Heart Disease Prediction Report

**Introduction:**

The given dataset contains the information about Heart disease. The dataset contains 14 columns. The Goal is to predict where the patient having the heart disease or not. The target variable is named as “Target” in dataset where 0- no presence of Heart disease and 1 – presence of Heart disease.

**Exploratory Data Analysis:**

The objectives of the EDA are as follows:

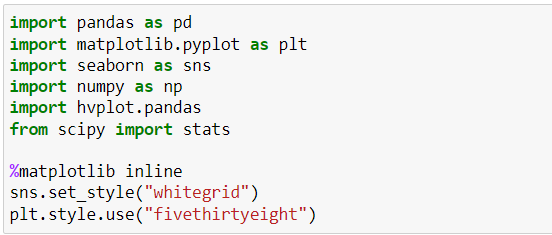
i. To get an overview of the distribution of the dataset.

ii. Check for missing numerical values, outliers or other anomalies in the dataset.

iii. Discover patterns and relationships between variables in the dataset.

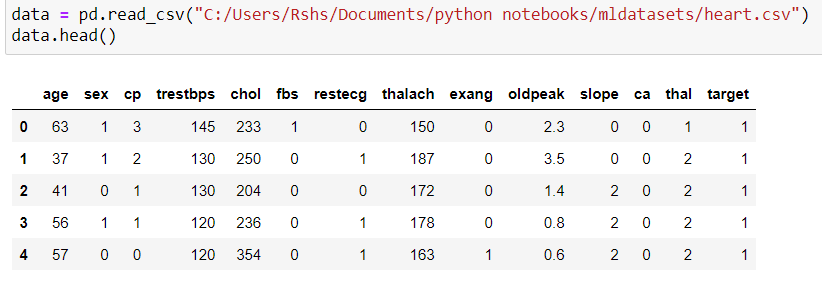
iv. Check the underlying assumptions in the dataset.

Importing the required Libraries

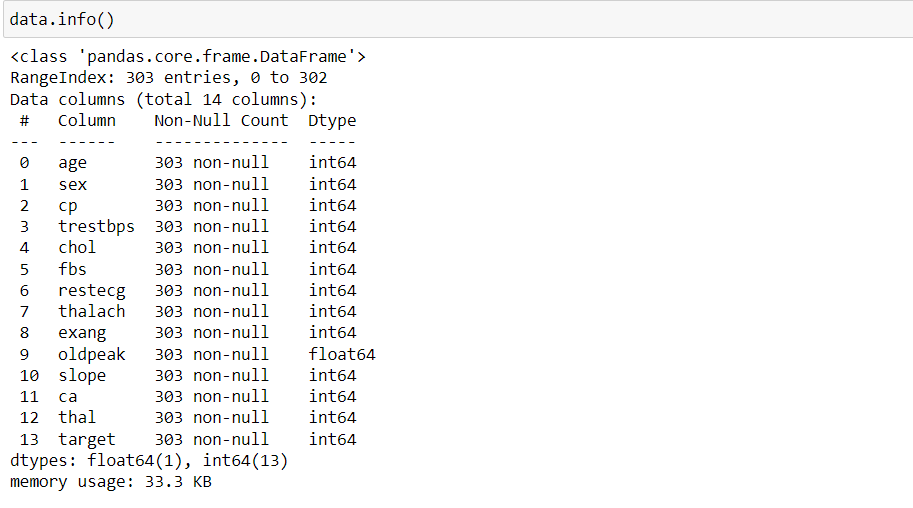


Next, load the data and let’s take a look,

Preview the dataset



Summary of the dataset

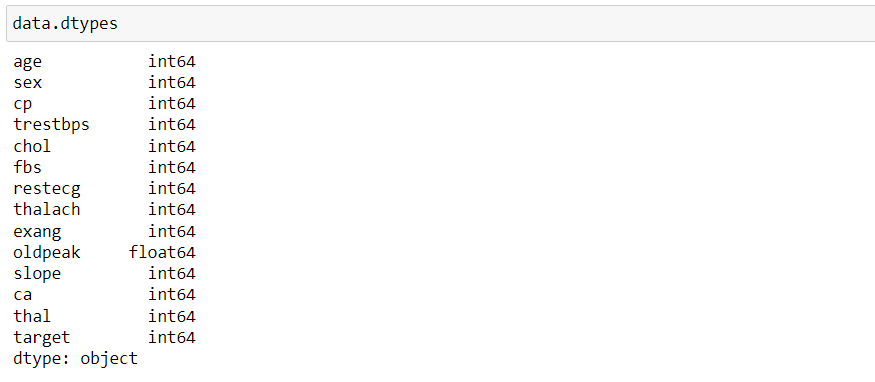


**Dataset Description**

The dataset contains several columns which are as follows -

* age: age in years
* sex: (1 = male; 0 = female)
* cp: chest pain type
* trestbps: resting blood pressure (in mm Hg on admission to the hospital)
* chol: serum cholestoral in mg/dl
* fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
* restecg: resting electrocardiographic results
* thalach: maximum heart rate achieved
* exang: exercise induced angina (1 = yes; 0 = no)
* oldpeak: ST depression induced by exercise relative to rest
* slope: the slope of the peak exercise ST segment
* ca: number of major vessels (0-3) colored by flourosopy
* thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
* target: 1 or 0

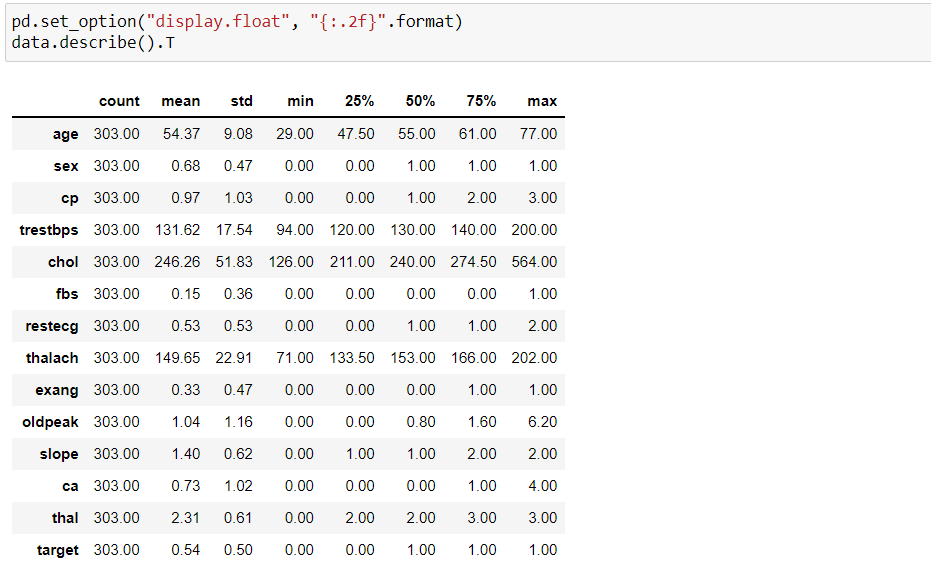
Check the data types of columns:



Findings:

* sex is a character variable. Its data type should be object. But it is encoded as (1 = male; 0 = female). So, its data type is given as int64.
* Same is the case with several other variables - fbs, exang and target.
* fbs (fasting blood sugar) should be a character variable as it contains only 0 and 1 as values (1 = true; 0 = false). As it contains only 0 and 1 as values, so its data type is given as int64.
* exang (exercise induced angina) should also be a character variable as it contains only 0 and 1 as values (1 = yes; 0 = no). It also contains only 0 and 1 as values, so its data type is given as int64.
* target should also be a character variable. But, it also contains 0 and 1 as values. So, its data type is given as int64.

Statistical properties of dataset:

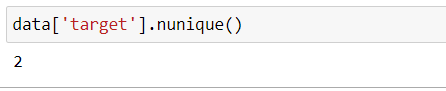


**Univariate analysis:**

Analysis of target feature variable:

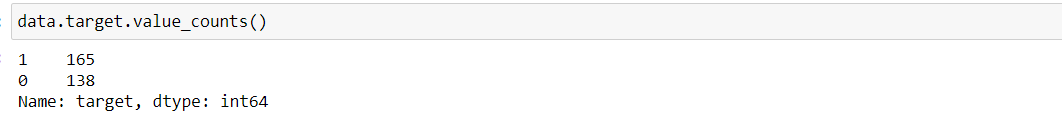
* Our feature variable of interest is target.
* It refers to the presence of heart disease in the patient.
* It is integer valued as it contains two integers 0 and 1 - (0 stands for absence of heart disease and 1 for presence of heart disease).
* So, in this section, I will analyze the target variable.

Check the number of unique values in target variable:



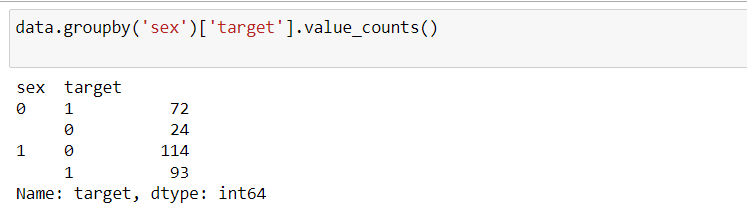
We can see that there are 2 unique values in the target variable.

Frequency distribution of target variable:



* 1 stands for presence of heart disease. So, there are 165 patients suffering from heart disease.
* Similarly, 0 stands for absence of heart disease. So, there are 138 patients who do not have any heart disease.

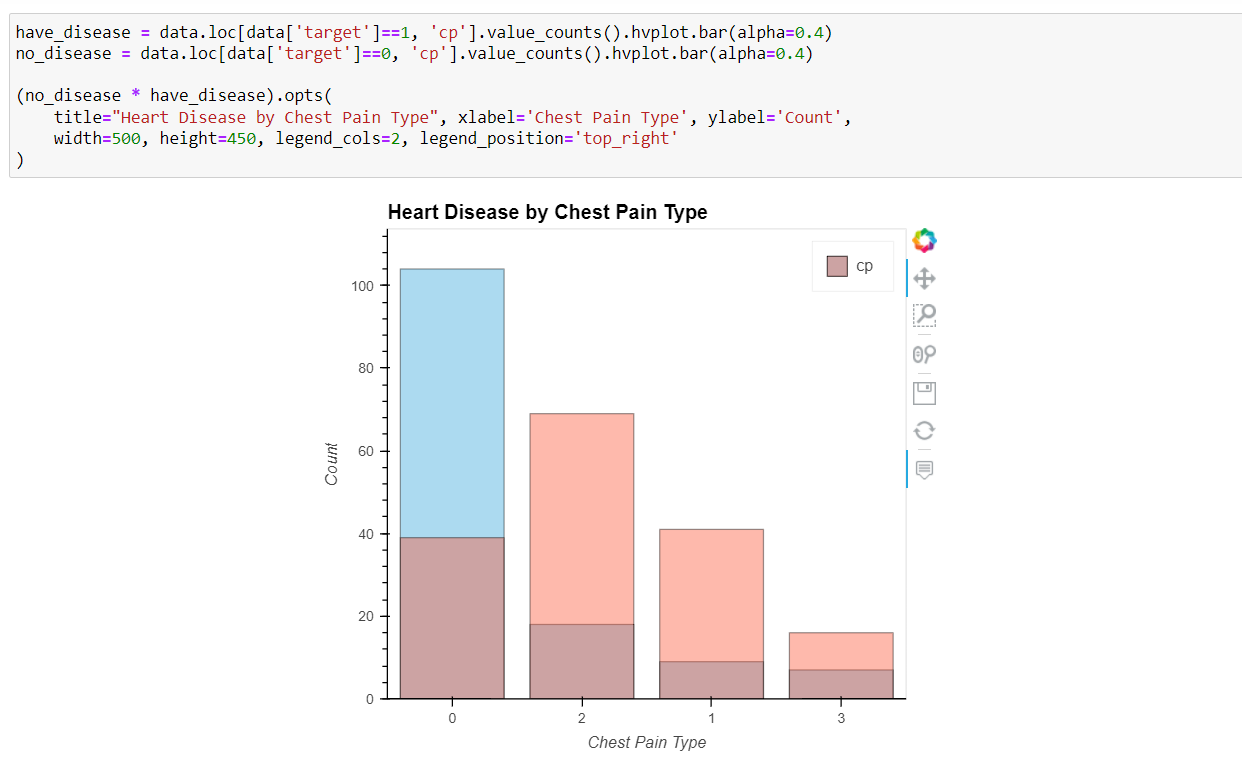
Frequency distribution of target variable with respect to sex:



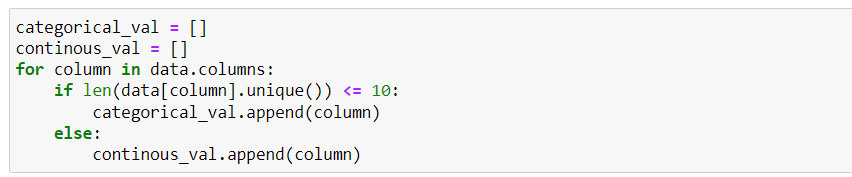
* sex variable contains two integer values 1 and 0 : (1 = male; 0 = female).
* target variable also contains two integer values 1 and 0 : (1 = Presence of heart disease; 0 = Absence of heart disease)
* So, out of 96 females - 72 have heart disease and 24 do not have heart disease.
* Similarly, out of 207 males - 93 have heart disease and 114 do not have heart disease.

