

## HEART DISEASE DETECTION REPORT


BY


Gurarshdeep Kaur , Student ID 103490976

Swinburne University of Technology

Contact Ph. No. : 0401537539.

---

 **ABSTRACT** : The aim of this assignment using “Heart Disease” dataset (from Kaggle) is to detect whether the people have heart disease or not by considering given number of attributes from the dataset. I have used graphical visualisations and descriptive statistics to explore all the attributes and the relationship among them , which contributes for heart disease or healthy heart. The data has been split into three suites with different ratios for training and testing test in each suite. In this assignment , Logistic Regression and Decision Tree Classifier algorithms have been used to detect heart disease. And among these two algorithms, Logistic Regression gives the best accuracy of 90.2% when 80% of data was used for training and 20% for testing. Heatmaps and Roc curves are used to compare the performances of these two models in each suite.

 **INTRODUCTION** : As healthiness is important in today’s world for every human being and people with a healthy heart are counted as healthy. Keeping health factor in mind , ‘Heart disease’ dataset, from Kaggle, has been chosen to detect the people with healthy heart or with a diseased heart. The dataset consists of categorical as well as numerical columns and appropriate cleaning steps are used to clean the dataset. Each chief feature and their relationship with each other have been explored using descriptive statistics and graphical visualisations. After standardising some features of dataset , data has been split into three suites, which are as follows :

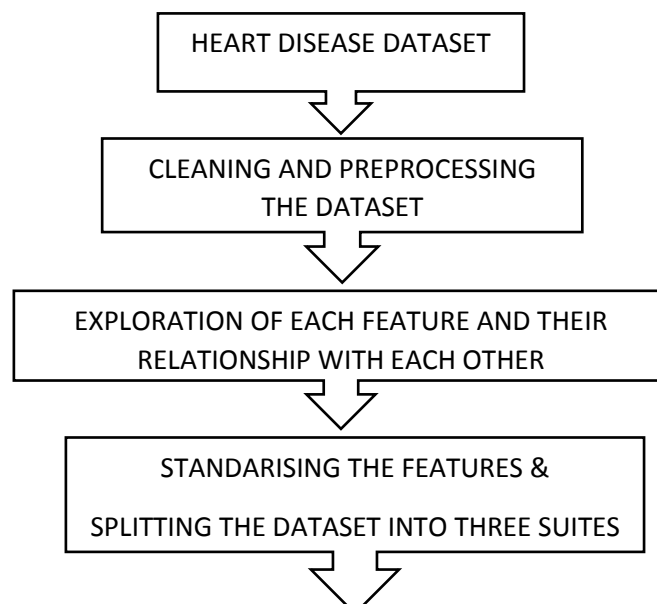
Suite1: 50% for training and 50% for testing

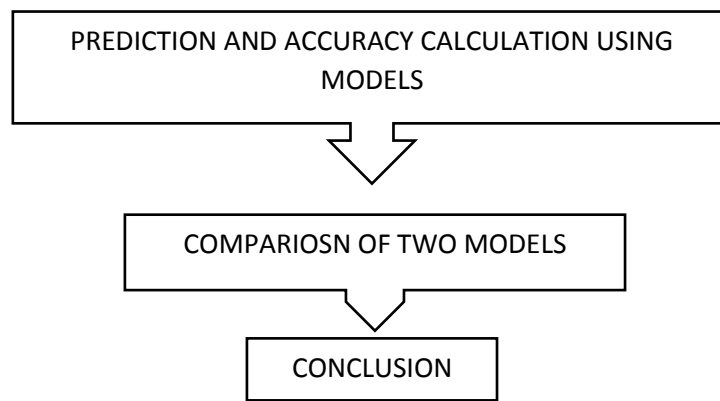
Suite2: 60% for training and 40% for testing

Suite3: 80% for training and 20% for testing

Logistic Regression and Decision Tree Classifier algorithms has been applied on each suite. After executing these steps, it is sensible to conclude that Logistic Regression algorithm gives the finest accuracy of 90.2% in the third suite to detect the heart disease.

The entire workflow of this assignment is given below :



















## **TASK 1 : Problem Formulation, Data Acquisition and Preparation**

### 1. **THE DATASET :**

Firstly, I explored UCI repository and Kaggle to choose one dataset for this assignment, which can satisfy the given conditions of containing categorical and numerical columns. According to the UCI, "This dataset contains 76 attributes, but all the published experiments refer to using a subset of 14 of them", which are easily available from Kaggle. So, I came to the decision of choosing dataset named Heart Disease from Kaggle, which is a subset of UCI repository data containing specific attributes.

