RUID: 197006702

Report on Causes of Heart Disease

Heart Disease:

Coronary illness depicts a scope of conditions that influence your heart. Ailments under the coronary illness umbrella incorporate vein maladies, for example, coronary supply route ailment; heart cadence issues (arrhythmias); and heart absconds you're brought into the world with (inborn heart surrenders), among others.

Numerous types of coronary illness can be anticipated or treated with solid way of life decisions.

Project Description:

The dataset contains 76 properties, yet all distributed tests allude to utilizing a subset of 14 of them. Specifically, the Cleveland database is the one in particular that has been utilized by ML analysts to this date. The "objective" field alludes to the nearness of coronary illness in the patient. It is the whole number esteemed from 0 (no nearness) to 4. Within the file there are 13 columns.

Overlook of Dataset columns.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Age: The individual's age in years

Sex: The individual's sex (1 = male, 0 = female)

Cp: The chest torment experienced (Value 1: run of the mill angina, Value 2: atypical angina, Value 3: non-anginal agony, Value asymptomatic)

Trestbps: The individual's resting pulse (mm Hg on admission to the clinic)

Chol: The individual's cholesterol estimation in mg/dl

Fbs: The individual's fasting glucose (> 120 mg/dl, 1 = genuine; 0 = bogus)

Restecg: Resting electrocardiographic estimation (0 = typical, 1 = having ST-T wave variation from the norm, 2 = indicating likely or distinct left ventricular hypertrophy by Estes' criteria)

Thalach: The individual's greatest pulse accomplished

Exang: Exercise prompted angina (1 = yes; 0 = no)

Oldpeak: ST misery actuated by practice comparative with rest ('ST' identifies with positions on the ECG plot. See progressively here)

Slant: the slant of the pinnacle practice ST portion (Value 1: upsloping, Value 2: level, Value 3: downsloping)

Ca: The quantity of significant vessels (0-3)

Thal: A blood issue called thalassemia (3 = normal; 6 = fixed deformity; 7 = reversable imperfection)

Target: Heart infection (0 = no, 1 = yes)

Objective:

The aim of the project is to build a model to predict the causes of Heart Disease , and to help achieve this objective there were 6 models used, the first being Logistic Regression modelling followed by a Classification algorithm called SVM, KNN, Naïve Bayes, Random Forest and Decision Tree to arrive at our predictions. The evaluation of each of the models were then compared and their results interpreted to come at a conclusion.

Techniques:

- **Data Cleaning:**
  We will check if there are any missing values in the dataset.

```
#To find if there are any missing values or null values.
data.isnull().sum()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Good to see there are no missing values.

  Results says there are no missing values which is good thing and will be easy to plot the further analysis.

- **EDA**
  Before diving into the analysis of the dataset, basic Exploratory data operations were performed on the dataset. The values of mean and median can give an idea where the outliers lie in a dataset, so to get an idea of the existing outliers in the dataset, python's inbuilt '*describe ()*' function was used, which gave the below output.

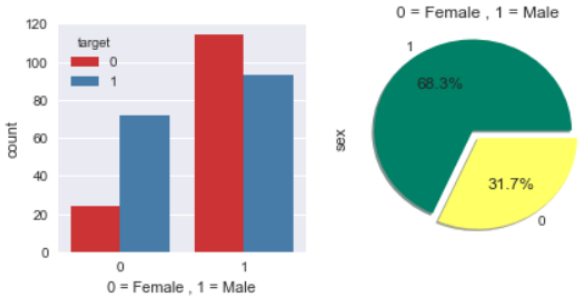
```
data.describe()
```

	age
count	303.000000
mean	54.366337
std	9.082101
min	29.000000
25%	47.500000
50%	55.000000
75%	61.000000
max	77.000000

From the above table we see that the mean worth is more prominent than the middle worth (list: 50 %), and notwithstanding that, there is a huge distinction between the estimations of the 75th %tile and the most extreme worth. In this manner, from these two perceptions we could state that there are outrageous qualities/exceptions in our dataset.

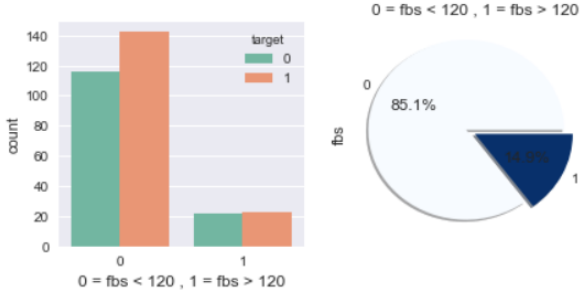
Further compared the columns and plotted the graphs to comprehend the overall analysis.