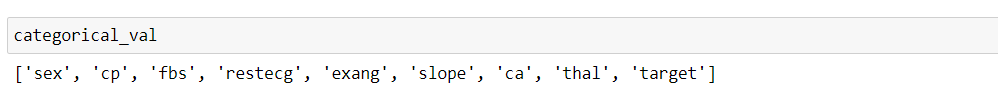
Some data visualizations,





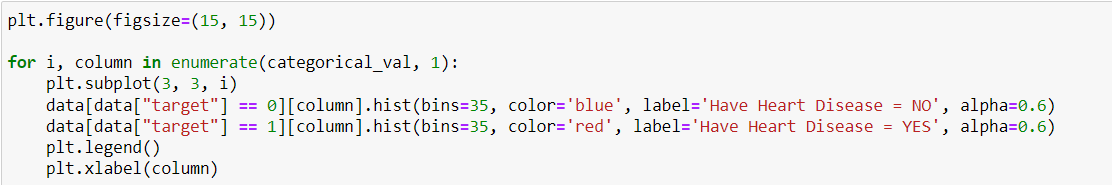
Finding the categorical variables in the dataset

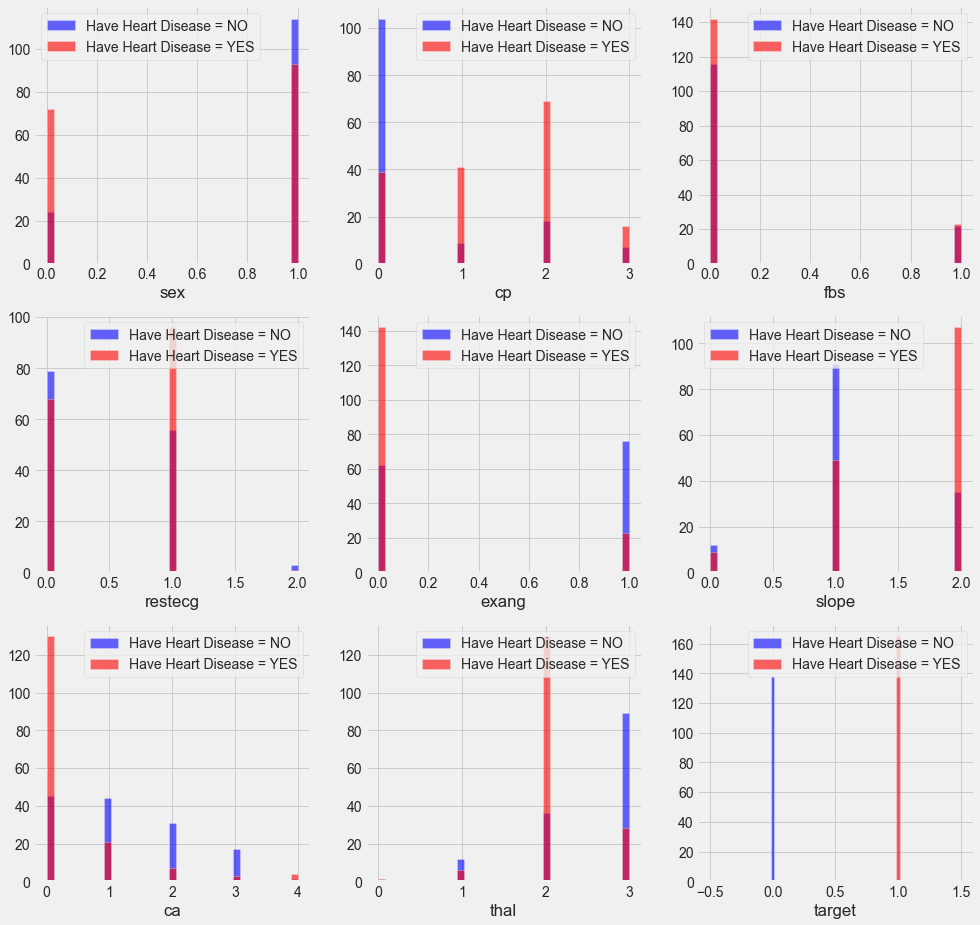




And these are the categorical variables in the dataset

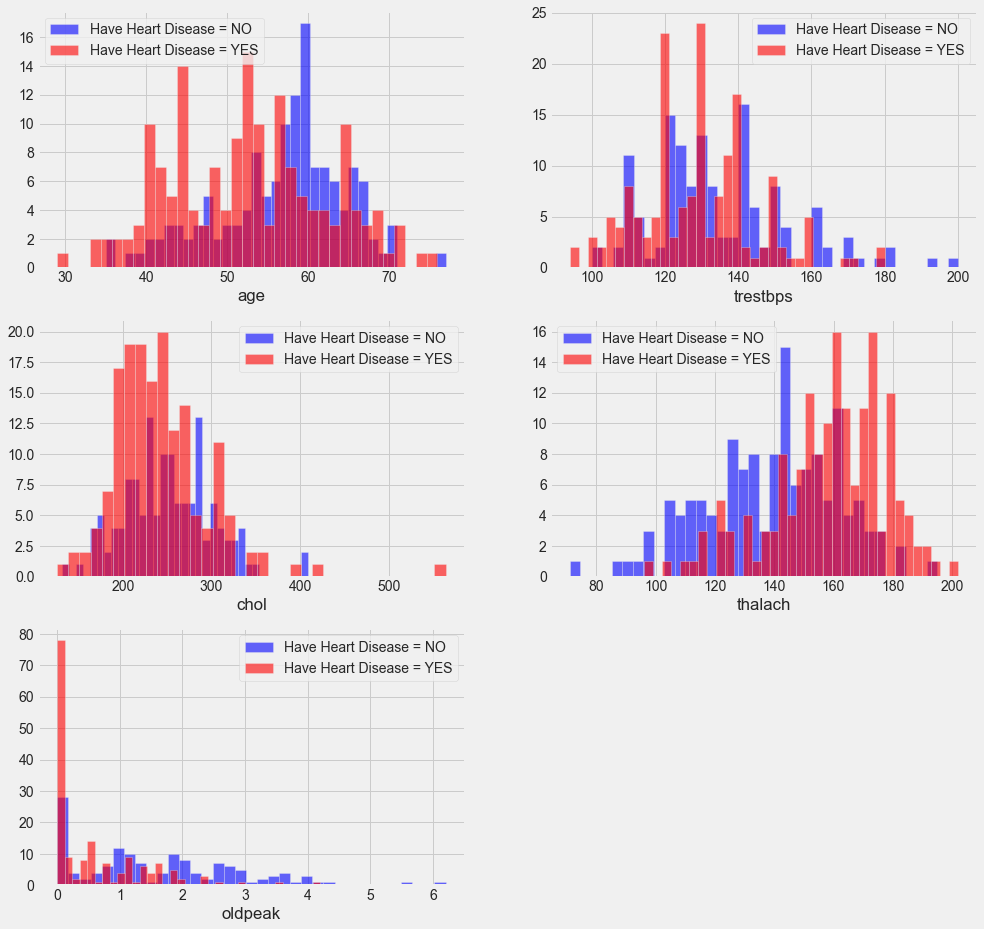
On further visualizing the dataset,





* cp {Chest Pain} : People with cp equl to 1, 2, 3 are more likely to have heart disease than people with cp equal to 0.
* restecg {resting electrocardiographic results} : People with value 1 (signals non-normal heart beat, can range from mild symptoms to severe problems) are more likely to have heart disease.
* exang {exercise induced angina} : People with value 0 (No ==> exercise induced angina) have heart disease more than people with value 1 (Yes ==> exercise induced angina)
* slope {the slope of the peak exercise ST segment} : People with slope value equal to 2 (Downslopins: signs of unhealthy heart) are more likely to have heart disease than people with slope value equal to 0 (Upsloping: better heart rate with exercise) or 1 (Flatsloping: minimal change (typical healthy heart)).
* ca {number of major vessels (0-3) colored by flourosopy} : the more blood movement the better so people with ca equal to 0 are more likely to have heart disease.
* thal {thalium stress result} : People with thal value equal to 2 (fixed defect: used to be defect but ok now) are more likely to have heart disease.

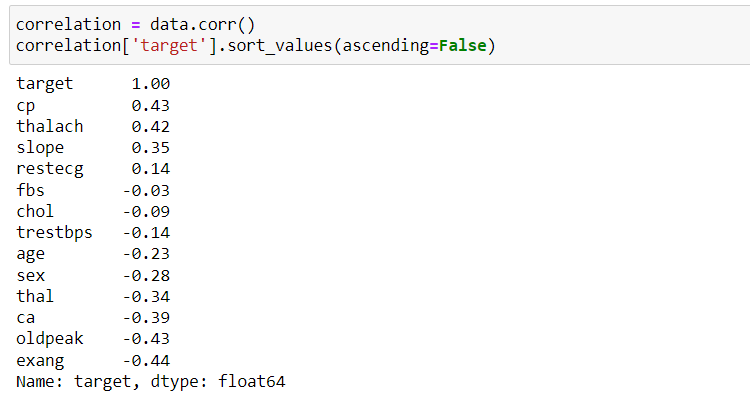




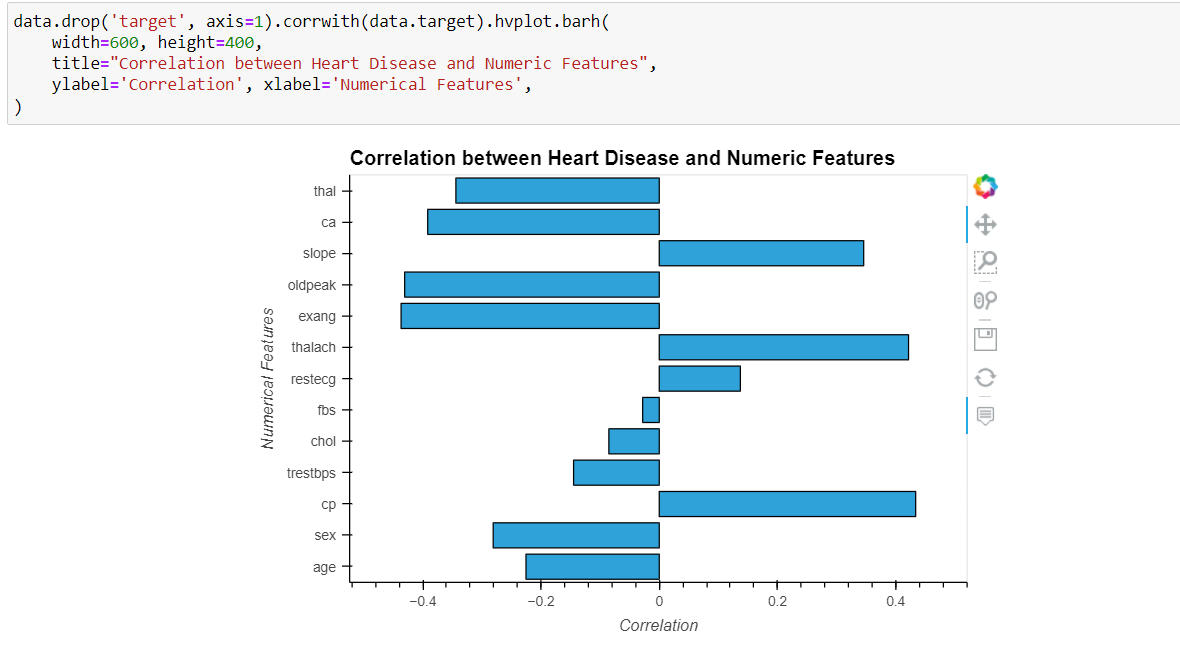
* trestbps : resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern
* chol {serum cholesterol in mg/dl} : above 200 is cause for concern.
* thalach {maximum heart rate achieved} : People how achieved a maximum more than 140 are more likely to have heart disease.
* oldpeak ST depression induced by exercise relative to rest looks at stress of heart during exercise unhealthy heart will stress more

**Bivariate Analysis:**

Estimate correlation coefficients and The target variable is target. So, we should check how each attribute correlates with the target variable.



**Visualizing the correlation of each attribute to the target**

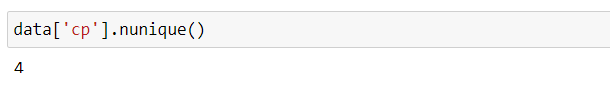


* The correlation coefficient ranges from -1 to +1.
* When it is close to +1, this signifies that there is a strong positive correlation. So, we can see that there is no variable which has strong positive correlation with target variable.
* When it is close to -1, it means that there is a strong negative correlation. So, we can see that there is no variable which has strong negative correlation with target variable.
* When it is close to 0, it means that there is very low correlation. So, there is very low correlation between target and fbs.
* We can see that the cp and thalach variables are mildly positively correlated with target variable. So, I will analyse the interaction between these features and target variable.

