Machine Learning Applied: Prediction of Heart Disease in individuals based on existing parameters of health

# Introduction

Heart Disease is one of the largest contributors to human deaths worldwide. It has been so for quite some time and has been the leading cause, historically. In fact, 2020 was the only year in the last decade that a new disease overtook heart disease in the number of deaths.

More than 11,800 COVID-19 deaths were reported, passing heart disease at 10,700 deaths, lung and tracheal cancer at nearly 4,000 deaths, chronic obstructive pulmonary disease at more than 3,700 deaths, and stroke at more than 3,600 deaths.

(Crist, 2020)

The reason Heart Disease causes so much damage is either late diagnosis, or faulty prognosis. Hard as that might be to digest, the stark truth is that individuals seldom realize what factors of their health attribute to the cause of heart disease.

But, truthfully, the nomenclature in itself is misleading – Heart Disease is one all encompassing term that people use to describe any medical issue related to the heart. Some of the major heart diseases include, but are not limited to, Coronary Heart Disease, High BP, Cardiac Arrest, Arrythmia and Stroke.

One of these, or a combination of these might cause the heart to either stop pumping, or become dysfunctional, leading to heart related death.

It is common for laypeople to simply look at only the external factors that might have been cause of the heart disease. Therefore, people may look at only common remedial steps to avoid heart disease – such as regular exercise, to avoid obesity and hypertension, healthy diets, and the avoidance of alcohol and or smoking. Unfortunately, the prognosis for Heart related diseases, is a little more complicated than that, and involves more factors than that which meets the eye.

In understanding this, I have tried to plot said factors and with the use of machine learning, tried and successfully arrived at a model that can detect the risk of heart disease in individuals bases these factors.

# Context of the project

Cleveland Database is a medical database that has been used my ML researchers globally to sort and figure out this problem of diagnosing heart disease in individuals. Using this data, I have attempted to

1. Import the database
2. Clean the data
3. Analyze the data
4. Standardize the data
5. Draw up the potential ML models
6. Shortlist / Finalize ML model
7. Hyper-tune the ML model
8. Predict the output
9. Save the ML model

This project is an attempt to realize a solution for the easy prediction of heart related diseases in individuals using the data available. Once the optimal ML model has been created / chosen, it can be deployed as a real-time prediction device for heart diseases.

# Problem definition

The Cleveland Database is the experimental set used for this purpose by ML researchers worldwide – but the data has several gaps that impede the creation of optimal ML models. Proper analysis of the data is required to create optimal ML model for this purpose.

Data Set Information:

This database contains of 14 attributes. In particular, the Cleveland database, is the only one that has been extensively used for this concern by ML researchers, to this date. The "goal" field in the data set refers to the presence of heart disease in the patient. It is integer-valued from 0 (no presence) to 4 (high-risk). My experiment with the Cleveland database is concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

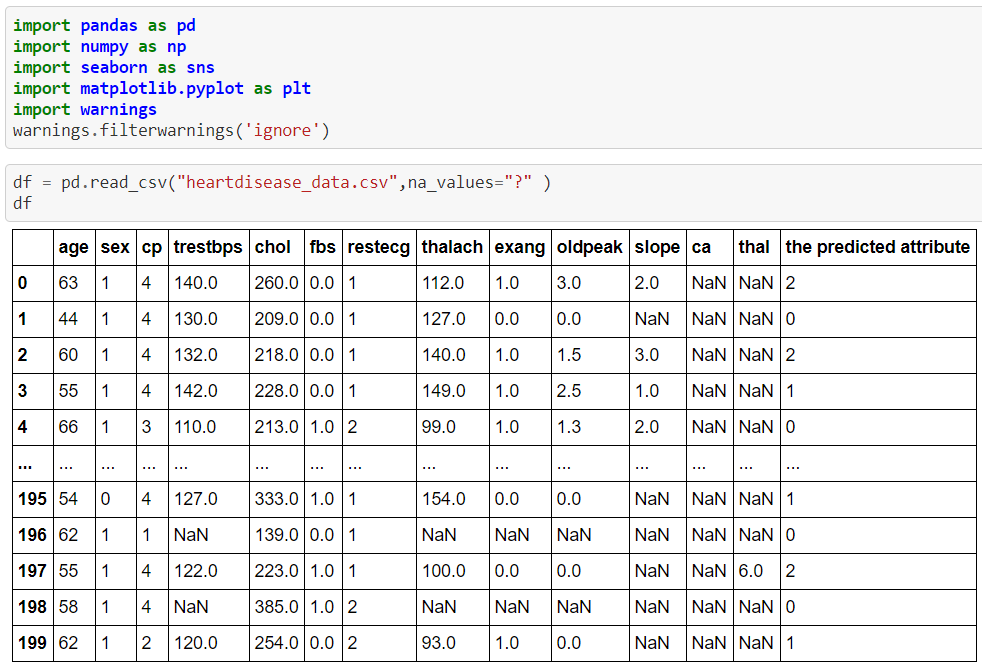
Attributes used:

1. age
2. sex
3. cp- chest pain type (4 values)
4. trestbps - resting blood pressure
5. chol - serum cholesterol in mg/dl
6. fbs - fasting blood sugar > 120 mg/dl
7. restecg - resting electrocardiographic results (values 0,1,2)
8. thalach - maximum heart rate achieved
9. exang - exercise induced angina
10. oldpeak - ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. ca - number of major vessels (0-3) colored by fluoroscopy
13. thal - 0 = normal; 1 = fixed defect; 2 = reversible defect

Note: You can find the dataset in the link below.

<https://github.com/dsrscientist/dataset1/blob/master/heartdisease_data.csv>

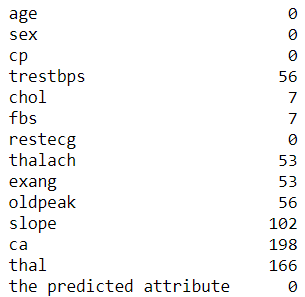
# Data Analysis

I have started my analysis of the data by importing the required libraries:  


**The database has 200 rows X 14 columns**

Based on initial scouting, there are several values missing in the dataset. To circumvent this, I have attempted to identify the missing data and fill in the gaps using the other data available.

Initial Analysis of the data shows the below numbers of data are missing:



Evidently, there is a lot of data missing. In order to fill in for the missing data, I have attempted the following steps:

1. Find the unique value in each column of the data set

the unique value of **age** is

[63 44 60 55 66 65 56 59 62 57 46 58 64 74 52 69 51 54 77 61 40 41 42 53

68 67 72 75 49 35 43 48 50 45 76 70 71 38 37]

the unique value of **sex** is

[1 0]

the unique value of **cp** is

[4 3 2 1]

the unique value of **trestbps** is

[140. 130. 132. 142. 110. 120. 150. 180. 160. 126. nan 128. 170. 152.

116. 124. 0. 122. 144. 154. 125. 104. 136. 134. 138. 178. 146. 135.

158. 106. 112. 102. 96. 172. 155. 156. 118. 100. 190. 114. 127.]

the unique value of **chol** is

[260. 209. 218. 228. 213. 0. 236. 267. 166. 220. 177. 186. 100. 171.

230. 281. 203. 277. 233. 240. 153. 224. 316. 311. 270. 217. 214. 252.

339. 216. 276. 458. 241. 384. 297. 248. 308. 208. 227. 210. 245. 225.

198. 195. 161. 258. 235. 305. 223. 282. 349. nan 160. 312. 283. 142.

211. 306. 222. 202. 197. 204. 274. 192. 298. 272. 200. 261. 181. 221.

175. 219. 310. 232. 273. 182. 292. 289. 193. 170. 369. 173. 271. 244.

285. 243. 237. 165. 287. 256. 264. 226. 207. 284. 337. 254. 300. 333.

139. 385.]

the unique value of **fbs** is

[ 0. 1. nan]

the unique value of **restecg** is

[1 2 0]

the unique value of **thalach** is

[112. 127. 140. 149. 99. 120. 105. 141. 157. 117. nan 148. 86. 84.

125. 118. 124. 106. 111. 180. 129. 110. 155. 122. 133. 131. 80. 165.

107. 128. 160. 97. 161. 130. 108. 123. 144. 102. 145. 69. 138. 150.

88. 132. 121. 135. 100. 162. 73. 154. 115. 119. 159. 94. 113. 98.