```
fig,ax=plt.subplots(1,2,figsize=(7,3))
sns.countplot(x='sex',data = data, hue='target',palette='Set1',ax=ax[0])
ax[0].set_xlabel("0 = Female , 1 = Male")
data.sex.value_counts().plot.pie(ax=ax[1],autopct='%1.1f%%',shadow=True, explode=[0.1,0], cmap = 'summer')
ax[1].set_title("0 = Female , 1 = Male")
plt.show()
```



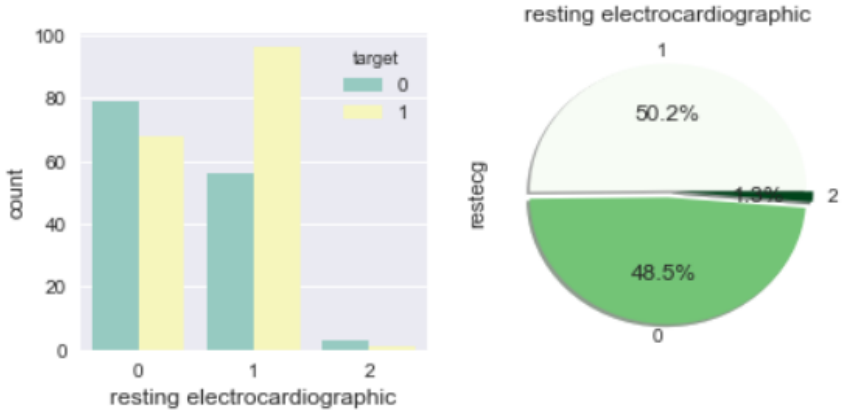
From the above graph we can see that, Number of Women experiencing Heart Disease are more than Men, yet Men populace is more than Women.

```
fig,ax=plt.subplots(1,2,figsize=(7,3))
sns.countplot(x='fbs',data=data,hue='target',palette='Set2',ax=ax[0])
ax[0].set_xlabel("0 = fbs < 120 , 1 = fbs > 120", size=12)
data.fbs.value_counts().plot.pie(ax=ax[1],autopct='%1.1f%%',shadow=True, explode=[0.1,0],cmap='Blues')
ax[1].set_title("0 = fbs < 120 , 1 = fbs > 120", size=12)
plt.show()
```



From the above graph its strange to see that people having less than 120 mg/dl has higher risk of causing Heart Disease and prediabetes.

```
fig,ax=plt.subplots(1,2,figsize=(7,3))
sns.countplot(x='restecg',data=data,hue='target',palette='Set3',ax=ax[0])
ax[0].set_xlabel("resting electrocardiographic",size=12)
data.restecg.value_counts().plot.pie(ax=ax[1],autopct='%1.1f%%',shadow=True,
                                     explode=[0.005,0.05,0.05],cmap='Greens')
ax[1].set_title("resting electrocardiographic",size=12)