**Analysis of target and cp variable:**

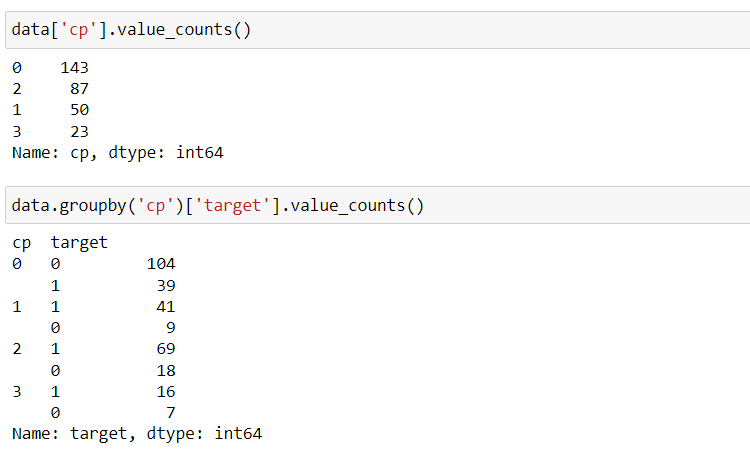
**Explore cp variable:**

cp stands for chest pain type. First, I will check number of unique values in cp variable.



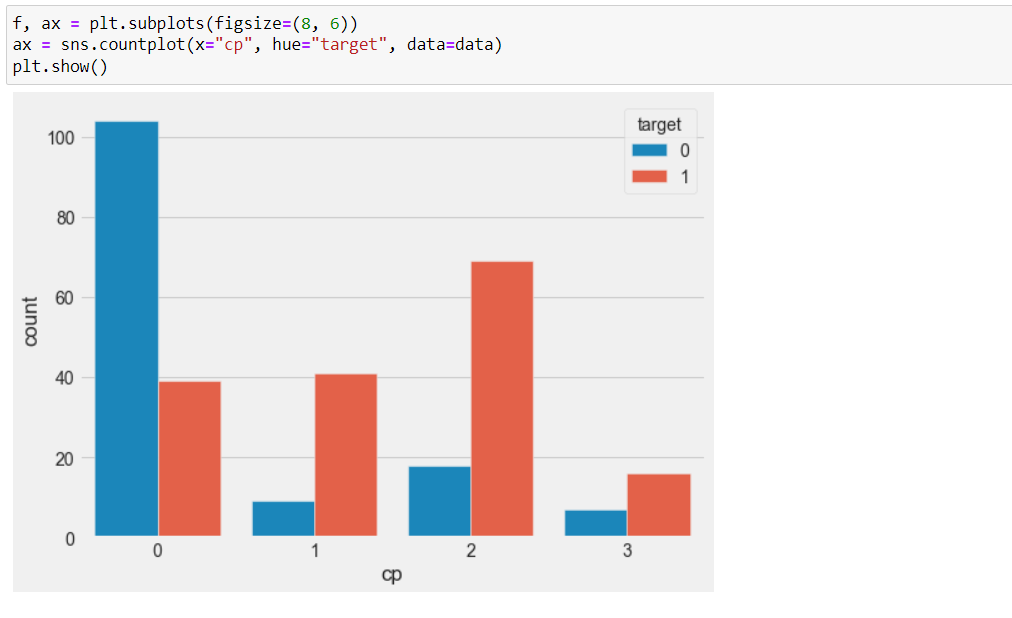
So, there are 4 unique values in cp variable. Hence, it is a categorical variable.

**Frequency Distribution of cp:**



* It can be seen that cp is a categorical variable and it contains 4 types of values - 0, 1, 2 and 3.
* target variable contains two integer values 1 and 0 : (1 = Presence of heart disease; 0 = Absence of heart disease)
* So, the above analysis gives target variable values categorized into presence and absence of heart disease and groupby cp variable values.

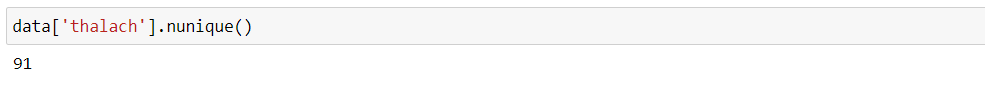
**Visualizing the value counts of the cp variable with respect to target:**



**Analysis of target and thalach variable:**

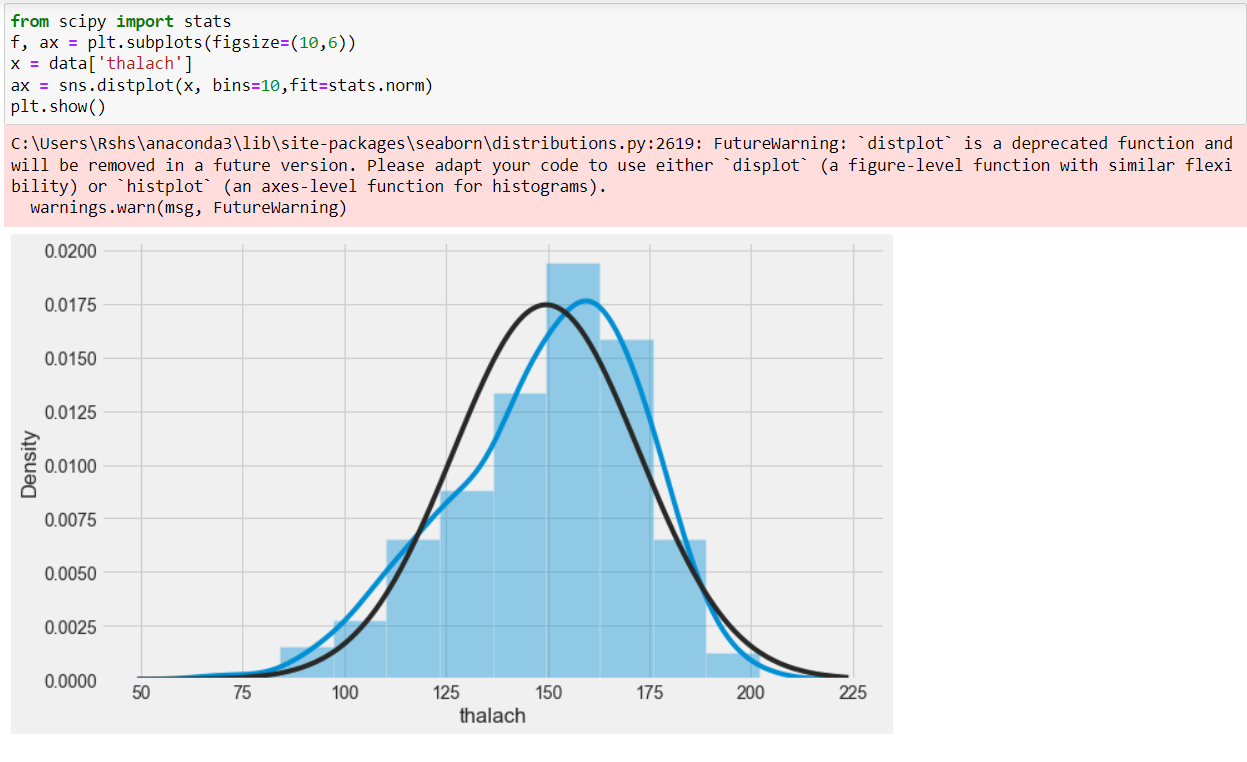
**Explore thalach variable:**

thalach stands for maximum heart rate achieved. The number of unique values in thalach is as follows



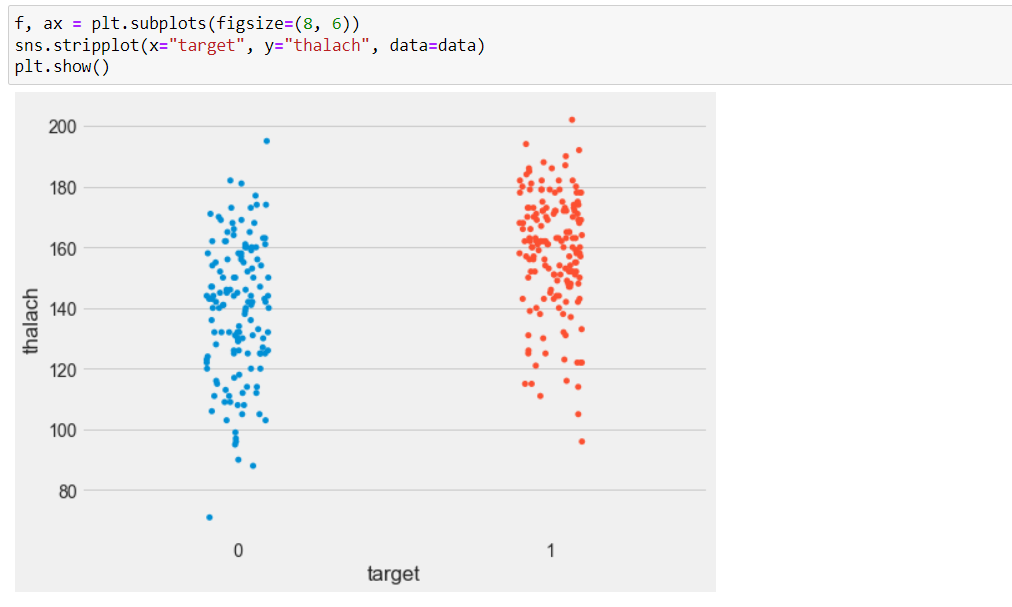
So, number of unique values in thalach variable is 91. Hence, it is numerical variable.

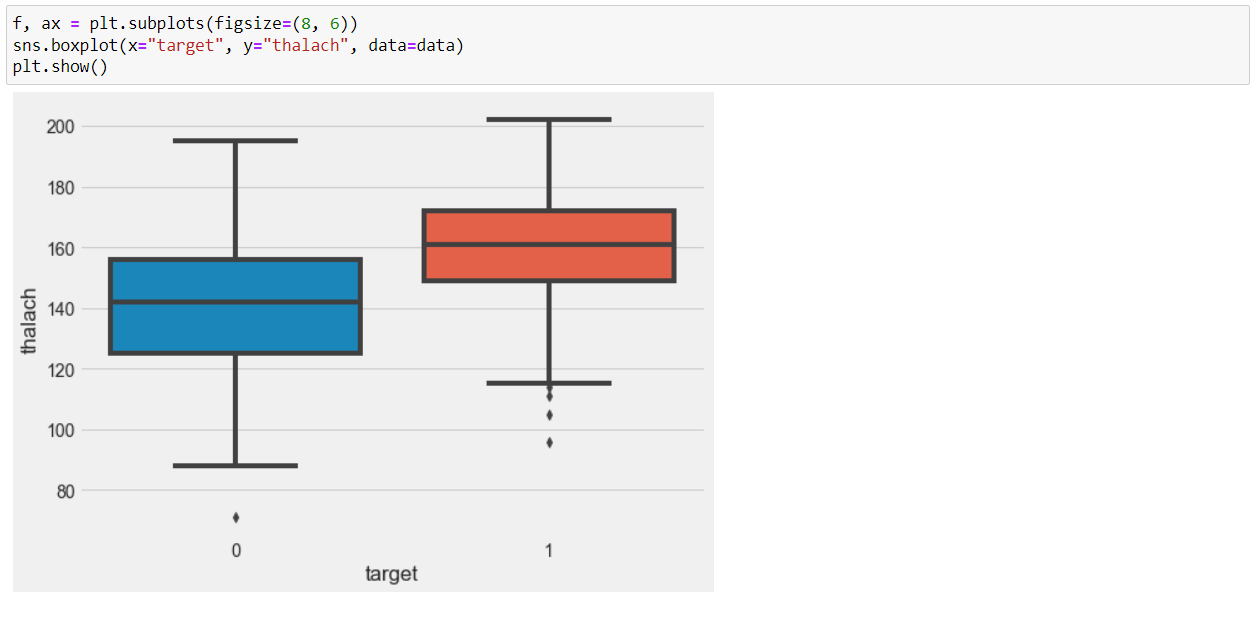
**Visualize the frequency distribution of thalach variable:**



We can see that the thalach variable is slightly negatively skewed.

**Visualize frequency distribution of thalach variable with respect to target:**





We can see that those people suffering from heart disease (target = 1) have relatively higher heart rate (thalach) as compared to people who are not suffering from heart disease (target = 0).

