**Predicting Heart Disease**

**Objective:**

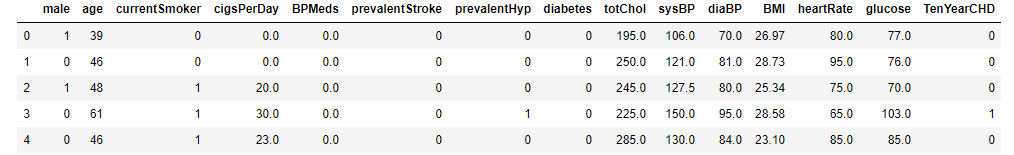
World Health Organization has estimated 12 million deaths occur worldwide: every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardiovascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression.

Data Overview:

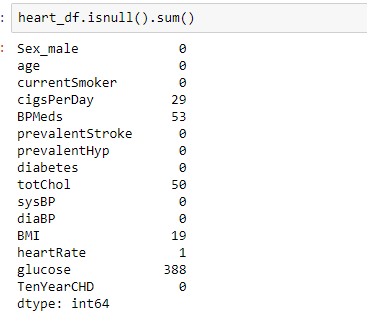
The dataset is publicly available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients’ information. It includes over 4,000 records and 15 attributes.

**Data Read:**

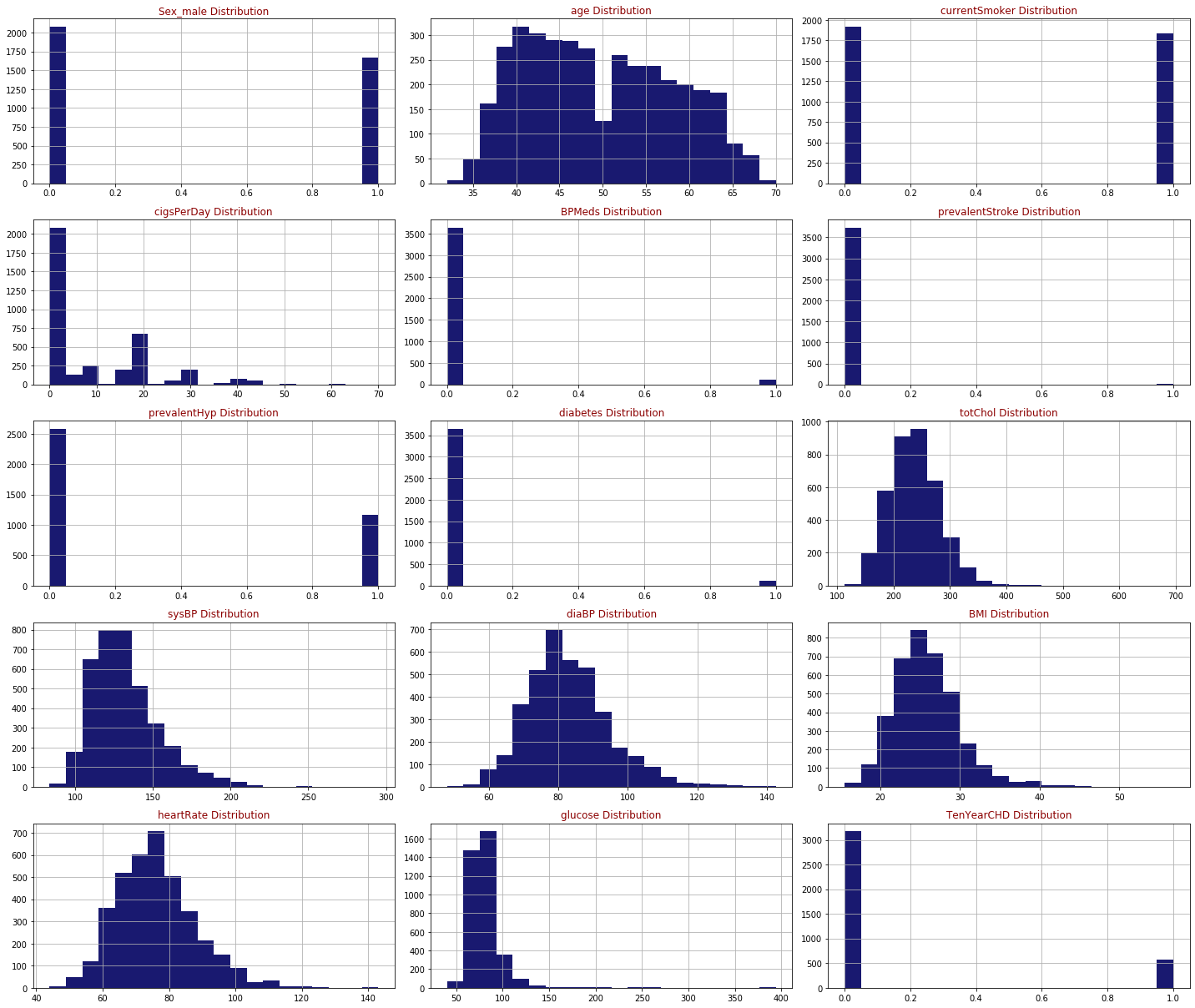
Using Pandas csv library reading the locally stored data