plt.show()
```

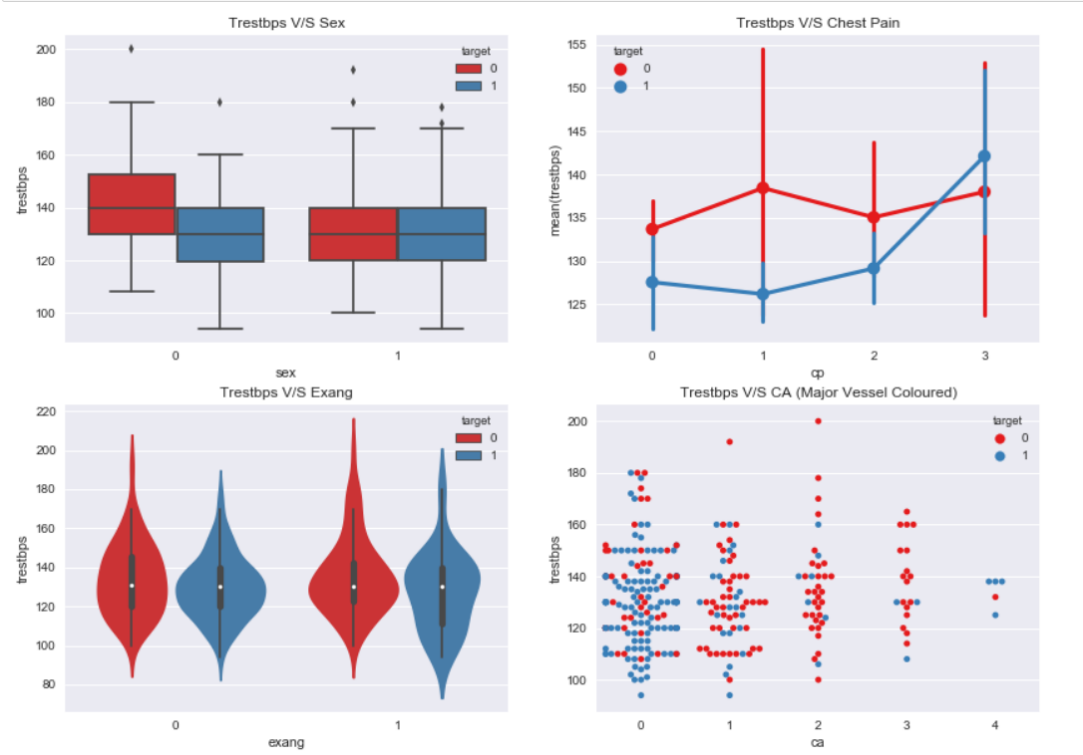


Analysing the target and ECG to find the sufferings of Heart Disease. An electrocardiogram (ECG) is a test which quantifies the electrical movement of your heart to show whether it is working ordinarily. An ECG records the heart's mood and action on a moving piece of paper or a line on a screen.
From the above graph we can understand that ECG = 1 will lead to person causing higher chance of Heart Disease. Similarly, I plotted for other columns as well.

Then I implemented the continuous feature which contains a limited number of classes or unmistakable gatherings. Unmitigated information won't have a legitimate request. Ceaseless factors are numeric factors that have a limitless number of qualities between any two qualities. A constant variable can be numeric or date/time.

I compared the trestbps with Gender, Chest Pain, Exang and Major Vessels in order to plot the graphs and prognosis the basic understanding of causes of Heart Disease.

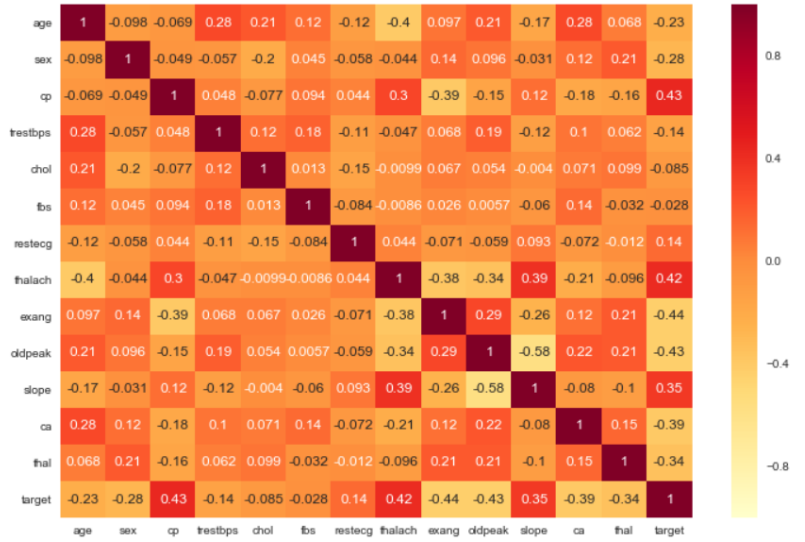
RUID: 197006702



In light of the above examination, we can say that Gender assumes a minor job as for Blood Pressure (trestbps). Yet, Chest Pain assumes Vital Role. As Chest torment builds Blood Pressure will likewise increment alongside odds of Heart Diseases. Similarly, plotted for Cholesterol and old peak column found the basic causes of the infection.

After that to find correlation I implemented the HeatMap where I found that the vast majority of the highlights are profoundly corresponded with one another

```
plt.figure(figsize=(12,8))
sns.heatmap(data.corr(),annot=True,cmap = 'YlOrRd')
plt.show()
```



- **Feature Transformation**
  It is utilized to cast a pandas item to a predefined dtype. astype() work additionally gives the capacity to change over any appropriate existing section to categorical type. DataFrame.astype() work comes helpful when we need to case a specific section data type to another data type.