This dataset has 14 attributes and 303 rows. The explanation of the attributes used in this dataset is given below:

-  age: The person's age in years
-  sex: The person's sex (1 = male, 0 = female)
-  cp: chest pain type
  - Value 0: asymptomatic
  - Value 1: atypical angina
  - Value 2: non-anginal pain
  - Value 3: typical angina
-  trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
-  chol: The person's cholesterol measurement in mg/dl
-  fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = yes; 0 = no)
-  restecg: resting electrocardiographic results
  - Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria
  - Value 1: normal
  - Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-  thalach: The person's 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 — 0: down sloping; 1: flat; 2: up sloping
-  ca: The number of major vessels (0–3)
-  thal: A blood disorder called thalassemia Value.
  - Value 1: fixed defect (no blood flow in some part of the heart)
  - Value 2: normal blood flow
  - Value 3: reversible defect (a blood flow is observed but it is not normal)
-  target: Heart disease (1 = no, 0 = yes)

### 2. **CLEANING THE DATASET :**

### DUPLICATES:

There were two duplicated rows in the dataset which has been dropped by keeping the one in the dataset, using the command “df.drop\_duplicates(inplace = True)” , where “df” is dataset used i.e., df = heart.csv

### OUT OF RANGE VALUES :

The possible range for the attributes ‘ca’ and ‘thal’ is 0-3 and (1,2,3) respectively. But in the dataset , there was some values which were out of range and that values are replaced as “NaN”. After this step, the next step was to fill the null values using this command “df.fillna(df.median())”.

Now, the whole dataset is properly cleaned using appropriate steps and is ready for exploration.

## **TASK 2 : DATA EXPLORATION**

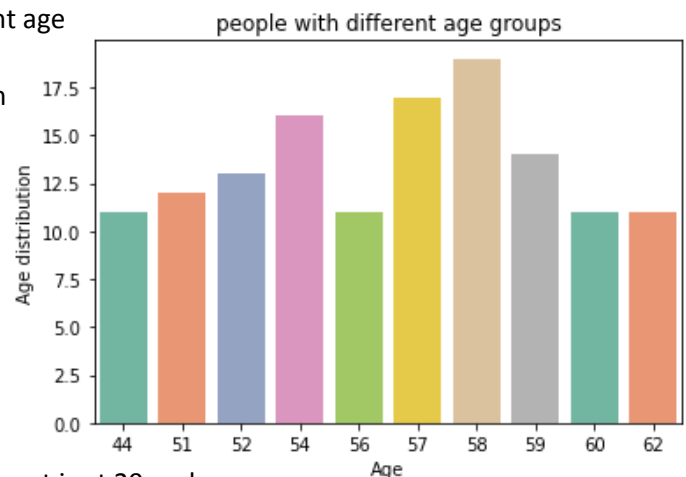
### • EXPLORATION OF EACH COLUMN :

#### 1. AGE:

The dataset focused on people with different age groups to detect whether they have heart disease or not . I have plotted a graph which focus only on the age group who have 10 or more than 10 people of that age in the dataset.

#### OBSERVATIONS :

- \*) Most people in the dataset have age between 50s and 60s.
- \*) People who have age of 58 are highest number of persons to be considered in the dataset.
- \*) The mean age is about 54 years and youngest is at 29 and oldest is at 77.

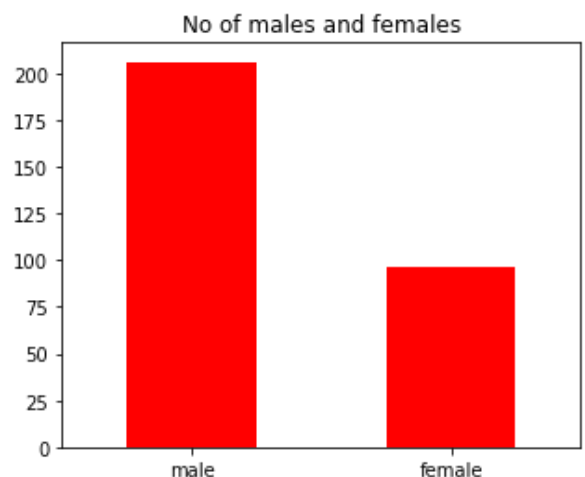


#### 2. SEX:

The dataset focused on people with different genders to detect whether they have heart disease or not .