96. 151. 126. 93.]

the unique value of **exang** is

[ 1. 0. nan]

the unique value of **oldpeak** is

[ 3. 0. 1.5 2.5 1.3 -0.5 2. 0.5 1. nan 1.6 4. 3.5 0.8

1.7]

the unique value of **slope** is

[ 2. nan 3. 1.]

the unique value of **ca** is

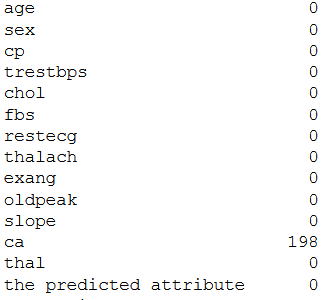
[nan 0.]

the unique value of **thal** is

[nan 3. 7. 6.]

the unique value of **the predicted attribute** is

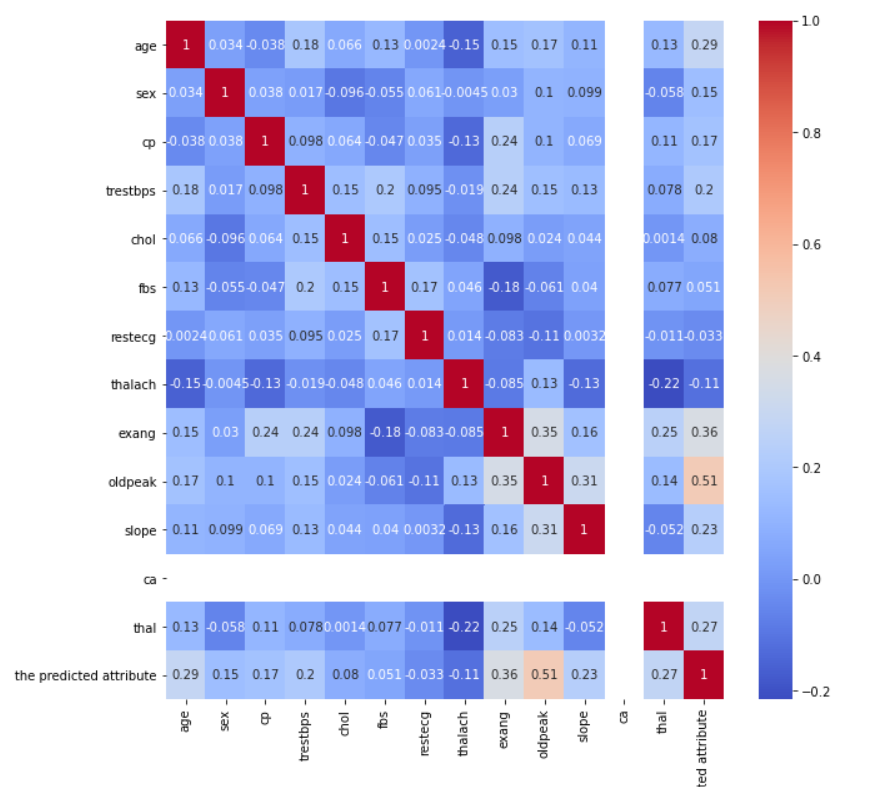
[2 0 1 3 4]

1. Once we have the unique values of each attribute, we can go ahead and break the data into smaller pieces according to the unique value to find the mean of the data in missing attribute and fill in the NaN values with mean.
2. We have filled all the NaN values but attribute **ca** has 99% of missing values.
3. So, we are deciding to drop ca to make our data predictable.



We have cleaned our data by filling the missing pieces in the data lets understand the correlation of our data

# Data Correlation.

 Correlation visualization using Heat Map.

We are seeing the data correlation, the **ca** attribute has more missing data its evident that it has no correlation, lets also see the correlation of the all the Feature variable with the Target variable.