```
data.sex=data.sex.astype('category')
data.cp=data.cp.astype('category')
data.fbs=data.fbs.astype('category')
data.restecg=data.restecg.astype('category')
data.exang=data.exang.astype('category')
data.ca=data.ca.astype('category')
data.slope=data.slope.astype('category')
data.thal=data.thal.astype('category')
```

```
X = data.drop(['target'], axis = 1)
y = data.target.values
```

- **Creating Dummy Variables**
  Pandas has a capacity which can transform an absolute variable into a progression of zeros and ones, which makes them significantly simpler to evaluate and analyze.

	age	sex	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	ca	...	cp_1	cp_2	cp_3	slope_0	slope_1	slope_2	thal_0	thal_1	thal_2	thal_3
0	63	1	145	233	1	0	150	0	2.3	0	...	0	0	1	1	0	0	0	1	0	0
1	37	1	130	250	0	1	187	0	3.5	0	...	0	1	0	1	0	0	0	0	1	0
2	41	0	130	204	0	0	172	0	1.4	0	...	1	0	0	0	0	1	0	0	1	0
3	56	1	120	236	0	1	178	0	0.8	0	...	1	0	0	0	0	1	0	0	1	0
4	57	0	120	354	0	1	163	1	0.6	0	...	0	0	0	0	0	1	0	0	1	0

5 rows × 21 columns

- **Implementing of Predictive Models**
  Predictive model is a procedure that utilizations information mining and the likelihood to figure results. Each model is comprised of various indicators, which are factors that are going to impact future outcomes. When information has been gathered for important indicators, a factual model is detailed.

  Further the dataset we use is typically part of training data and test data. The training set contains a known yield and the model learns on this data so as to be summed up to other information later on. We have the test dataset (or subset) so as to test our model's expectation on this subset.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state=0)
```

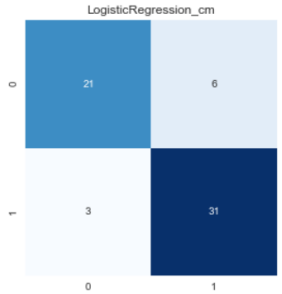
- **Evaluation of Model using Cross Validator.**

  Each AI calculation works best under a given arrangement of conditions. Ensuring your calculation fits the presumptions/prerequisites guarantees unrivalled execution. You can't utilize any calculation in any condition. For e.g.: We can't utilize straight relapse on a categorical dependent variable. Since we won't be acknowledged for getting incredibly low estimations of balanced R^2 and F measurement. Rather, in such circumstances, we should

take a stab at utilizing calculations, for example, Logistic Regression, Decision Trees, Support Vector Machine (SVM), Random Forest, and so on.

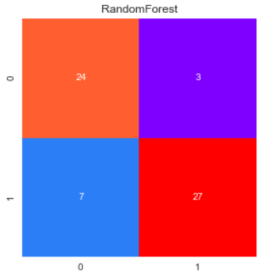
Logistic regression is a measurable model that in its fundamental structure utilizes a calculated capacity to demonstrate a parallel ward variable, albeit a lot of increasingly complex expansions exist. In regression examination, strategic relapse (or logit relapse) is assessing the parameters of a calculated model (a type of twofold relapse).

```
#LogisticRegression
lr_c=LogisticRegression(random_state=0)
lr_c.fit(X_train,y_train)
lr_pred=lr_c.predict(X_test)
lr_cm=confusion_matrix(y_test,lr_pred)
lr_ac=accuracy_score(y_test, lr_pred)

plt.figure(figsize=(20,10))
plt.subplot(2,4,1)
plt.title("LogisticRegression_cm")
sns.heatmap(lr_cm,annot=True,cmap="Blues",fmt="d",cbar=False)
```




```
#RandomForest
rdf_c=RandomForestClassifier(n_estimators=10,criterion='entropy',random_state=0)
rdf_c.fit(X_train,y_train)
rdf_pred=rdf_c.predict(X_test)
rdf_cm=confusion_matrix(y_test,rdf_pred)
rdf_ac=accuracy_score(rdf_pred,y_test)

plt.figure(figsize=(20,10))
plt.subplot(2,4,4)
plt.title("RandomForest")
sns.heatmap(rdf_cm,annot=True,cmap="rainbow",fmt="d",cbar=False)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x270be62cd68>
```