**Findings of Bivariate Analysis:**

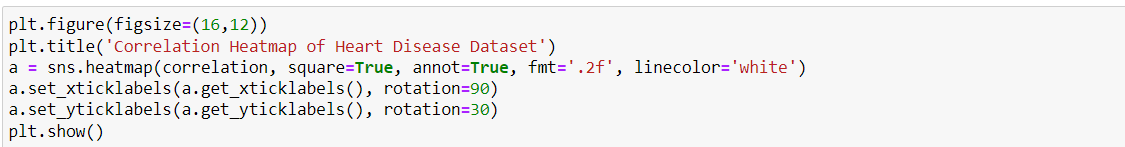
* There is no variable which has strong positive correlation with target variable.
* There is no variable which has strong negative correlation with target variable.
* There is very low correlation between target and fbs.
* The cp and thalach variables are mildly positively correlated with target variable.
* We can see that the thalach variable is slightly negatively skewed.
* The people suffering from heart disease (target = 1) have relatively higher heart rate (thalach) as compared to people who are not suffering from heart disease (target = 0).
* The people suffering from heart disease (target = 1) have relatively higher heart rate (thalach) as compared to people who are not suffering from heart disease (target = 0)

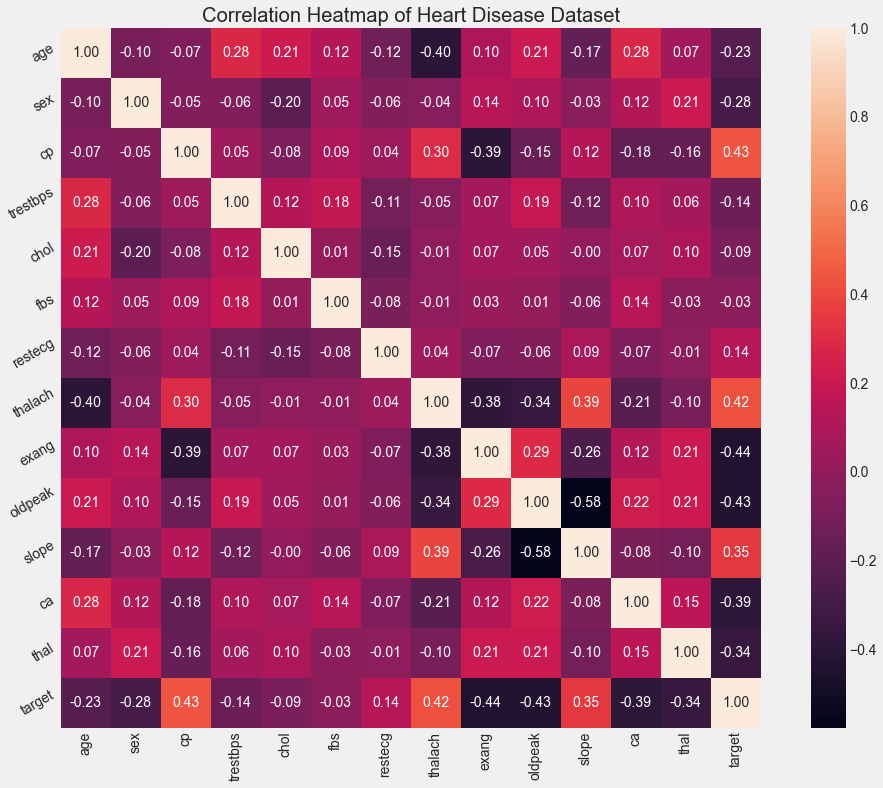
**Multivariate analysis:**

The objective of the multivariate analysis is to discover patterns and relationships in the dataset.

**Discover patterns and relationships:**

* An important step in EDA is to discover patterns and relationships between variables in the dataset.
* I will use heat map and pair plot to discover the patterns and relationships in the dataset.





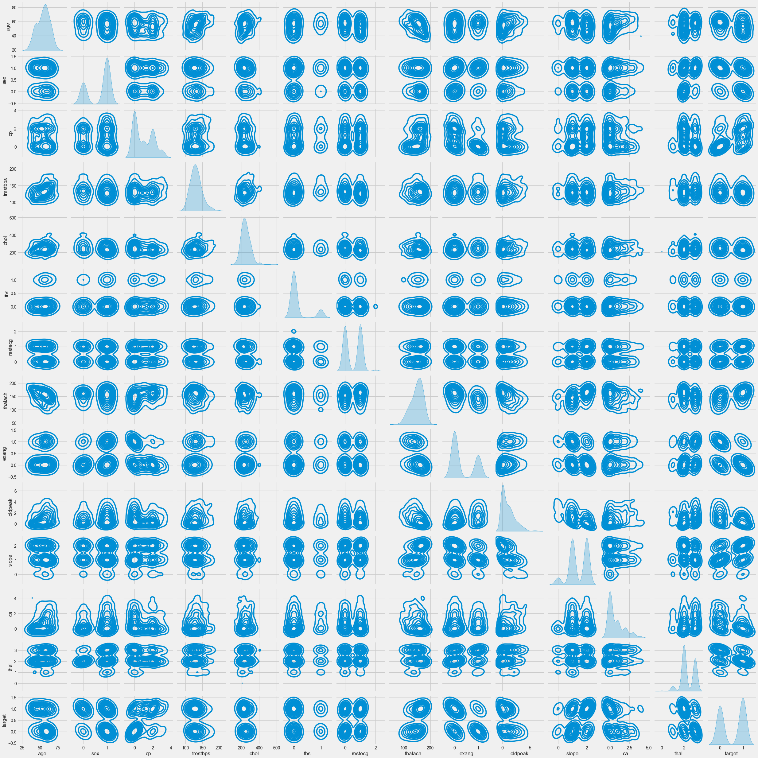
Interpretation:

From the above correlation heat map, we can conclude that,

* target and cp variable are mildly positively correlated (correlation coefficient = 0.43).
* target and thalach variable are also mildly positively correlated (correlation coefficient = 0.42).
* target and slope variable are weakly positively correlated (correlation coefficient = 0.35).
* target and exang variable are mildly negatively correlated (correlation coefficient = -0.44).
* target and oldpeak variable are also mildly negatively correlated (correlation coefficient = -0.43).
* target and ca variable are weakly negatively correlated (correlation coefficient = -0.39).
* target and thal variable are also weakly negatively correlated (correlation coefficient = -0.34).

**Pairplot:**



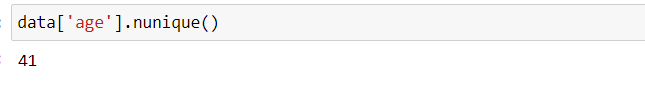


From this pairplot we can see that there are no relationship between the variables and they all have a clustered relation.

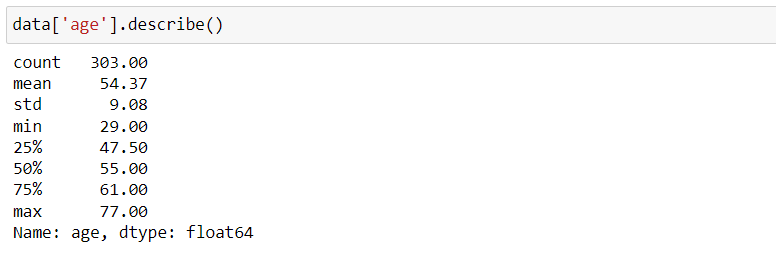
**Analysis of Variables:**

**Age:**

Check the number of unique values in age variable:

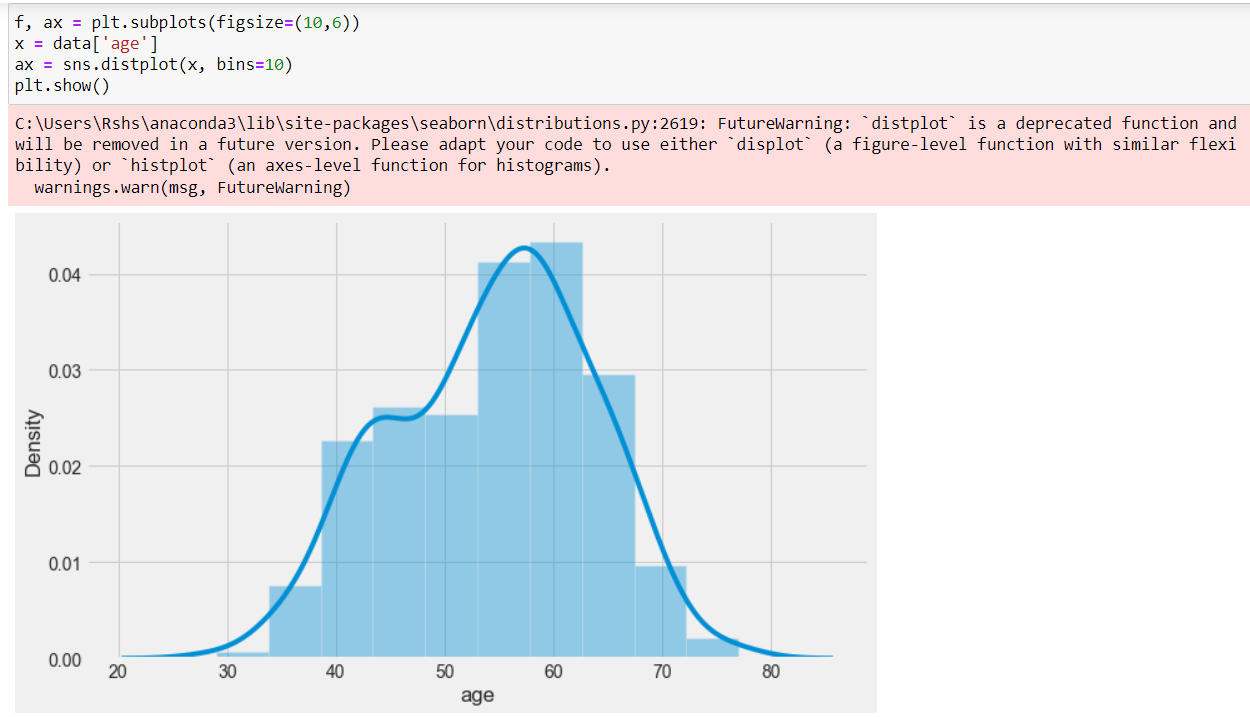


View statistical summary of age variable



* The mean value of the age variable is 54.37 years.
* The minimum and maximum values of age are 29 and 77 years.

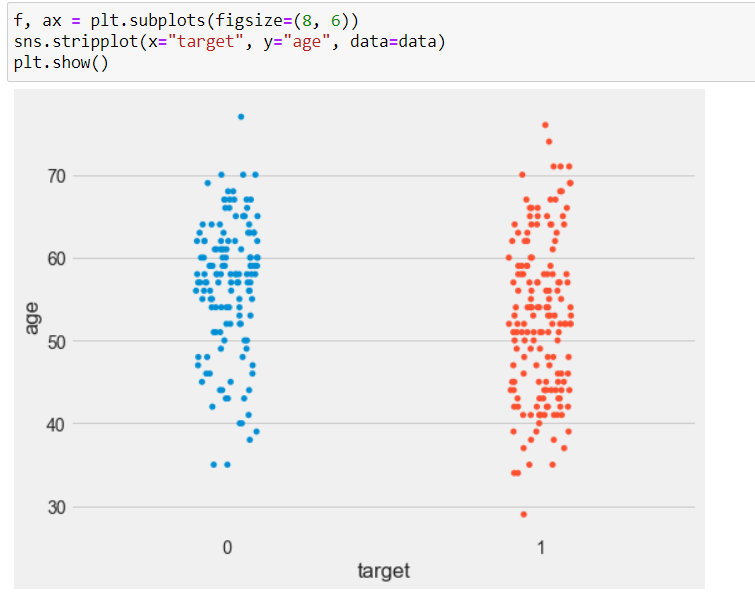
**Plot the distribution of age variable:**



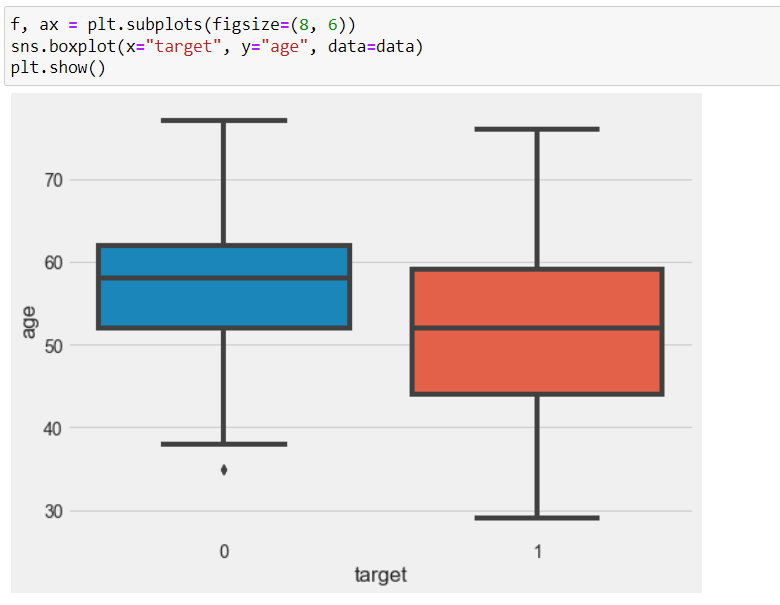
The age variable distribution is approximately normal.