#### The correlation with the Target Variable is:

oldpeak 0.514634

exang 0.364374

age 0.287289

thal 0.274930

slope 0.231266

trestbps 0.203937

cp 0.168210

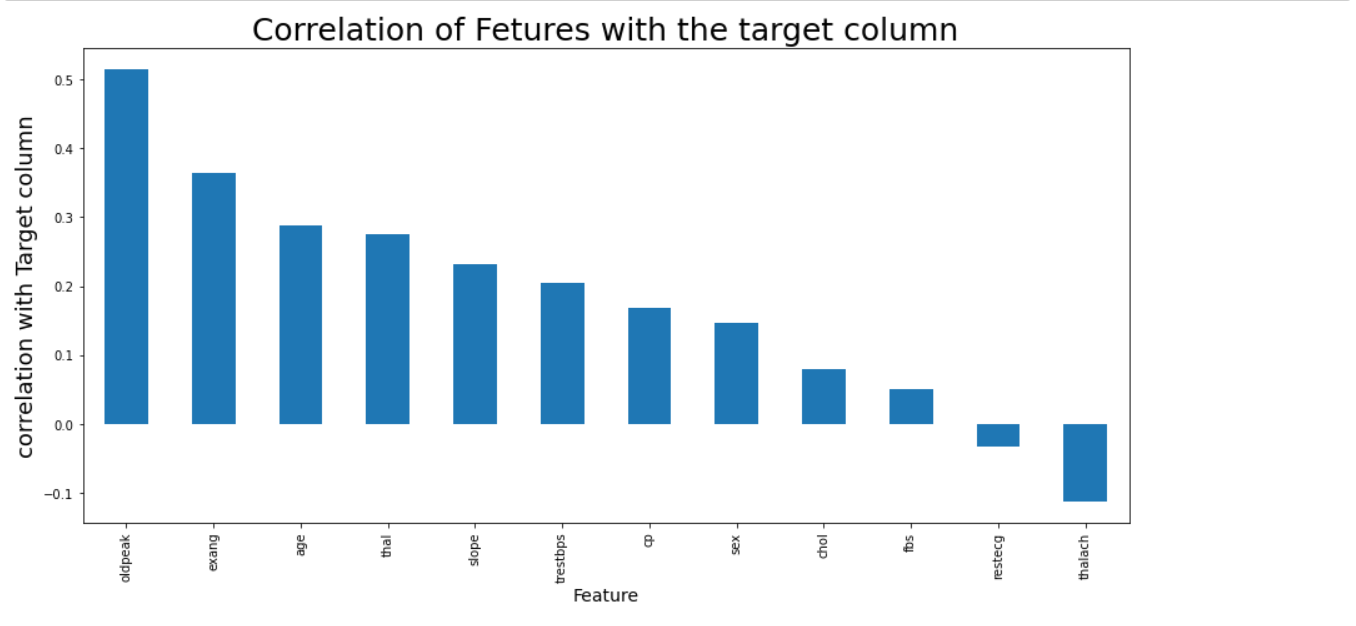
sex 0.147470

chol 0.080084

fbs 0.051016

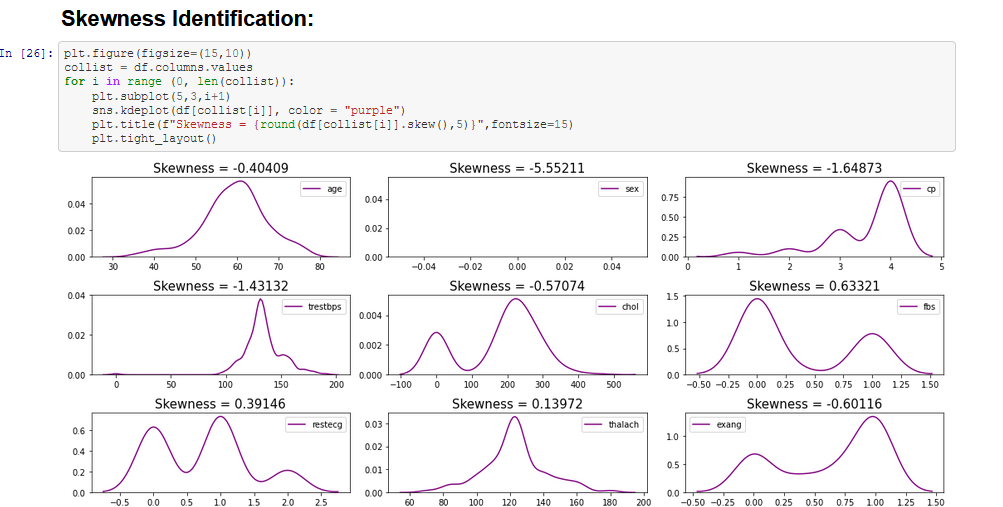
restecg -0.032800

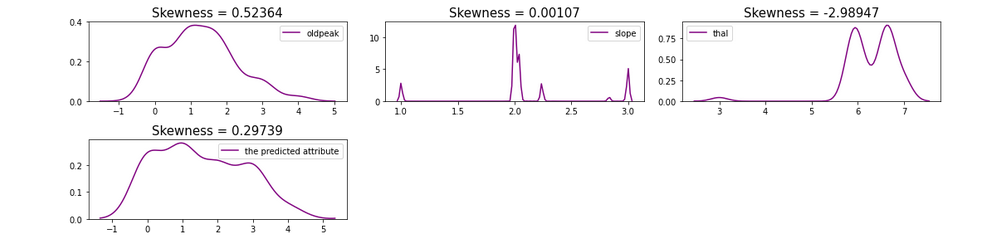
thalach -0.111939



We observe that “oldpeak” have highest correlation with Target Variable which defines that the Heart Disease are caused mostly by “oldpeak”.