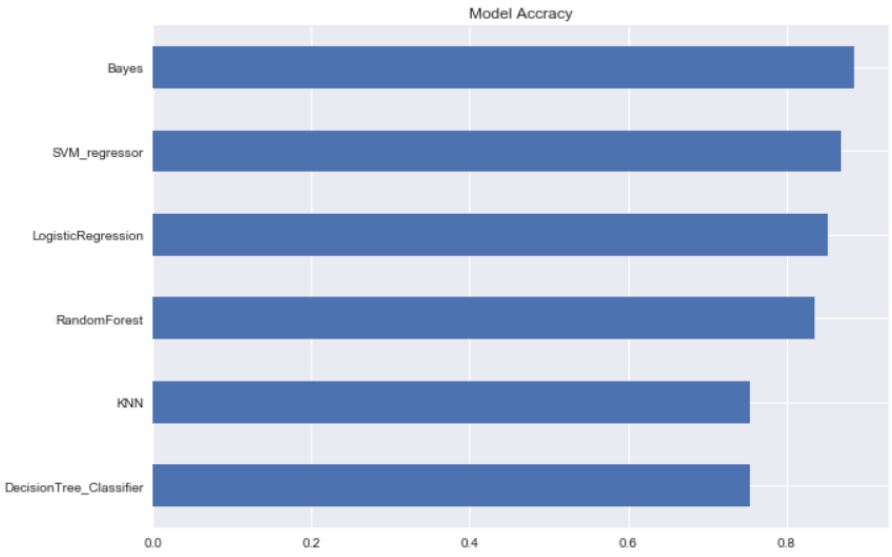


Similarly plotted the heatmap for model, SVM regression, Random Forest, Decision Tree, Bayes Classification to find the correlation between the columns and accuracy level of the predictions. Machine learning version accuracy is the size used to decide which version is high-quality at identifying relationships and styles between variables in a dataset based totally at the enter, or schooling, statistics. The better a model can generalize to 'unseen' statistics, the better predictions, and insights it could produce, which in flip supply extra business price.

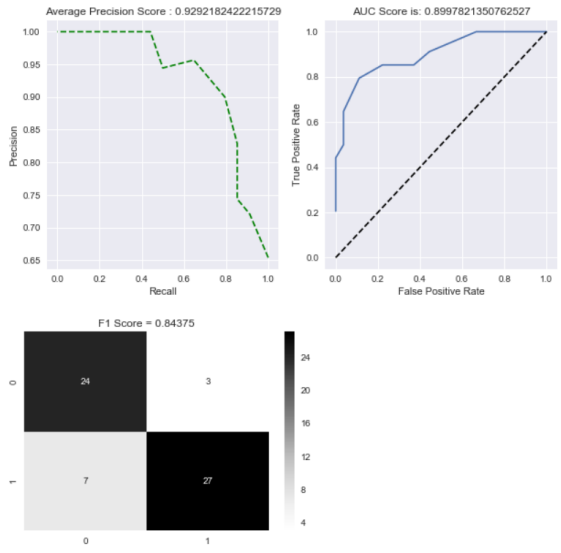
```
print('LogisticRegression_accuracy:\t',lr_ac)
print('SVM_regressor_accuracy:\t\t',svr_ac)
print('RandomForest_accuracy:\t\t',rdf_ac)
print('DecisionTree_accuracy:\t\t',dtree_ac)
print('KNN_accuracy:\t\t\t',knn_ac)
print('BayesClassifier_accuracy:\t',bayes_ac)
```

```
LogisticRegression_accuracy:     0.852459016393
SVM_regressor_accuracy:          0.868852459016
RandomForest_accuracy:           0.83606557377
DecisionTree_accuracy:           0.754098360656
KNN_accuracy:                    0.754098360656
BayesClassifier_accuracy:        0.885245901639
```

From the above results we can see that Naive Bayes Classifier algorithm gives best accuracy of 88.5% as compared to another machine learning algorithm.



