**Overview**

**• Objective**

The objective of this project is to predict the possibility of Heart Attack and to find out who’s more likely to have heart attack. It will help individual and health care personal to find ways to prevent heart attack of human being.

The scope of the project extends to all health care or diagnostic centre to prevent Heart Attack.

➢ Whether Increasing in age has any effect towards heart attack?

➢ Does increase in cholesterol level in body have any effect towards the heart attack?

➢ Does Increase in blood pressure have any relation with heart attack?

➢ Does family history have any effect towards heart attack?

**• Data Science Proposal**

➢ Draw some insights by plotting above mentioned variables against the target variable

➢ Check for missing values in the data, and if any, will process the data accordingly.

➢ Understand how the features are related with our target variable – Target

➢ Apply feature selection and feature engineering to make it model ready

➢ Apply Random forest Decision Tree & Gradient Based Classifier model to check which one is the most   
 suitable

➢ Draw out recommendations based on our analysis

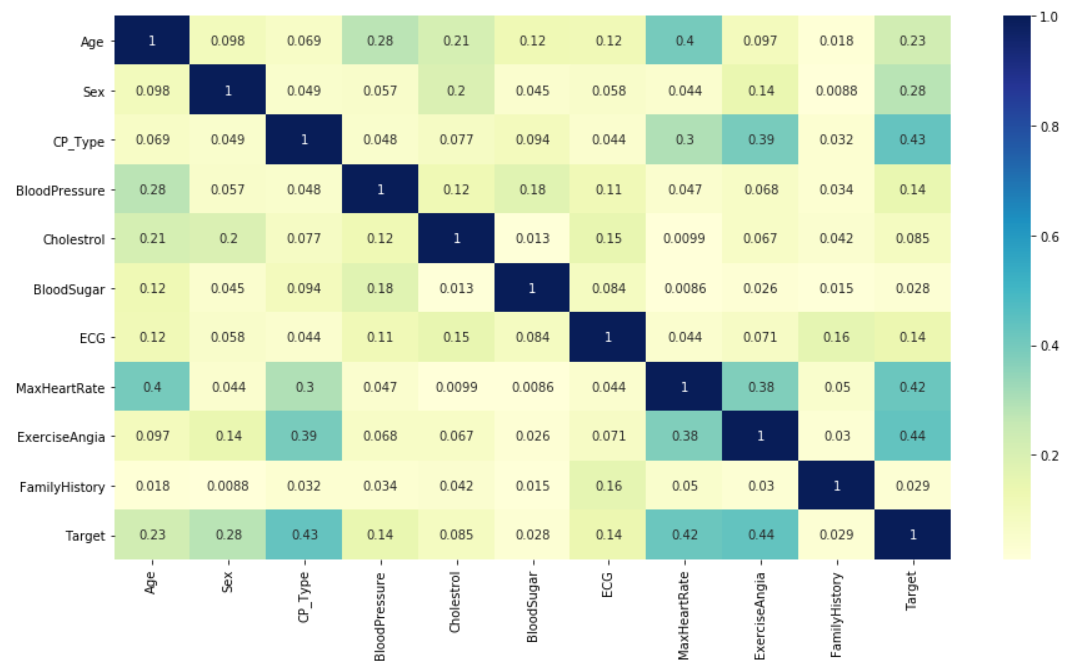
**Wrangling/Pre-Processing**

* No Missing values were there, so pre-processing was not required.

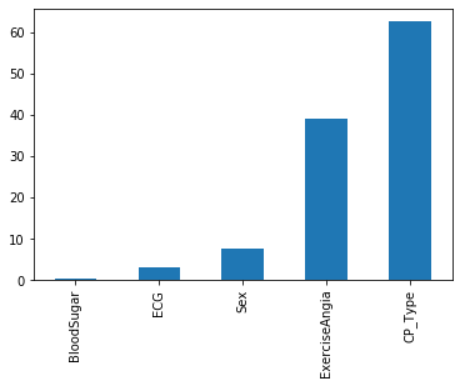
**Feature Selection**

Two feature selection technique have been used:

1. **Pearson correlation:** Pearson’s Correlation Coefficient helps you find out the relationship between two quantities. It gives you the measure of the strength of association between two variables. The value of Pearson’s Correlation Coefficient can be between -1 to +1. 1 means that they are highly correlated and 0 means no correlation. -1 means that there is a negative correlation. Think of it as an inverse proportion.