# Pre-processing pipeline

 Let’s further cleanse the data by identifying the Skewness and Outliers in the data.



|  | **Feature\_names** | **Skew** |
| --- | --- | --- |
| **0** | fbs | 0.633207 |
| **1** | oldpeak | 0.523641 |
| **2** | restecg | 0.391463 |
| **3** | the predicted attribute | 0.297393 |
| **4** | thalach | 0.139718 |
| **5** | slope | 0.001068 |
| **6** | age | -0.404091 |
| **7** | chol | -0.570738 |
| **8** | exang | -0.601159 |
| **9** | trestbps | -1.431317 |
| **10** | cp | -1.648734 |
| **11** | thal | -2.989473 |
| **12** | sex | -5.552105 |

We observe that most of our attributes have skewness but if skewness is more than +/-0.5 will reduce the prediction when we train the model lets filter those data separately.

Feature names with Skewness is present more than +/-0.5 as follows:

Postive Skewed data:

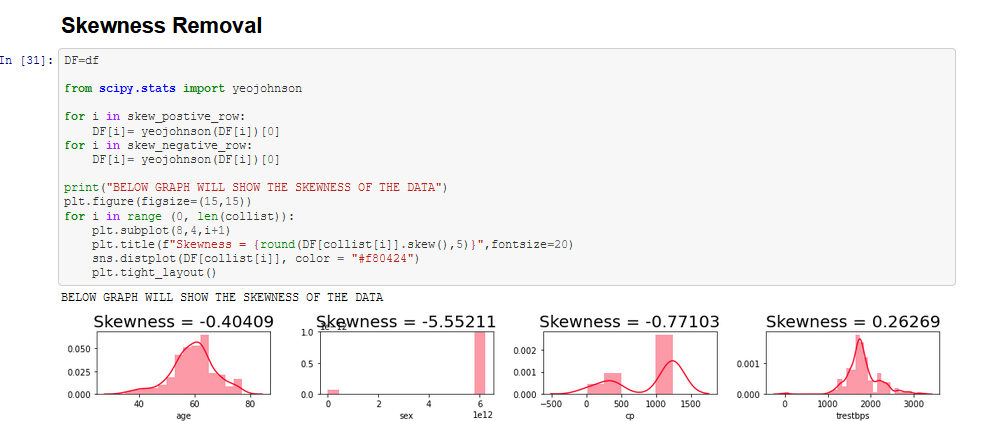
['fbs', 'oldpeak']

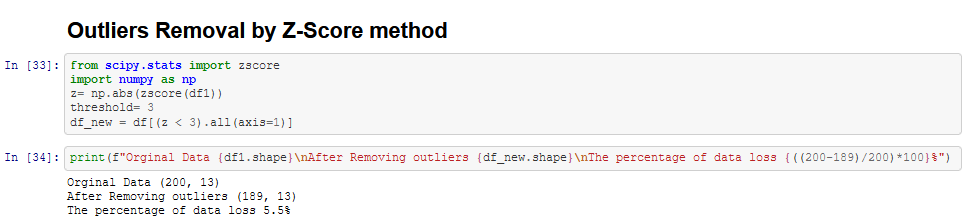
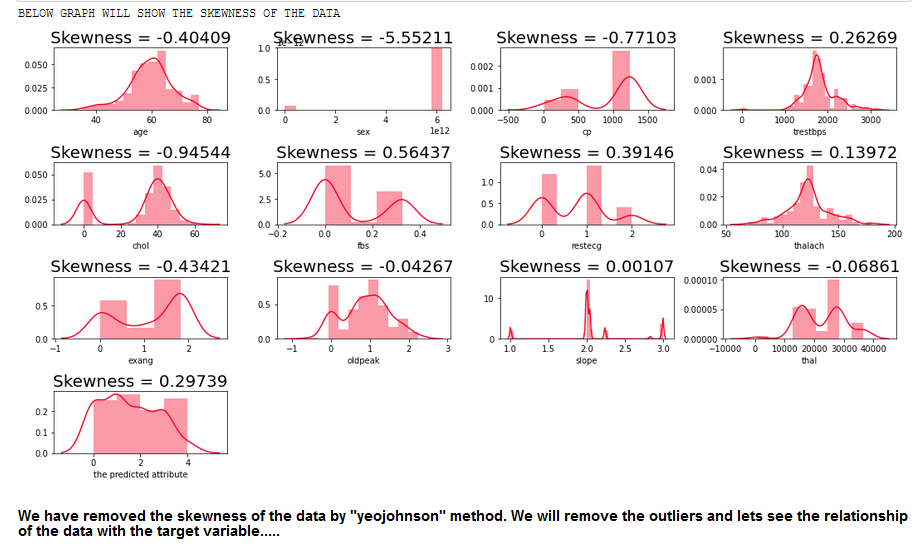
negative Skewed data:

['chol', 'exang', 'trestbps', 'cp', 'thal', 'sex']

We have to remove the skewness of the above filtered attributes but lets also see the outliers present in the data.

# Standardizing the Data

We also observed that data have more outliers lets figure out a method in correcting skewness and outliers without much loss in our data.

Skewness in our data is corrected below +/-0.5 let’s figure out a method in removing outliers in the data without much loss in the data.

 Loss of data post outlier’s removal is only 5.5% from the data so we will adopt the Z-Score Outliers removed data for our analysis. And finally, in our preprocessing pipeline we will scale the data with MinMaxScaler before we pass the data for training.

# Building Machine Learning Models.

## Model Training.

 We will train our data with eight different Classification algorithm and lets also find out best random\_state parameter before we actually split and train our data.

We have the Best Accuracy in GaussianNB as 85.41% on randomstate 263. We are splitting the data as x\_train, x\_test, y\_train, y\_test and test data size is 25% from the original data. Now we will train our data in following Classification algorithm.

1. "Naive Bayes Gaussian",
2. "K Neighbors Classifier",
3. "Random Forest",
4. "Decision Tree",
5. "Extra Tree",
6. "Ada Boost",
7. "Gradient Boosting",
8. "XGBoost"

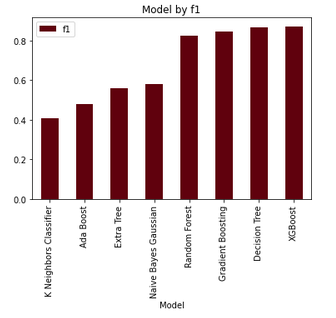
Post the training or our data we are sorting and filtering our model based on F1-Score, Accuracy, Precision, Recall.

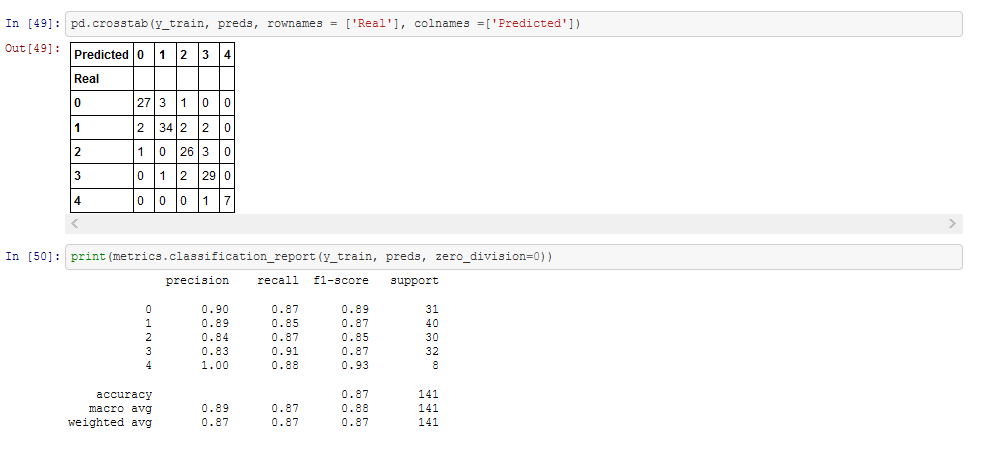


We are finalizing the best model with the highest F1-Score.