**Analyse age and target variable:**

Visualize frequency distribution of age variable with respect to target

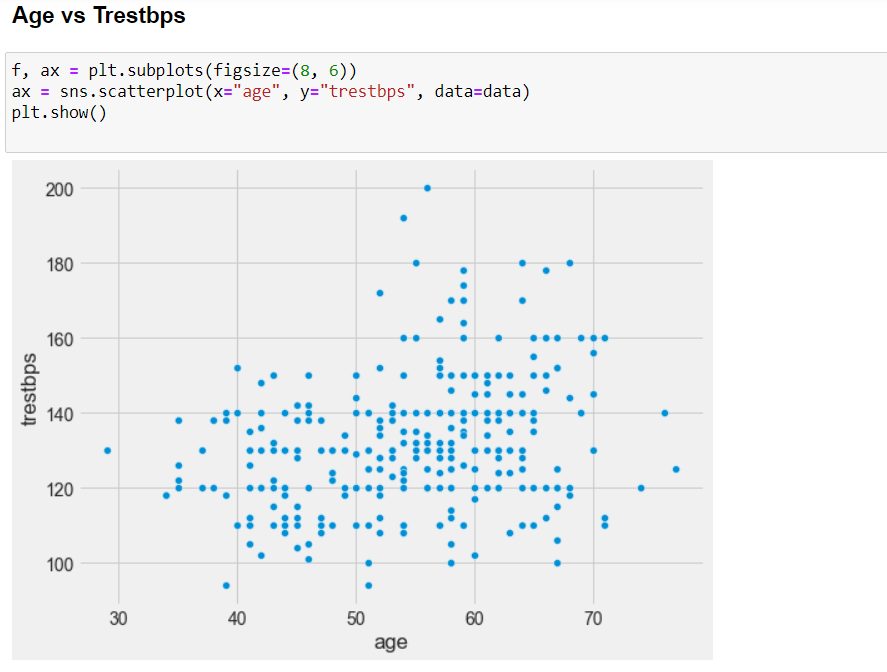


We can see that the people suffering from heart disease (target = 1) and people who are not suffering from heart disease (target = 0) have comparable ages.



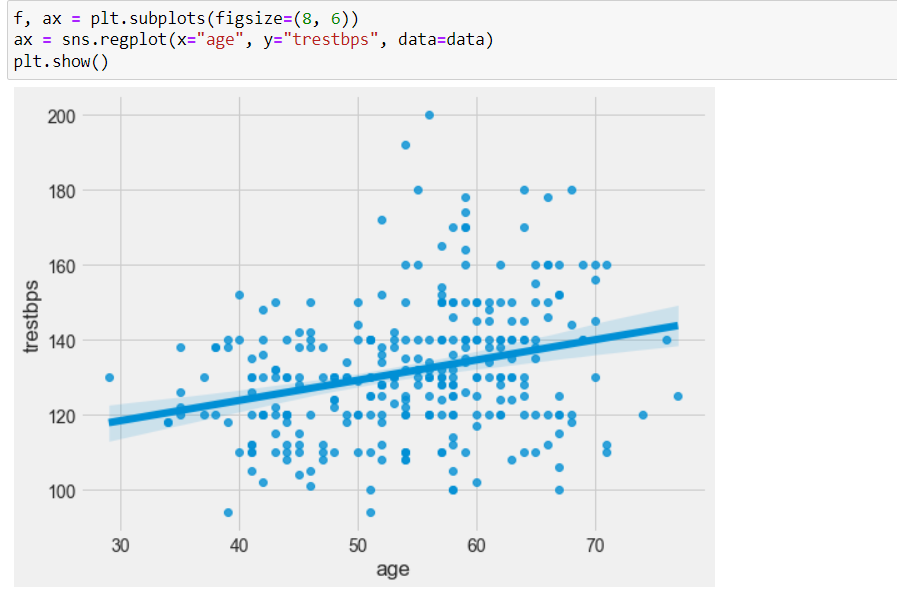
* The mean age of the people who have heart disease is less than the mean age of the people who do not have heart disease.
* The dispersion or spread of age of the people who have heart disease is greater than the dispersion or spread of age of the people who do not have heart disease

**Analyze age and trestbps variable:**

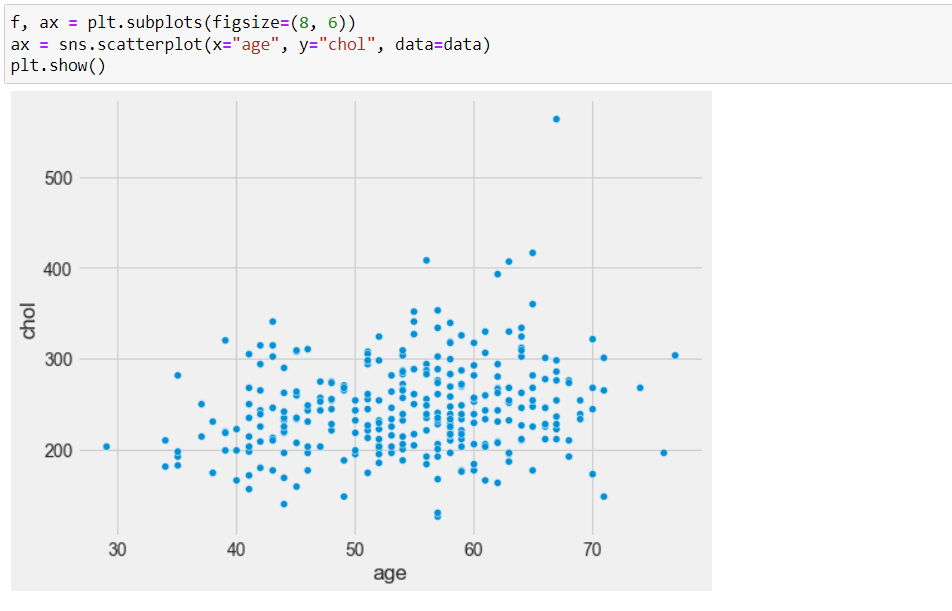


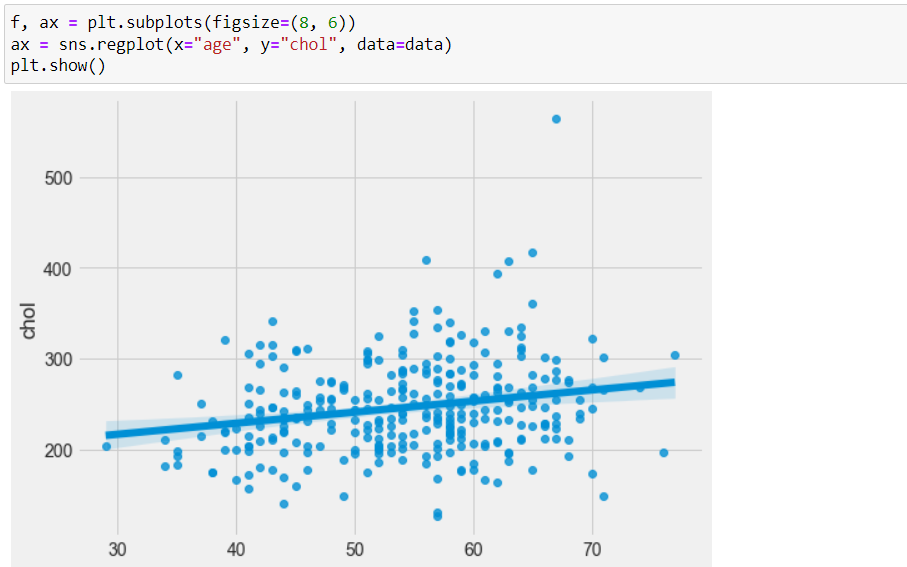
The above scatter plot shows that there is no correlation between age and trestbps variable.

On further visualisation we can see that the linear regression model is not good fit to the data.



**Analyze age and chol variable:**

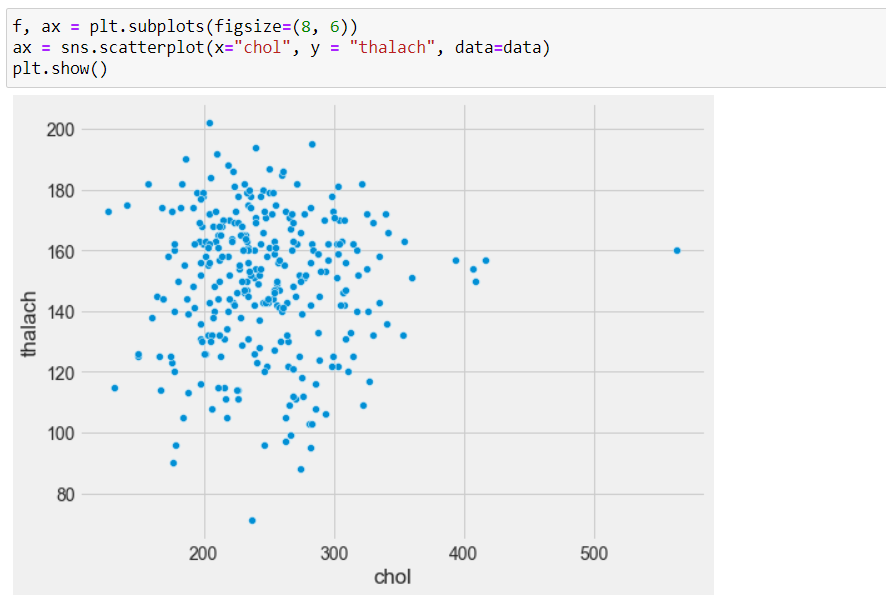




Similarly, The above scatter plot shows that there is no correlation between age and chol variable.

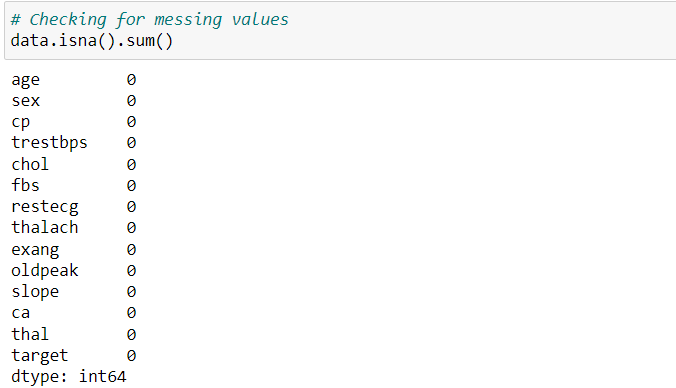
On further visualisation we can see that the linear regression model is not good fit to the data.

**Analyze chol and thalach variable:**



The above plot shows that there is no correlation between chol and thalach variable.

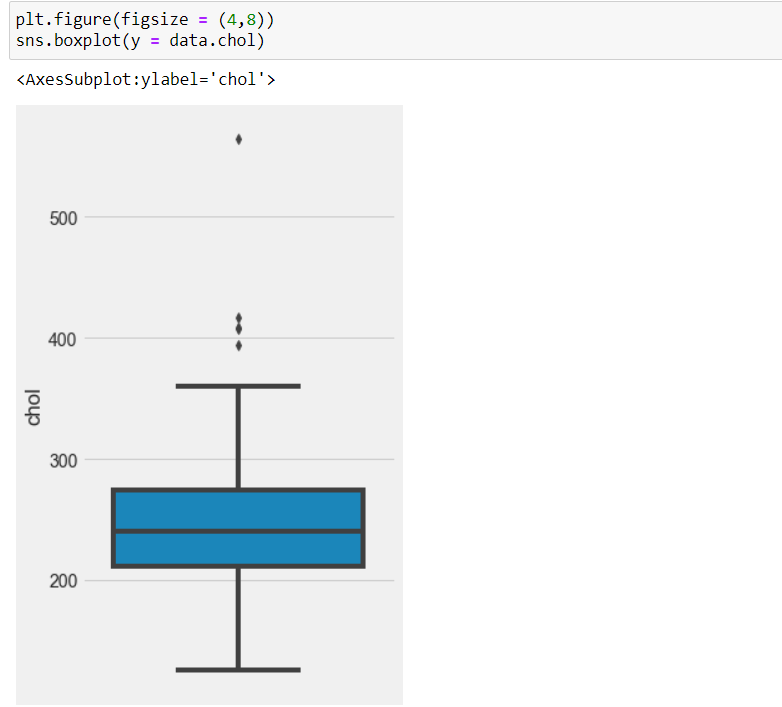
**Check for missing values in the dataset:**



We can see that there are no missing values in the dataset.