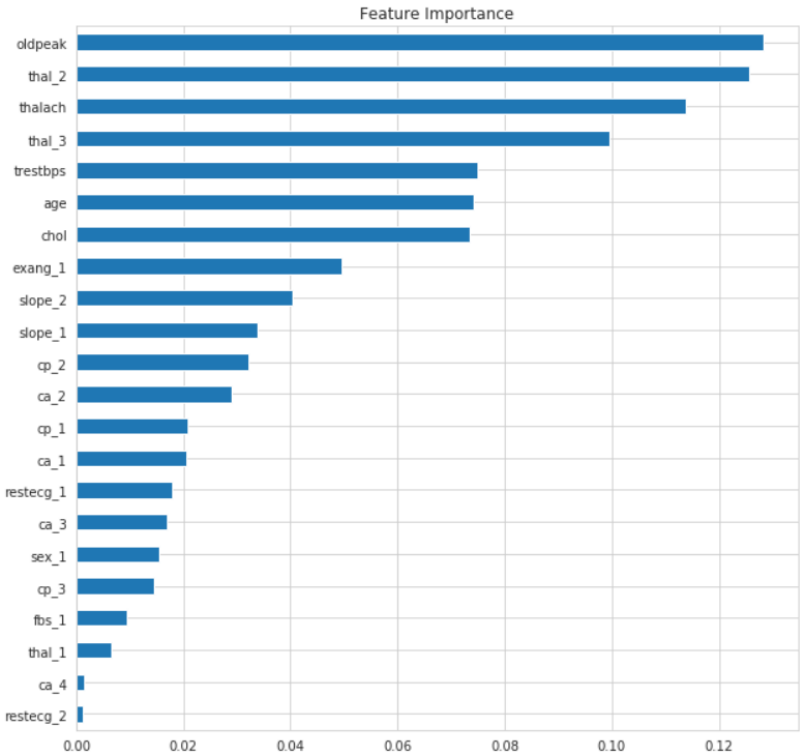
From the above graph we can see that accuracy of Bayes and SVM regression are doing better than other ML algorithms. In any case, for Classification task ACCURACY isn't significant. Rather than exactness model ought to be made a decision on premise of AUC (Area under Curve), ROC CURVE, High Precision and High Recall esteems. F1 score likewise assume important job which is equivalents to 2/ (1/exactness + 1/Recall) score.
So, I implemented the Precision Score, Roc Curve and F1 score for all the models.



ROC Curves summarize the alternate-off among the actual effective charge and false tremendous price for a predictive model the usage of different possibility thresholds. ROC curves are suitable while the observations are balanced between every magnificence, whereas precision-consider curves are suitable for imbalanced datasets.

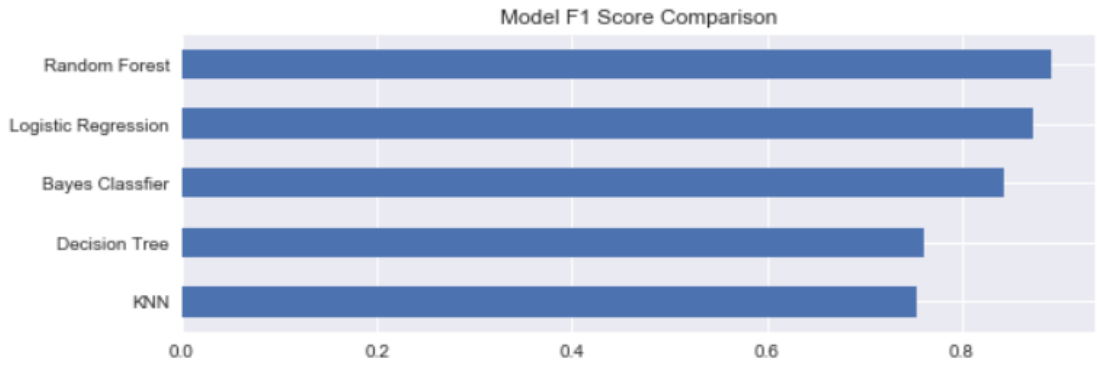
Results and Conclusion:

Implemented the Feature importance to predict the actual cause of Heart disease.



From the above graph we can see **thal_2 and old peak** are most important feature in prediction.

Further, I plotted the F1 Score Comparison which gave me different results.



From the above graph we can conclude that Random Forest is the best model to give accurate results. It is providing 92% of precision level.

Reference:

https://towardsdatascience.com/the-basics-knn-for-classification-and-regression-c1e8a6c955

https://towardsdatascience.com/understanding-logistic-regression-9b02c2aec102

https://towardsdatascience.com/understanding-random-forest-58381e0602d2

https://www.microstrategy.com/us/resources/introductory-guides/predictive-modeling-the-only-guide-you-need