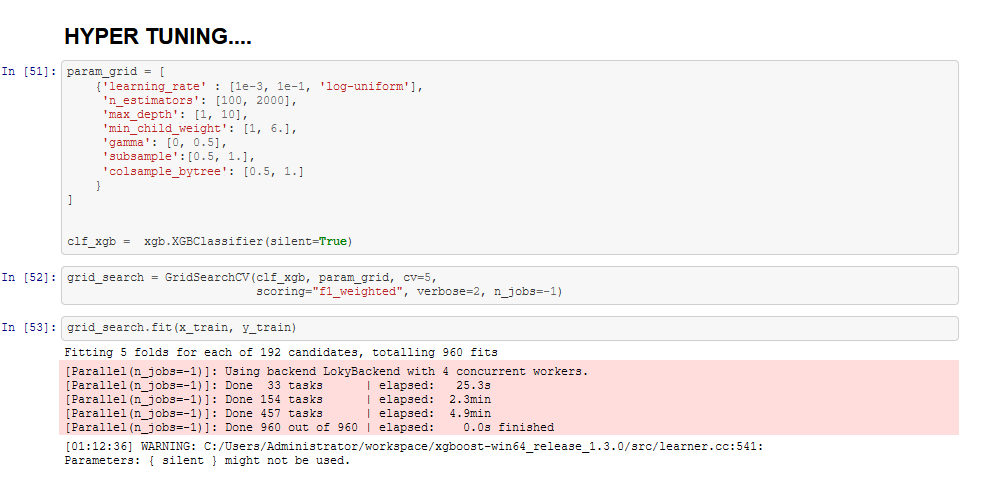
|  | **Model** | **accuracy** | **precision** | **recall** | **f1** |
| --- | --- | --- | --- | --- | --- |
| **7** | XGBoost | 0.872414 | 0.883742 | 0.872414 | 0.872099 |
| **3** | Decision Tree | 0.865271 | 0.886909 | 0.865271 | 0.865291 |
| **6** | Gradient Boosting | 0.850985 | 0.847042 | 0.850985 | 0.844062 |
| **2** | Random Forest | 0.830296 | 0.845531 | 0.830296 | 0.822390 |
| **0** | Naive Bayes Gaussian | 0.596305 | 0.585353 | 0.596305 | 0.579212 |
| **4** | Extra Tree | 0.568719 | 0.570484 | 0.568719 | 0.559285 |
| **5** | Ada Boost | 0.568473 | 0.461805 | 0.568473 | 0.477868 |
| **1** | K Neighbors Classifier | 0.426108 | 0.416452 | 0.426108 | 0.405382 |

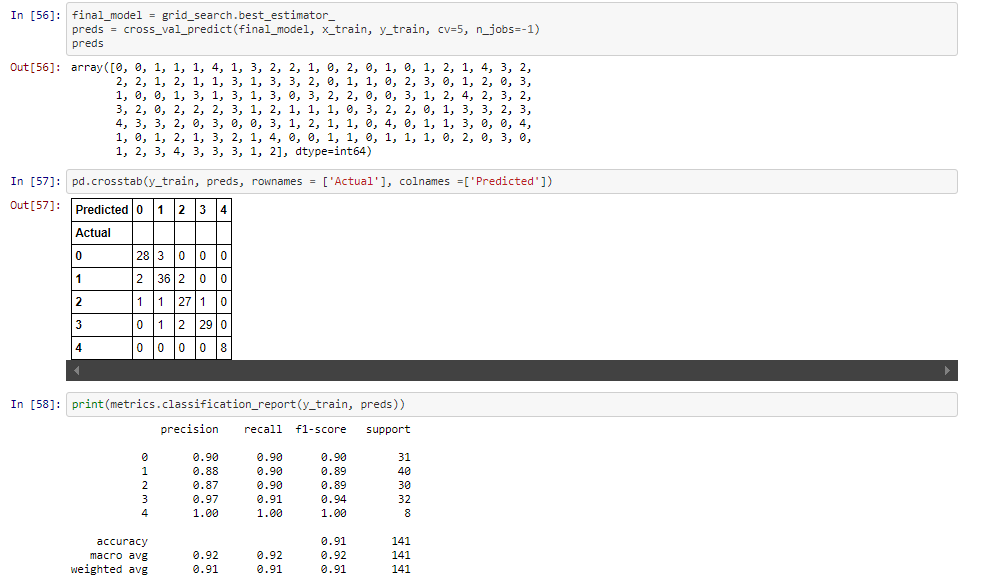
We observe that the XGBoost gives high accuracy and F1-Score.