**Outlier detection:**

Examine chol variable for outliers by plotting a box plot

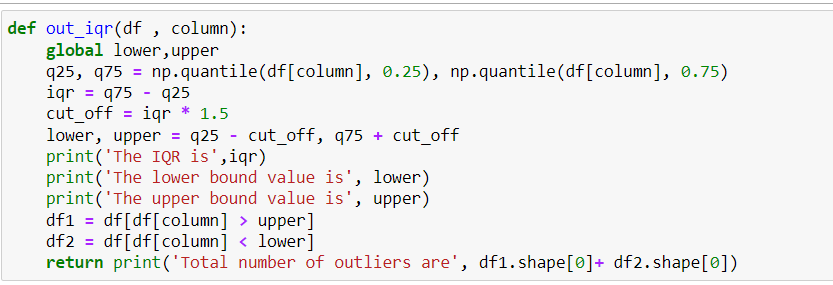


From the above box plot, we can surely observe that there are outliers in it.

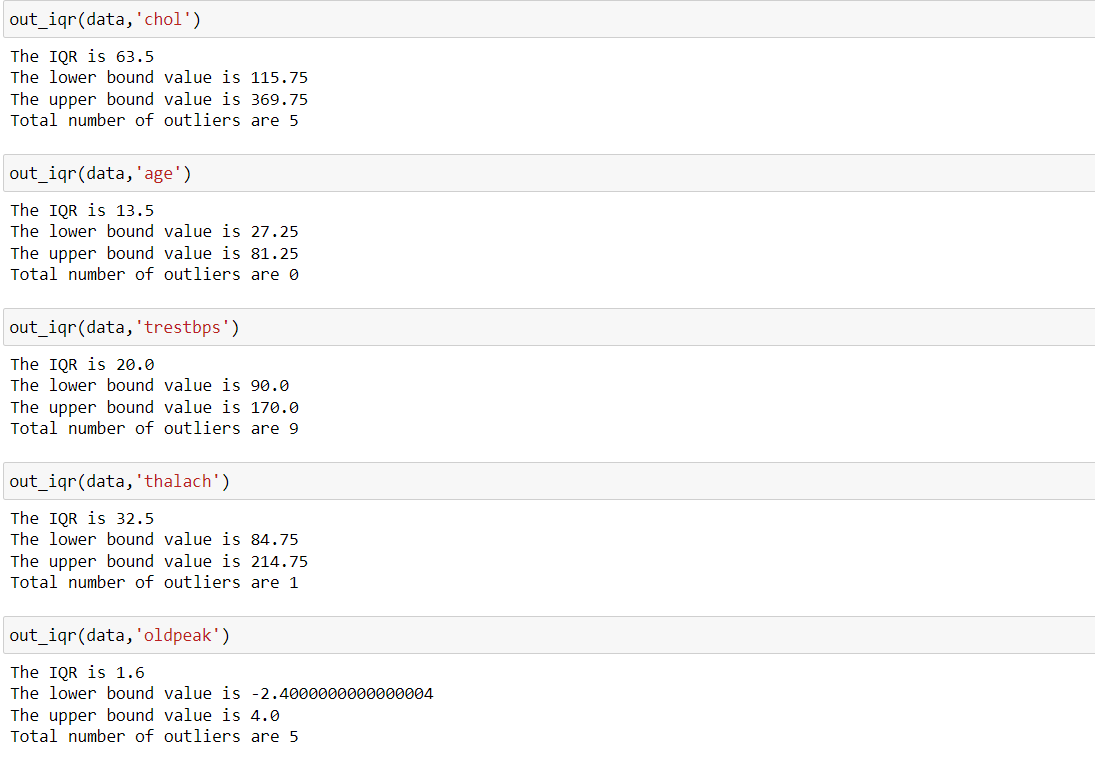
**Detect outliers using Interquartile Range Method:**

IQR is a concept in statistics that is used to measure the statistical dispersion and data variability by dividing the dataset into quartiles.

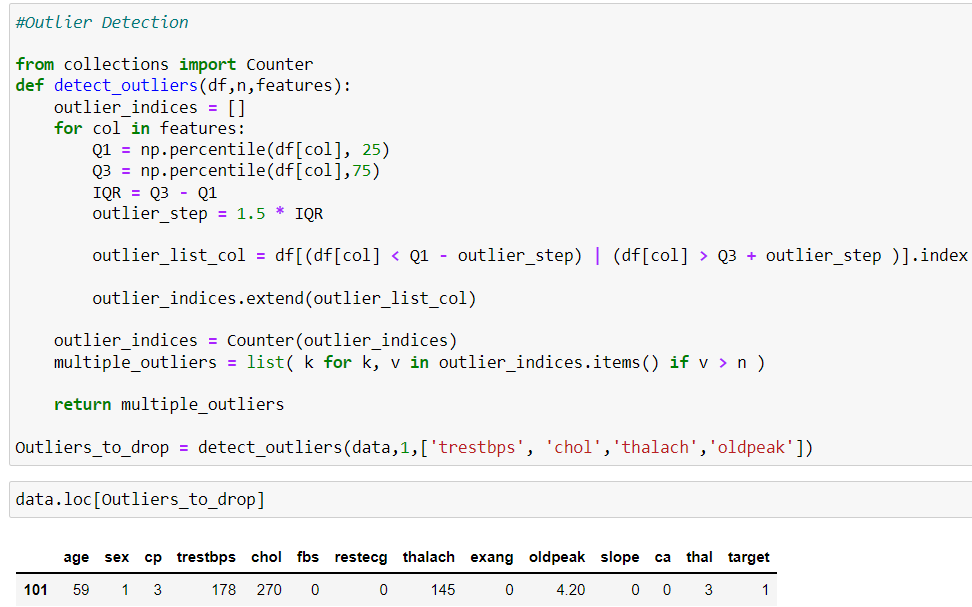
It is the difference between the third quartile and the first quartile (IQR = Q3 -Q1). Outliers in this case are defined as the observations that are below (Q1 − 1.5x IQR) or boxplot lower whisker or above (Q3 + 1.5x IQR) or boxplot upper whisker. It can be visually represented by the box plot.



By using this function, we can detect the outliers in the dataset



From the above observations its clear that this dataset contains outliers in it,

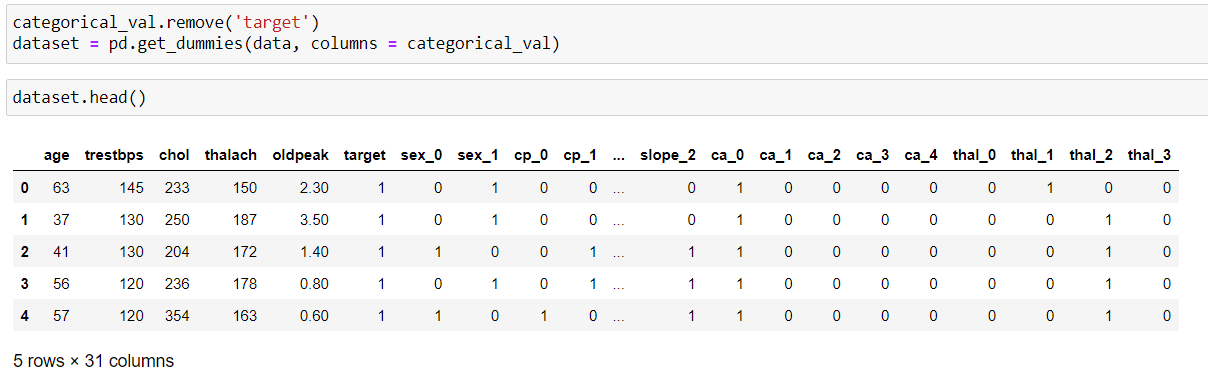


As the values of all the columns are almost in ranges. The presence of outliers doesn’t affect the dataset.

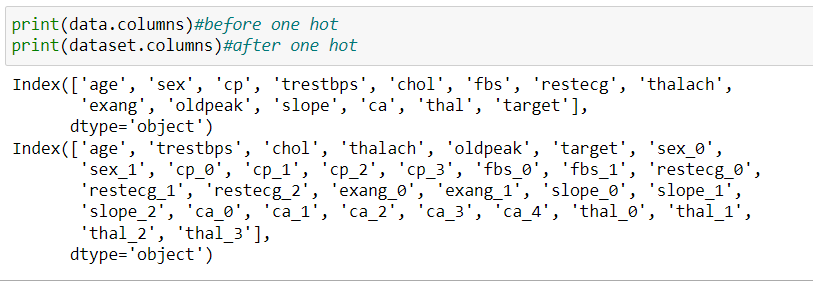
**Data Pre-processing:**

After exploring the dataset, I observed that I need to convert some categorical variables into dummy variables and scale all the values before training the Machine Learning models.

1. Remove Categorical variables using one hot encoding method:



The columns in the dataset before and after one hot encoding,

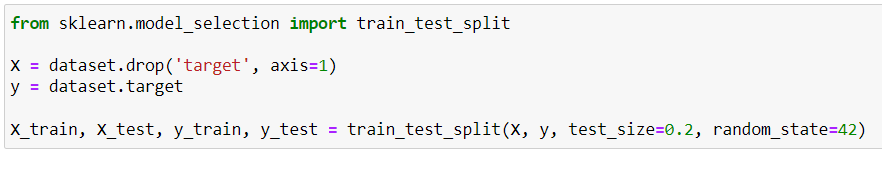


1. After handling categorical variables scale all the values using standard scaler.

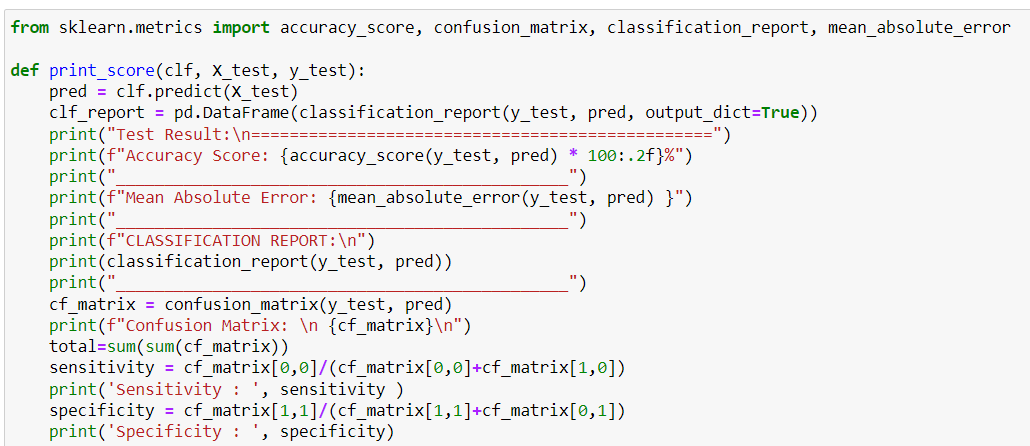


This is the preview of dataset after scaling.

1. After Scaling split the dataset into training and testing sets to train and evaluate the models.



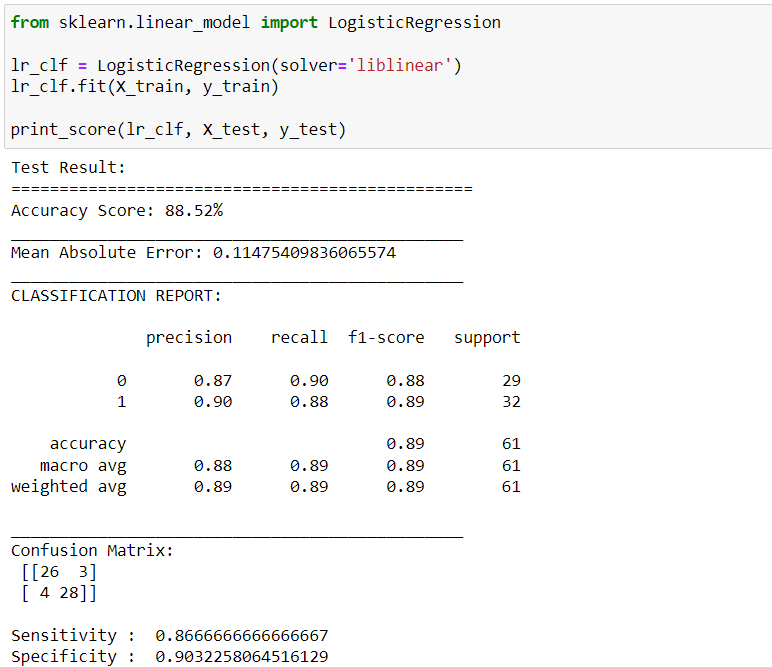
**Building ML Models:**



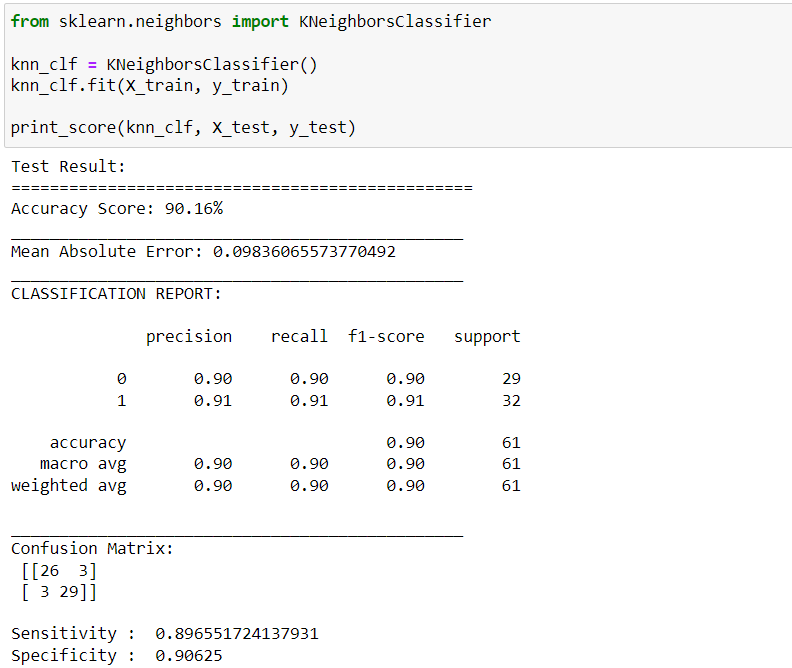
Models to be trained,

* Logistic Regression
* K-Nearest Neighbours Classifier
* Support Vector machine
* Decision Tree Classifier
* Random Forest Classifier
* XGBoost Classifier