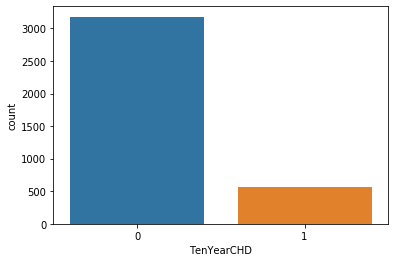
Check the number of Null values



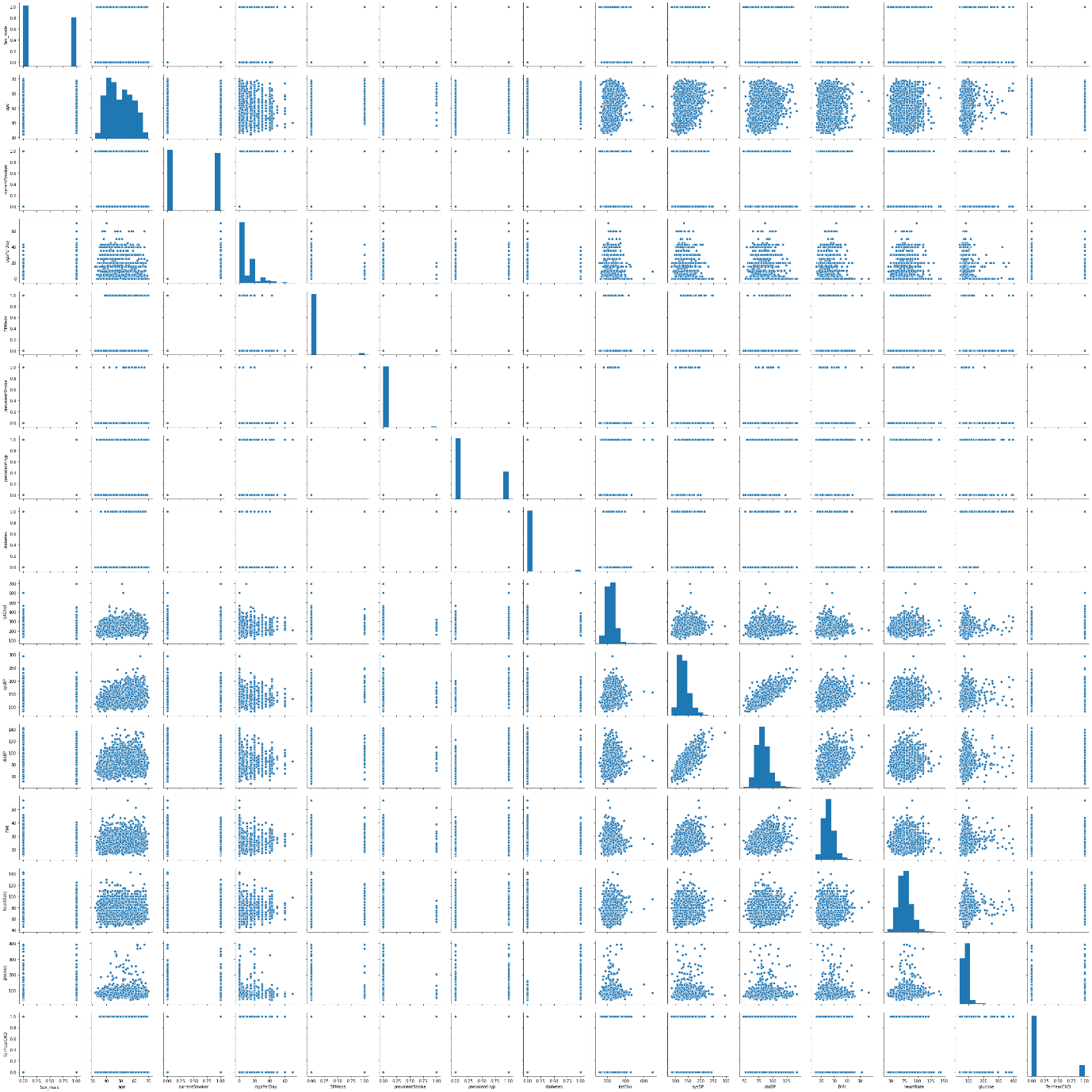
**Exploratory Data Analysis:**



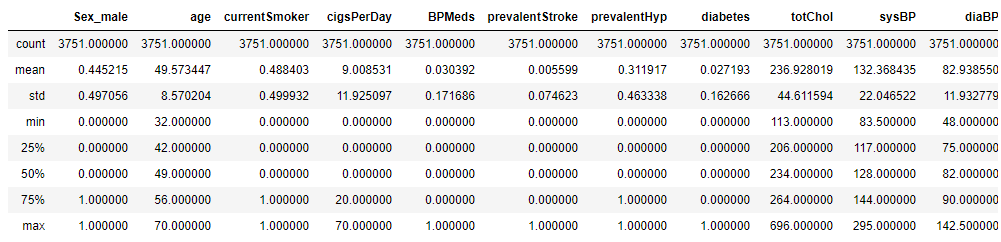
**Check the Male and Female Count**



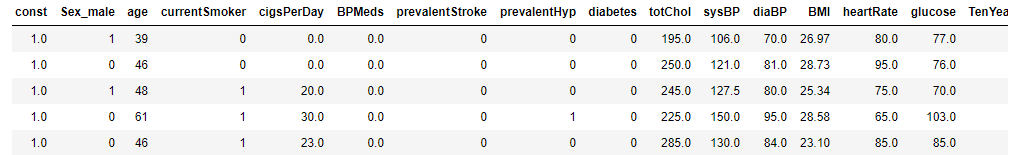
**Generating Pair Plots**



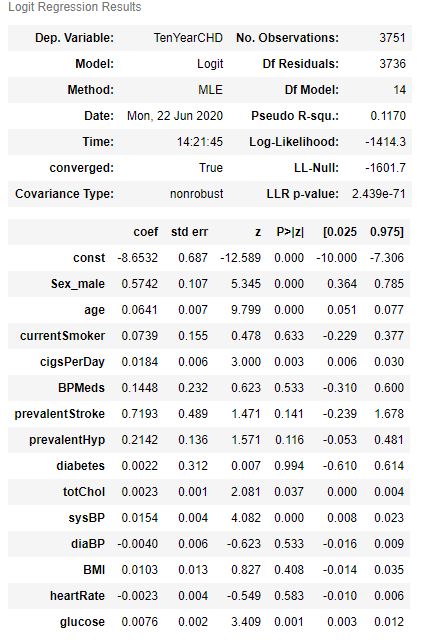
**Describing the Data**



Based on the data, applying Logistic Regression for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. In logistic regression the dependent variable is always binary. Logistic regression is mainly used to for prediction and also calculating the probability of success.

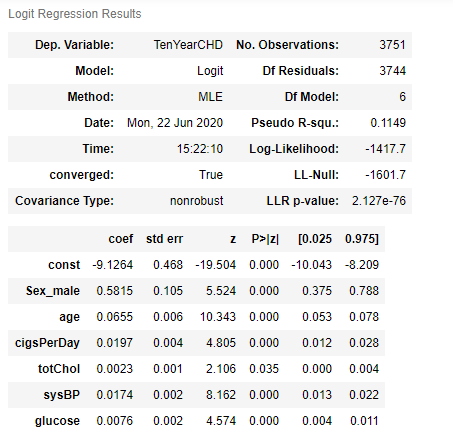


**Logistic Regression Results:**

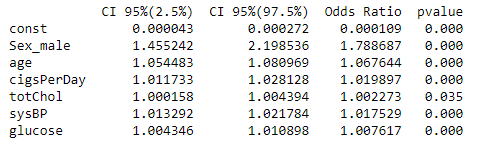


The results above show some of the attributes with P value higher than the preferred alpha (5%) and thereby showing low statistically significant relationship with the probability of heart disease. Backward elimination approach is used here to remove those attributes with highest Pvalue one at a time followed by running the regression repeatedly until all attributes have P Values less than 0.05.

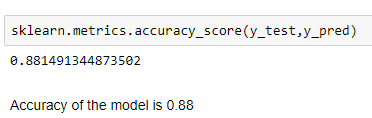
**Feature Selection: Backward elimination (P-value approach)**



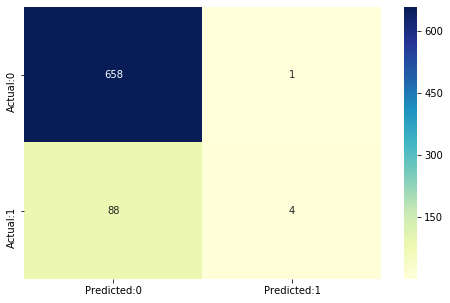
Interpreting the results: Odds Ratio, Confidence Intervals and Pvalues



**Model Evaluation**



Generate Heatmap on the predicted model:



The confusion matrix shows 658+4 = 662 correct predictions and 88+1= 89 incorrect ones.

