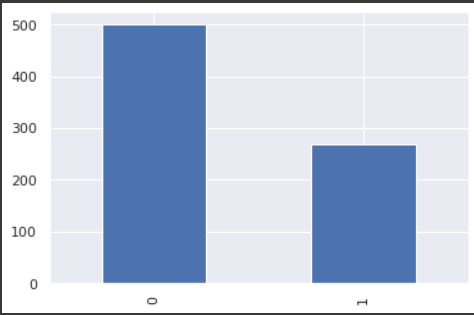
This is done so as to make the data efficient and it maintains the sample size and is also easy to use.

**VISUALIZATION:**

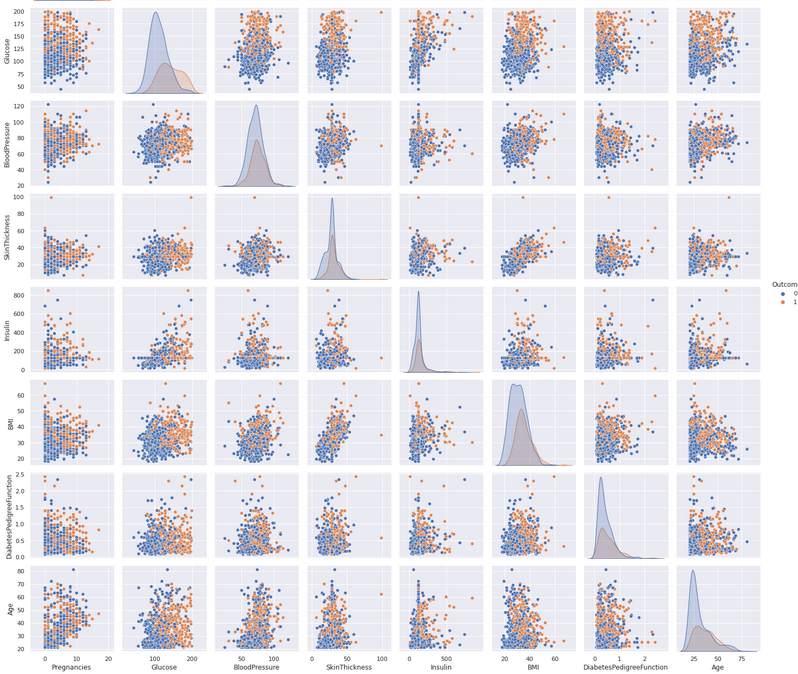
**Showing how each value in the variable changes**



****

****

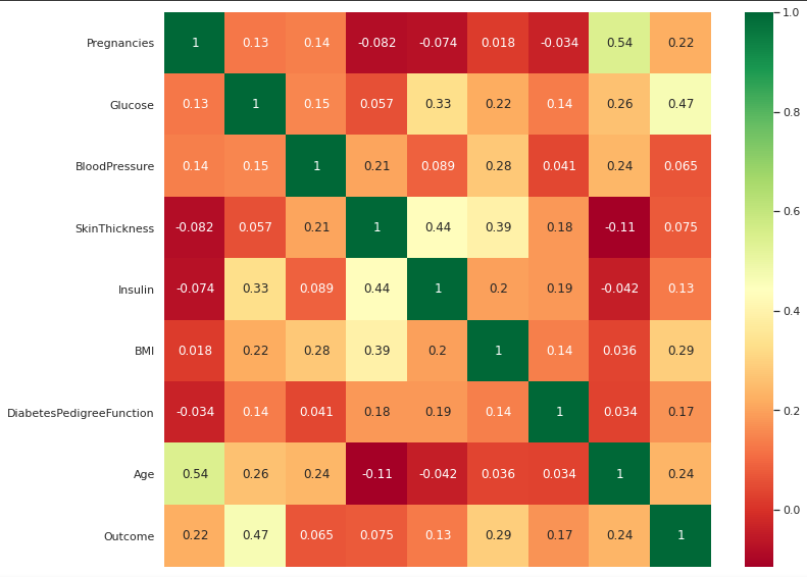
**Visualizing the relationship between variables**