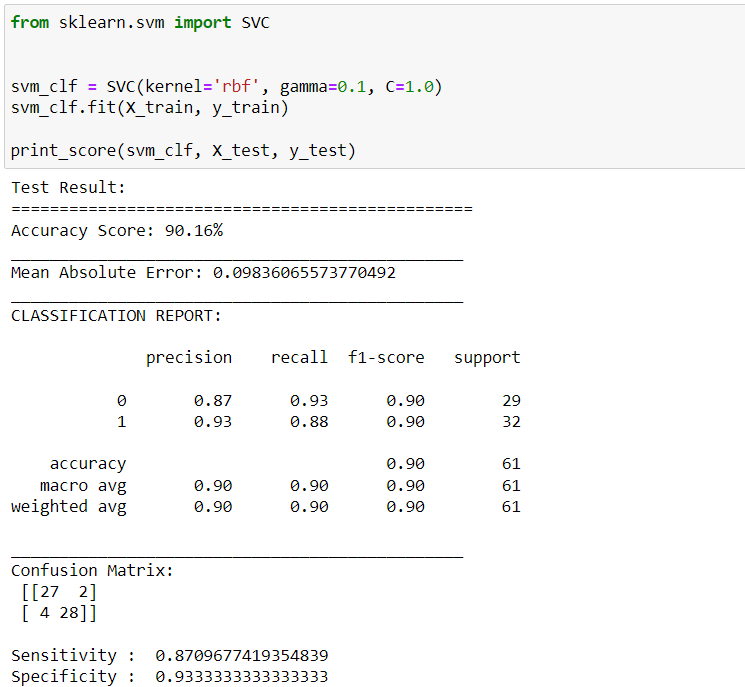
**Logistic Regression:**



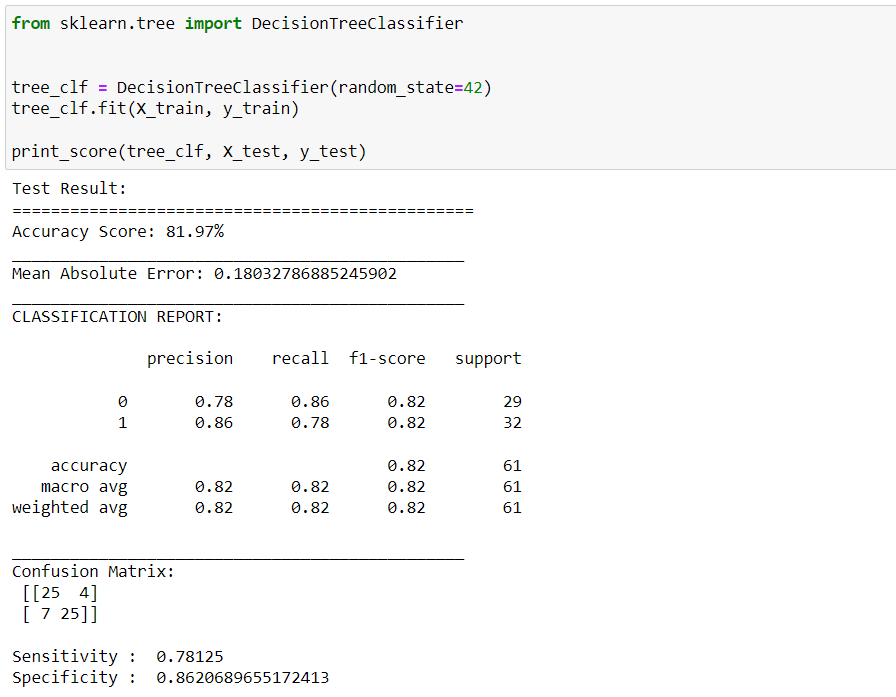
**K-Nearest Neighbours Classifier:**



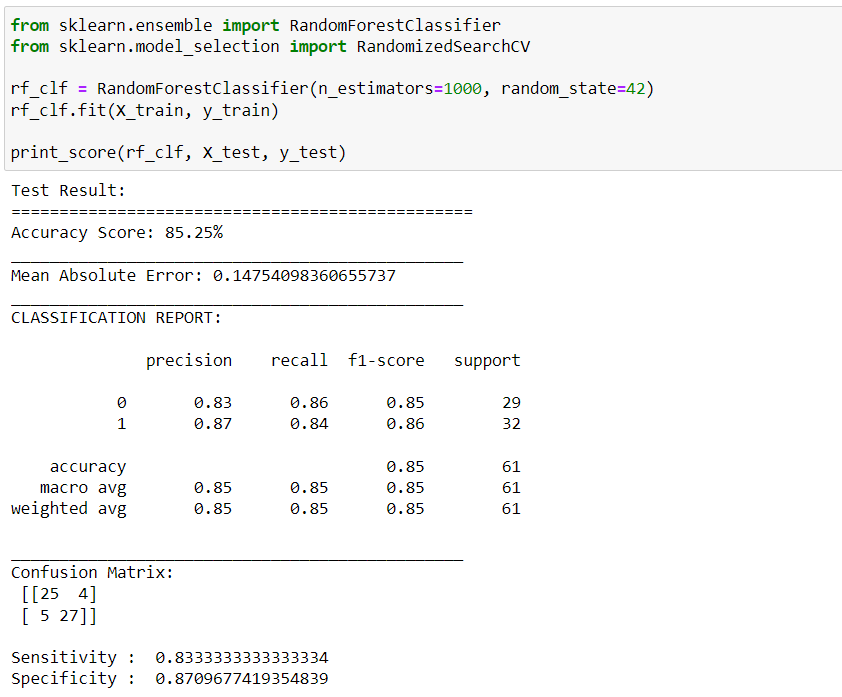
**Support Vector Machine:**



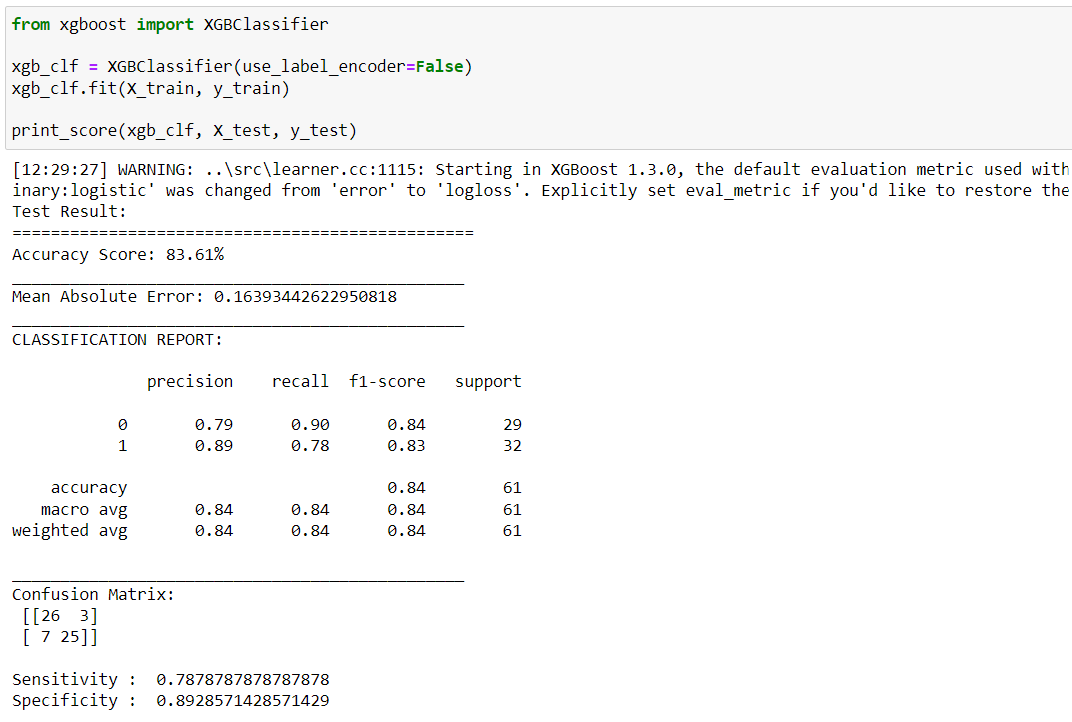
**Decision Tree Classifier:**



**Random Forest Classifier:**



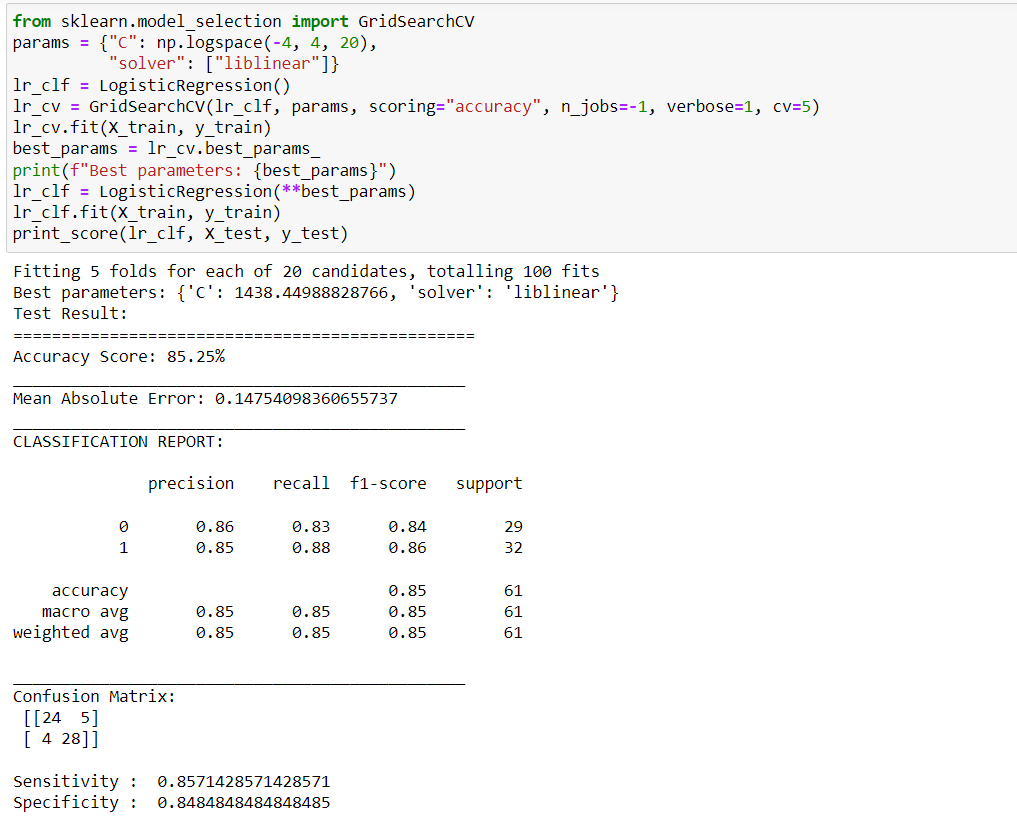
**XGBoost Classifier:**



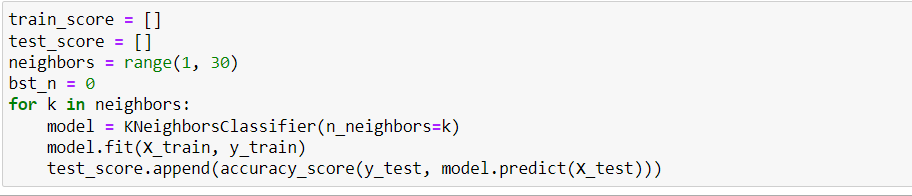
**Hyperparameter Tuning:**

The optimum hyperparameters can be found using GridSearchCV. In that machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values.