#### OBSERVATIONS :

- \*) From the graph, it can be clearly seen that the number of males in the dataset are highest than the number of females.
- \*) The number of males are higher than 200 while female can be counted to be approximately near 100 in the dataset.

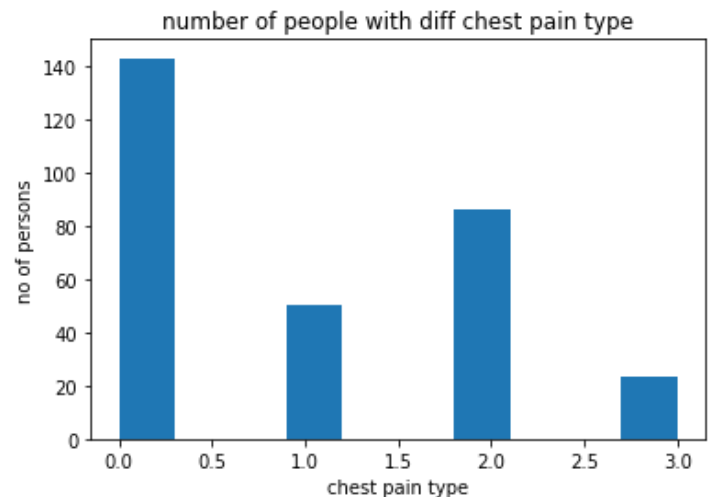


### 3. CHEST PAIN TYPE :

The dataset has attribute 'cp' to detect the chest pain type among different persons.

The values of chest pain type are as follows:

- Value 0: asymptomatic
- Value 1: atypical angina
- Value 2: non-anginal pain
- Value 3: typical angina



#### OBSERVATIONS:

\*) Most of the people have asymptomatic chest pain i.e.

140 people have asymptomatic chest pain type.

\*) The chest pain type 'typical angina' is least among the people.

\*) Most people are suffered from asymptomatic and atypical angina chest pain type

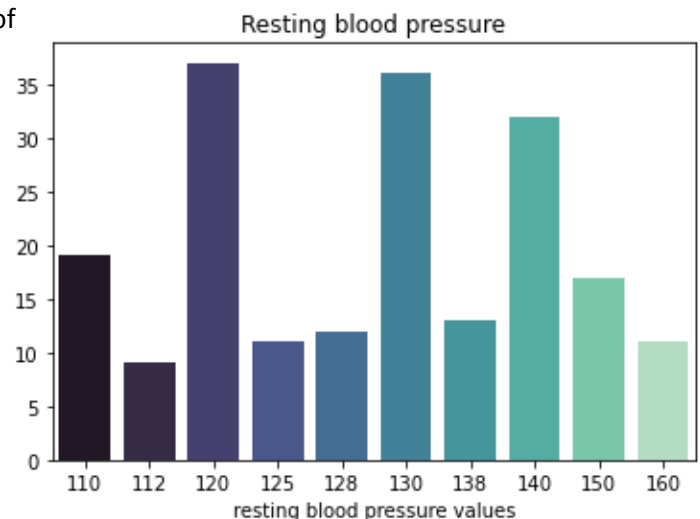
### 4. RESTING BLOOD PRESSURE (trestbps) :

There were fluctuations in the values of the resting blood pressure of the people. I focused on mapping the most common resting blood pressure values of the people.

#### OBSERVATIONS :

\*) The most common blood pressure of people is either 120 or 130, which is the normal resting blood pressure in the world of medical science.

\*) The graph shows that 9 people have 112 blood pressure.



### 5. RESTECG :

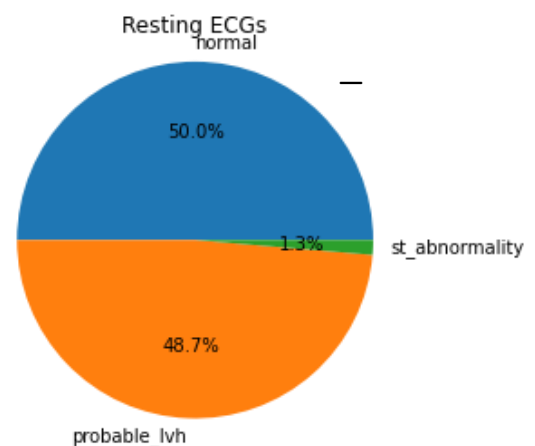
Restecg attribute is the resting electro cardio graphic results which have the following values:

- Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria
- Value 1: normal
- Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of  $> 0.05$  mV)

#### OBSERVATIONS :

\*) Almost 50% of the people have normal ECG results.