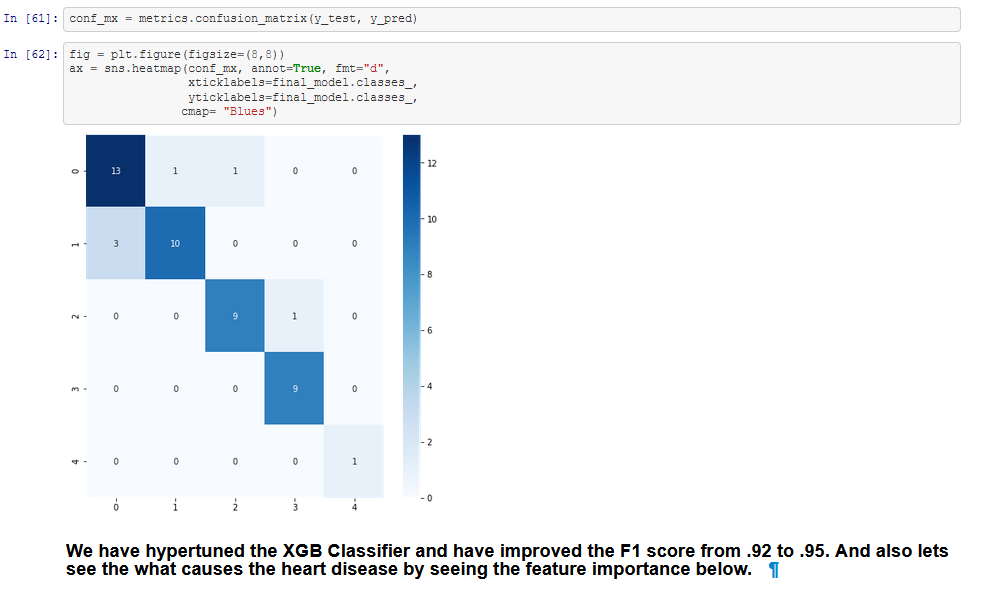
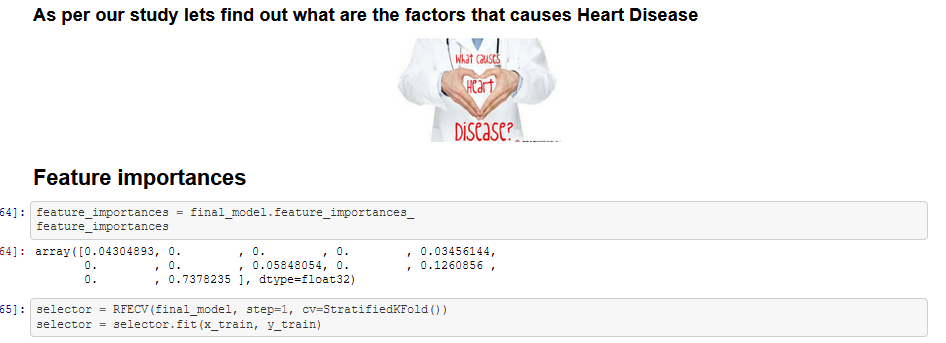


The above is the classification report of XGBoost model, now we will Hyper-parameter Tune the XGBoost model to increase the performance of the model.





Post the Hyper Parameter Tuning we can see that the F1-Score has increased from 89% - 90%



|  | **Feature\_Variables** | **Importance** |
| --- | --- | --- |
| **3** | resting blood pressure | 0.94 |
| **4** | cholestoral | 0.93 |
| **5** | fasting blood sugar | 0.93 |
| **6** | resting electrocardiographic | 0.91 |
| **7** | maximum heart rate achieved | 0.91 |
| **8** | exercise induced angina | 0.91 |
| **9** | oldpeak | 0.91 |
| **10** | the slope of the peak exercise ST segment | 0.91 |
| **11** | thal | 0.91 |
| **0** | age | 0.94 |
| **1** | sex | 0.90 |
| **2** | chest pain | 0.92 |



# Conclusion

Heart disease is the leading cause of the death in the United States as like said before. It is also a major cause of disability. There are many things that can raise your risk for heart disease. They are called risk factors. Some of them you cannot control, but there are many that you can control. Learning about them can lower your risk of heart disease. Our study analysis 13 factors with 200 records of different patients which caused the “Heart-Disease”. From which we figured the following 5 have more risk in causing heart disease.

1 Age 94%

2 Blood pressure 94%

3 Cholesterol 93%

4 Blood Sugar 93%

5 Chest pain 92%

# As suggested on internet Factors that could reduce risk of heart disease:

1. Control your blood pressure. High blood pressure is a major risk factor for heart disease. ...
2. Keep your cholesterol and triglyceride levels under control. ...
3. Stay at a healthy weight. ...
4. Eat a healthy diet. ...
5. Get regular exercise. ...
6. Limit alcohol. ...
7. Don't smoke. ...
8. Manage stress.