****

**FEATURE SELECTION:**

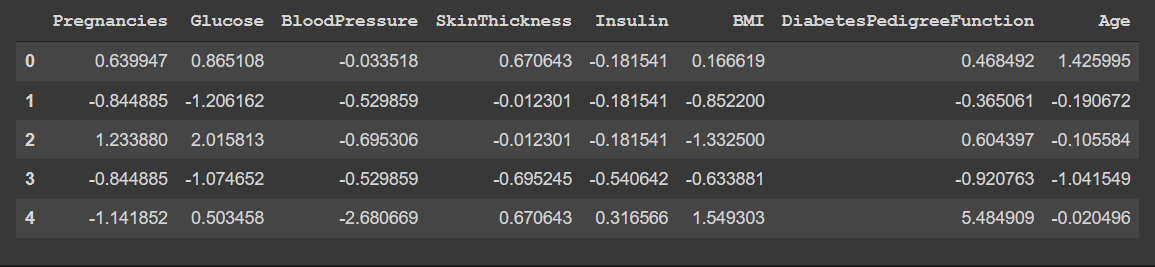
It is the process of reducing the number of parameters to maintain the optimality.In this I have used the method named correlation matrix. It is noting but the tables showing the relation between the n independent variables with one dependent variable.

From this we can able to identify which features are really necessary for our method. It reduces the attributes by retaining the information that is necessary.



**SCALING:**

This is a method used to normalize the features of the data. This is done because the higher values starts dominating . In this StandardScaler() , which removes the mean and scales of each features to unit variance. It results in a distribution with a standard deviation equals to 1.

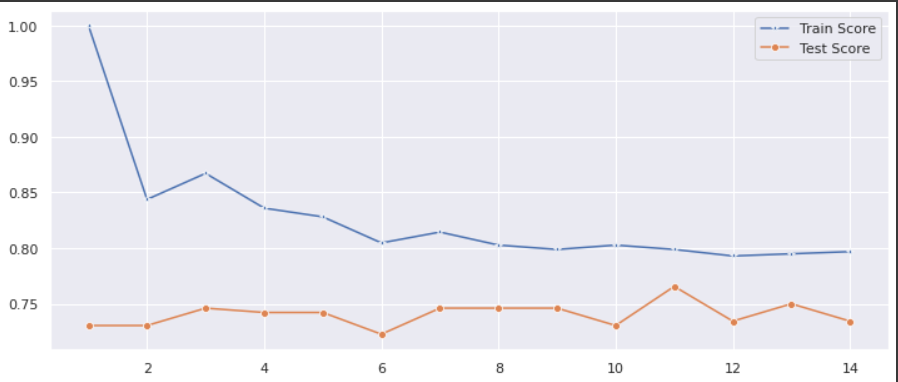


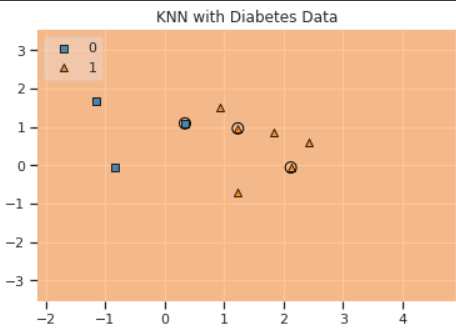
**FITTING THE MODEL:**

k-NN is a type of instance-based learning, or lazy learning, where the function

is only approximated locally and all computation is deferred until classification.

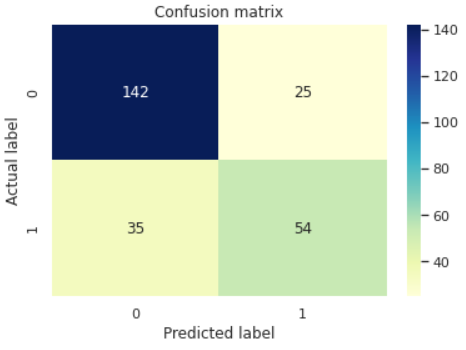
**k-NN : Varying Number of Neighbours**





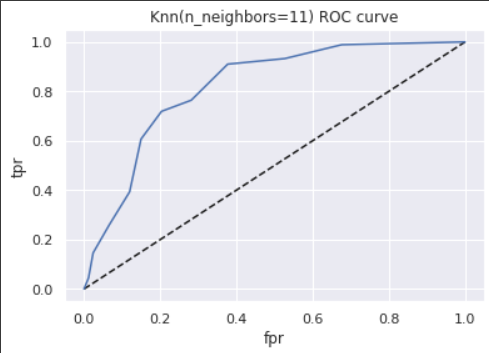
**CONFUSION MATRIX:**

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known.



**ACCURACY:**

ROC\_AUC\_SCORE : **0.8193500639171096**







**INTERPRETATION:**

From the test score and train score we can able to conclude that the graph dose’nt deviated much indicating that the model is fitted best for the data.