True Positives: 4

True Negatives: 658

False Positives: 1 (Type I error)

False Negatives: 88 (Type II error)

**Model Evaluation – Statistics**

The acuuracy of the model = TP+TN/(TP+TN+FP+FN) = 0.881491344873502

The Missclassification = 1-Accuracy = 0.118508655126498

Sensitivity or True Positive Rate = TP/(TP+FN) = 0.043478260869565216

Specificity or True Negative Rate = TN/(TN+FP) = 0.9984825493171472

Positive Predictive value = TP/(TP+FP) = 0.8

Negative predictive Value = TN/(TN+FN) = 0.8820375335120644

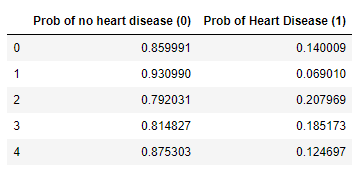
Positive Likelihood Ratio = Sensitivity/(1-Specificity) = 28.65217391304351

Negative likelihood Ratio = (1-Sensitivity)/Specificity = 0.9579754195850403

From the above statistics the model is highly specific than sensitive. The negative values are predicted more accurately than the positives.

Predicted probabilities of 0 (No Coronary Heart Disease) and 1 (Coronary Heart Disease: Yes) for the test data with a default classification threshold of 0.5

**Predicting Heart Disease between Male and Female**



Lower the threshold Since the model is predicting Heart disease too many type II errors is not advisable. A False Negative (ignoring the probability of disease when there actually is one) is more dangerous than a False Positive in this case. Hence to increase the sensitivity, threshold can be lowered.

With 0.1 threshold the Confusion Matrix is

[[240 419]

[ 11 81]]

with 321 correct predictions and 11 Type II errors(False Negatives)

Sensitivity: 0.8804347826086957 Specificity: 0.36418816388467373

With 0.2 threshold the Confusion Matrix is

[[519 140]

[ 43 49]]

with 568 correct predictions and 43 Type II errors(False Negatives)

Sensitivity: 0.532608695652174 Specificity: 0.787556904400607

With 0.3 threshold the Confusion Matrix is

[[617 42]

[ 70 22]]

with 639 correct predictions and 70 Type II errors( False Negatives)

Sensitivity: 0.2391304347826087 Specificity: 0.936267071320182

With 0.4 threshold the Confusion Matrix is

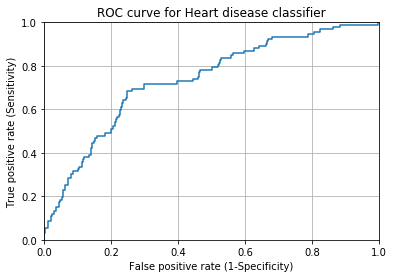
[[652 7]

[ 86 6]]

with 658 correct predictions and 86 Type II errors (False Negatives)

Sensitivity: 0.06521739130434782 Specificity: 0.9893778452200304

**ROC Curve**

****

A common way to visualize the trade-offs of different thresholds is by using an ROC curve, a plot of the true positive rate (# true positives/ total # positives) versus the false positive rate (# false positives / total # negatives) for all possible choices of thresholds. A model with good classification accuracy should have significantly more true positives than false positives at all thresholds.

The optimum position for roc curve is towards the top left corner where the specificity and sensitivity are at optimum levels

Area Under The Curve (AUC) The area under the ROC curve quantifies model classification accuracy; the higher the area, the greater the disparity between true and false positives, and the stronger the model in classifying members of the training dataset. An area of 0.5 corresponds to a model that performs no better than random classification and a good classifier stays as far away from that as possible. An area of 1 is ideal. The closer the AUC to 1 the better.

**ROC Curve score:**

**sklearn.metrics.roc\_auc\_score(y\_test,y\_pred\_prob\_yes[:,1])**

0.7355182423962525

* All attributes selected after the elimination process show Pvalues lower than 5% and thereby suggesting significant role in the Heart disease prediction.
* Men seem to be more susceptible to heart disease than women. Increase in Age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease.
* Total cholesterol shows no significant change in the odds of CHD. This could be due to the presence of 'good cholesterol (HDL) in the total cholesterol reading. Glucose too causes a very negligible change in odds (0.2%)
* The model predicted with 0.88 accuracy. The model is more specific than sensitive.
* \**The Area under the ROC curve is 73.5 which is somewhat satisfactory. \**
* \*\* Overall model could be improved with more data. \*\*