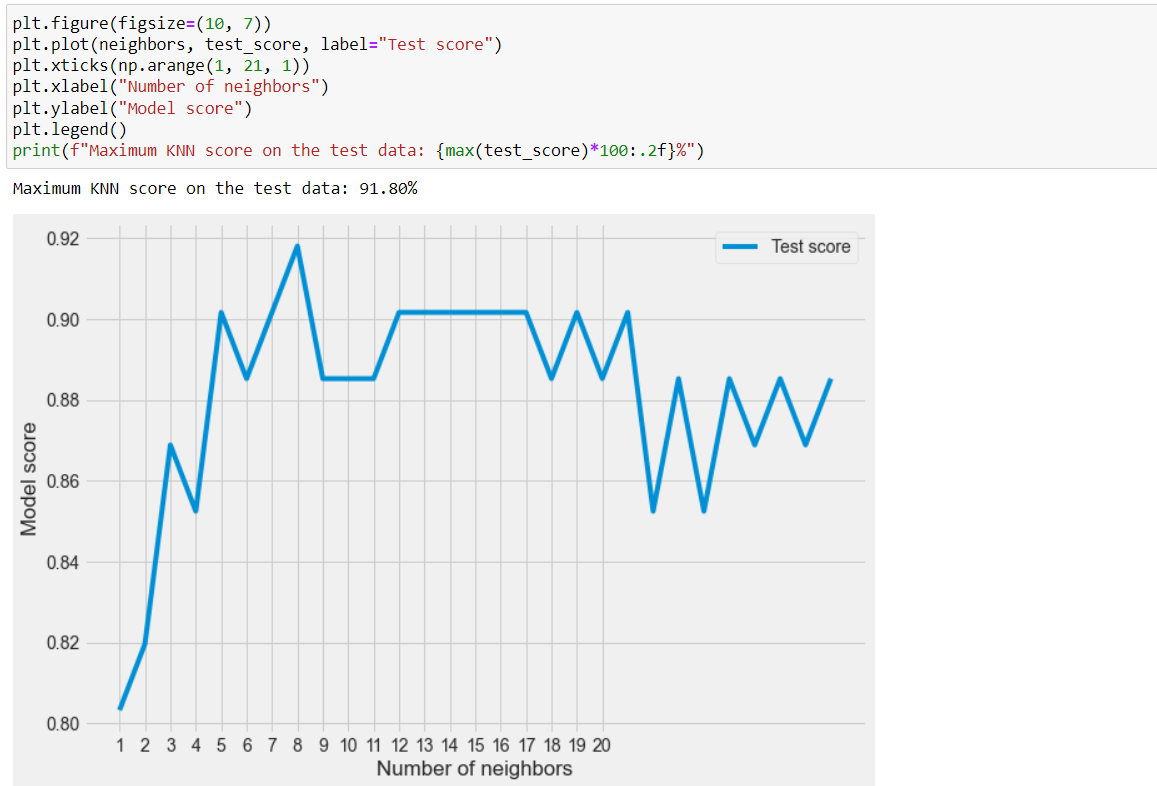
**Tuned Logistic Regression**



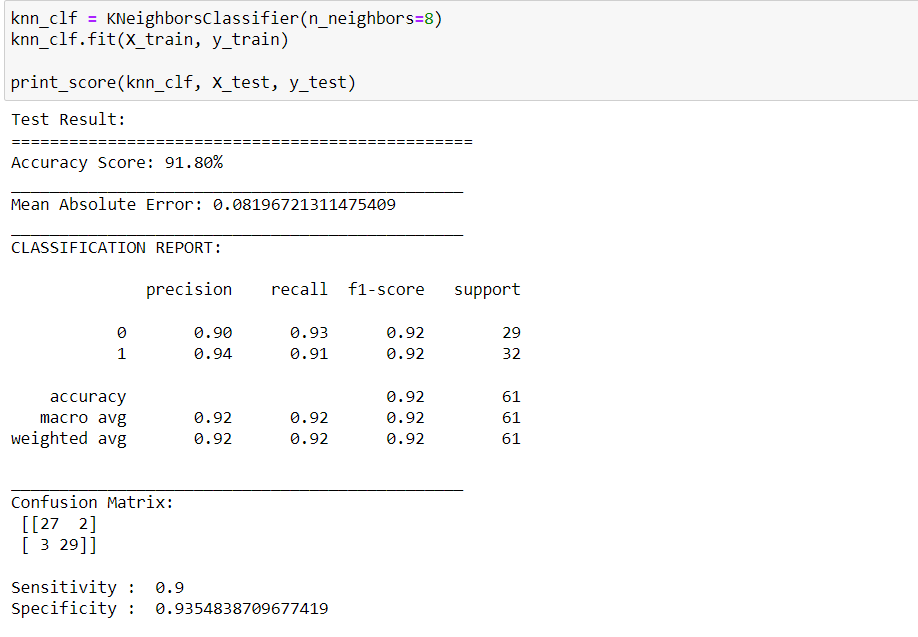
**Tuned KNN Classifier**



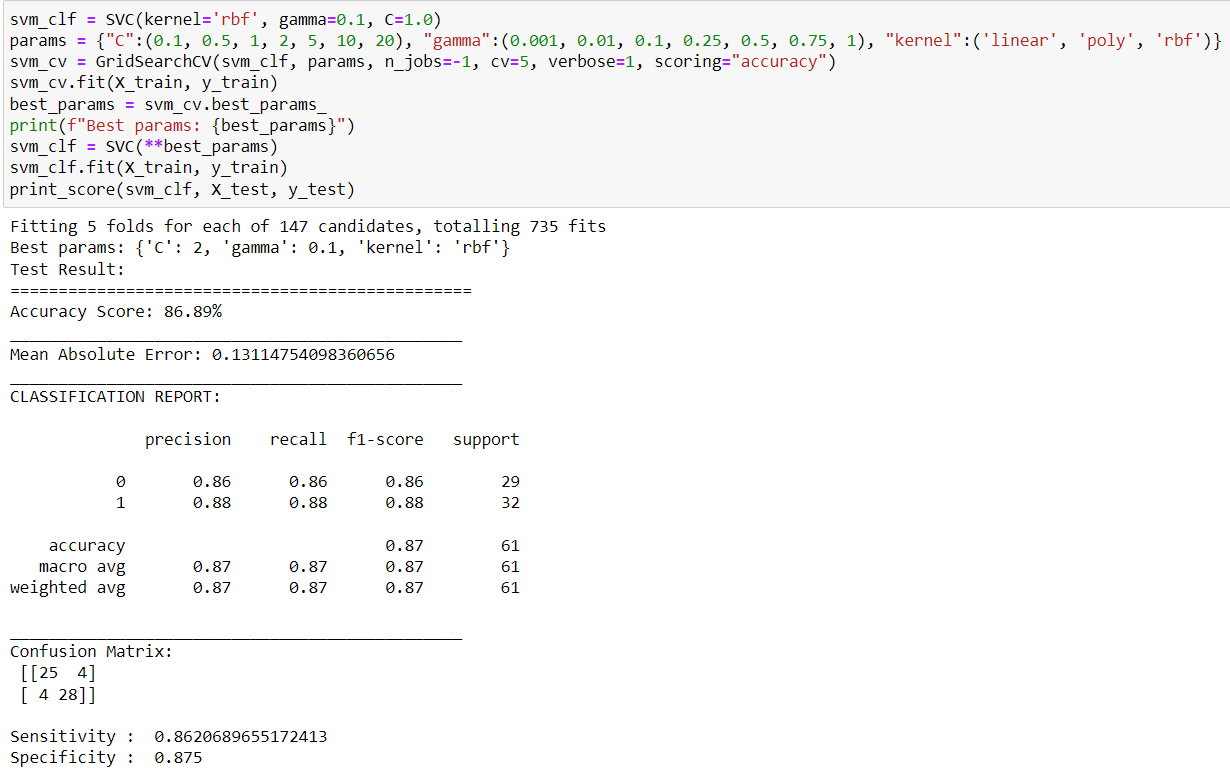
To find the best neighbour, we are iterating the neighbour value over a range of 1 to 30 and find the accuracy score. Plot the scores of all the neighbours and identify the best neighbour.



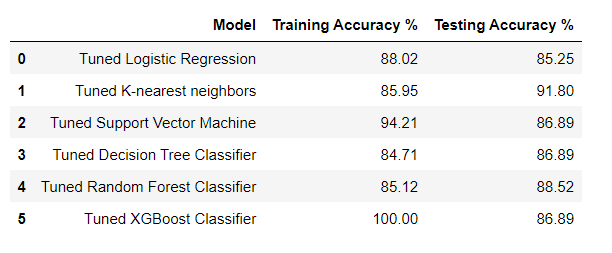
From the graph we can identify that 8 neighbour has the highest accuracy score among all the others.



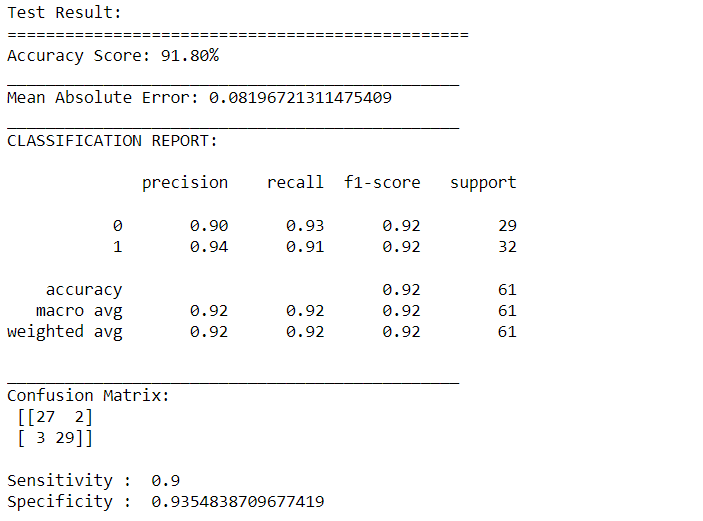
**Tuned Support Vector Machine classifier:**



Similarly, by tuning Decision Tree, Random Forest and XGBoost Classifiers we get the respective accuracy scores.



Comparing all the accuracy scores we can see that **KNN classifier** has the highest accuracy score.

On Examining the report and confusion matrix of KNN model, 

We can see the precision and recall scores and f1 scores. For this heart disease prediction model the recall score should be higher and it’s important than precision score since the model should correctly predict the person who dosen’t have heart disease, if the person actually have heart disease if the model predicts the person dosen’t have heart disease then it’s not good and it can become worse. So, **The Recall score should be higher in this model**.

**Check for Overfitting or underfitting of models:**

It’s important to understand prediction errors (bias and variance) and it can be done by using bias variance tradeoff

**Bias:**

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

**Variance:**

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.

It’s important to know they are a trade-off between these two concepts, and the goal is to balance

The Total error is calculated as,

Total Error = Bias2 + Variance+ Irreducible Error

Bias and Variance of the models **before** parameter tuning are as follows,

|  |  |  |  |
| --- | --- | --- | --- |
| S.no | Model | Bias | Variance |
| 1 | Logistic Regression | 0.096 | 0.039 |
| 2 | KNN classifier | 0.076 | 0.056 |
| 3 | Support Vector Machine | 0.102 | 0.036 |
| 4 | Decision Tree | 0.112 | 0.127 |
| 5 | Random Forest Classifier | 0.113 | 0.032 |
| 6 | XGBoost | 0.109 | 0.051 |

**Findings:**

* We can see that Logistic regression, KNN, Support Vector Machine, Random Forest and XGBoost models have high bias values and variance is not high so they are underfitted. So, these models are aiming off the target, but being consistent.
* These undefitted models bias can be increased by adding more features to the model and a balance can be achieved between bias and variance.
* In the Decision Tree model, we can see that both the bias and variance have high values, here this model is aiming off the target and being inconsistent. Its not a good choice to choose this model.

Bias and Variance of the models **after** parameter tuning are as follows,

|  |  |  |  |
| --- | --- | --- | --- |
| S.no | Model | Bias | Variance |
| 1 | Logistic Regression | 0.106 | 0.056 |
| 2 | KNN classifier | 0.087 | 0.048 |
| 3 | Support Vector Machine | 0.112 | 0.039 |
| 4 | Decision Tree | 0.093 | 0.082 |
| 5 | Random Forest Classifier | 0.106 | 0.028 |
| 6 | XGBoost | 0.108 | 0.049 |

**Findings:**

* We can see that Logistic Regression, Support Vector Machine, Random Forest and XGBoost models have high bias and the variance is not high. So, these models are aiming off the target, but being consistent. Thus, they are underfitted.
* These undefitted models bias can be increased by adding more features to the model and a balance can be achieved between bias and variance.
* In the Decision Tree model, we can see that both the bias and variance have high values, here this model is aiming off the target and being inconsistent. Its not a good choice to choose this model.
* In KNN classifier the bias value is high(but not much higher) and the variance is not much higher, here its not a best model but it performs well compared to other models. The high bias can be reduced by feeding more features and data to the model.