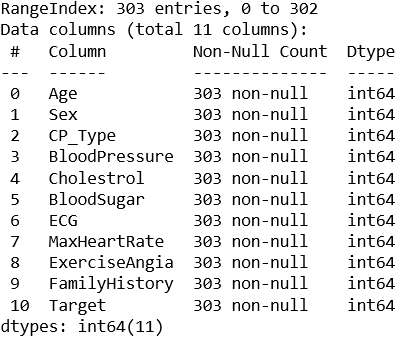


2. **Chi-Squared:** This is filter-based method. In this method, we calculate the chi-square metric between the target and the *Categorical* variable and only select the variable with the maximum chi-squared values. Best suited when we have categorical input data when the target variable is also categorical (e.g. classification predictive modelling) are the chi-squared statistic.



**Dataset**

* **How many features** :11 (includes Target feature)
* **Size of the dataset**: 303 (rows),11(columns)
* **Multiple files**: No (“Heart\_Attack\_Analysis\_Data.csv”)
* **What kind of data**: numerical (Integers)
* **what is the distribution**: Imbalanced
* **Distribution of Training set and Testing set**: 80% & 20%
* **Missing data and Preprocessing challenges**: No missing data



**Methodology**

The process of modelling means training a machine learning algorithm to predict the labels from the features, tuning it for the analysis need, and validating it on holdout data.

Models used for Heart Attack Analysis:

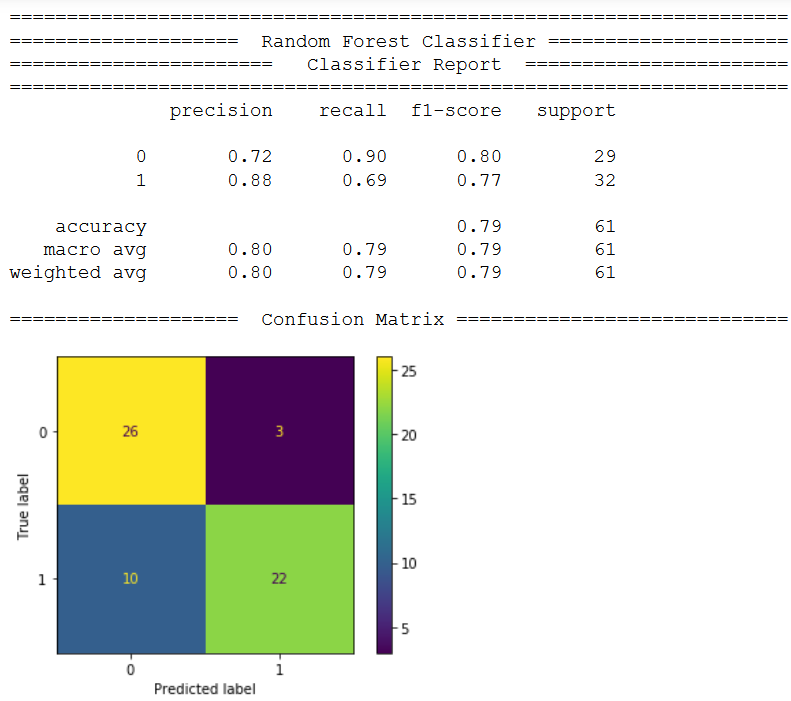
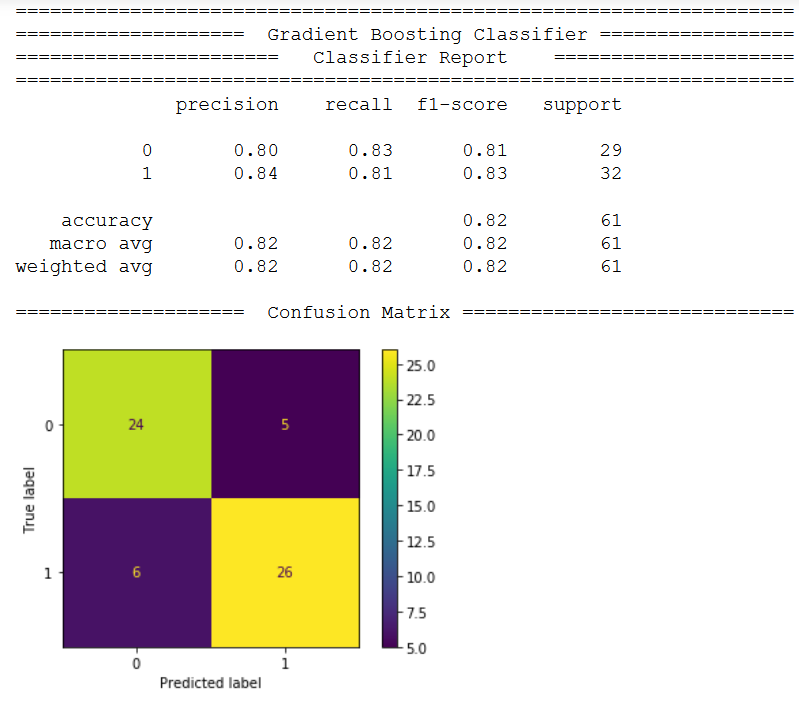
* **Gradient Boosting classifier** - The Gradient Boosting Classifier depends on a loss function. A custom loss function can be used, and many standardized loss functions are supported by gradient boosting classifiers, but the loss function has to be differentiable. Classification algorithms frequently use logarithmic loss, while regression algorithms can use squared errors. Gradient boosting systems don't have to derive a new loss function every time the boosting algorithm is added, rather any differentiable loss function can be applied to the system. Gradient boosting systems have two other necessary parts: a weak learner and an additive component. Gradient boosting systems use decision trees as their weak learners. Regression trees are used for the weak learners, and these regression trees output real values. Because the outputs are real values, as new learners are added into the model the output of the regression trees can be added together to correct for errors in the predictions.
* **Random Forest** - Random Forest is a trademark term for an ensemble of decision trees. In Random Forest, we’ve collection of decision trees (so known as “Forest”). To classify a new object based on attributes, each tree gives a classification, and we say the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Each tree is planted & grown as follows: If the number of cases in the training set is N, then sample of N cases is taken at random but with replacement. This sample will be the training set for growing the tree. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing. Each tree is grown to the largest extent possible. There is no pruning.

**Results**

1. The data consists of more than twice the number of people with sex = 1 than sex = 0.
2. The scatterplot heatmap matrix suggests that there might be some correlation between Target and CP\_Type and ECG.
3. It is intuitive that elder people might have higher chances of heart attack but according to the distribution plot of age wrt Target, it is evident that this isn't the case.
4. According to the distribution plot of MaxHeartRate wrt Target, people with higher maximum heart rate achieved have higher chances of heart attack.
5. According to the distribution plot of FamilyHistory wrt Target, number of person affected in family has no influence.
6. The categorical count plot tells about the following -
   * People with chest pain, that is with CP\_Type = 2 have higher chances of heart attack.
   * People with ECG = 1 have high chance of heart attack.
   * People with sex = 0 have higher chance of heart attack.
   * People with BloodSugar = 0 have much higher chance of heart attack.
   * People with no exercise induced angina, that is with ExerciseAngia = 0 have higher chance of heart attack.

**Comparison/Conclusion**

* Table and Plot for the evaluation metric of Random Forest and Gradient Boosting Classifier

* **Conclusion:** It can be considered that Gradient Boosting Classifier is showing higher accuracy compared to Random Forest Classifier from the confusion Matrix for the given dataset.