\*) A little wave abnormality has been seen among 1.3% of people.



\*) 48.7% of people have probable or definite left ventricular hypertrophy.

#### 6. TARGET:

The aim of this assignment is to detect that whether people have a heart disease or not. The target attribute has values which are given below:

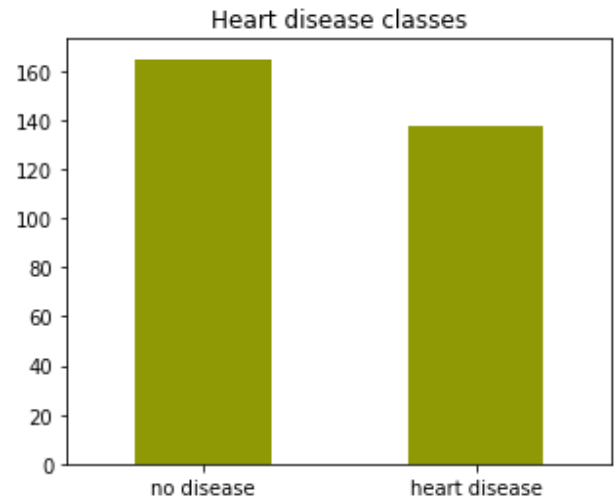
1 = no heart disease

0 = heart disease

#### OBSERVATIONS:

\*) More than 160 people have no heart disease.

\*) Nearly 140 people have heart disease.



#### 7. FASTING BLOOD SUGAR (fbs):

There were many people with fasting blood sugar which is greater than 120 mg/dl and the range of fbs attribute is given below :

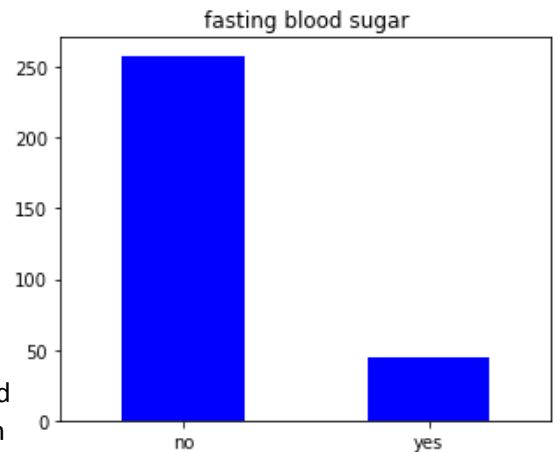
1 = yes

0 = No

#### OBSERVATIONS :

\*) Most of the people does not has fasting blood sugar.

\*) There are very few people with fasting blood sugar >120 mg/dl. The number of persons with fbs is 45.



#### 8. THAL :

Thal is a blood disorder value called thalassemia value which have following possible range :

Value 1: fixed defect (no blood flow in some part of the heart)

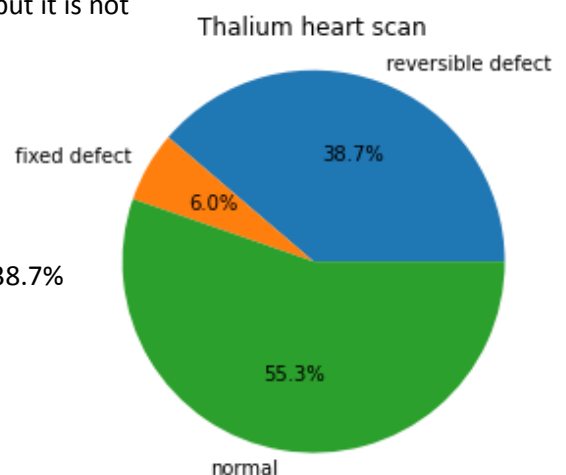
Value 2: normal blood flow

Value 3: reversible defect (a blood flow is observed but it is not normal)

#### OBSERVATIONS :

\*) From the graph it is clearly seen that more than 50% of people have normal thal value.

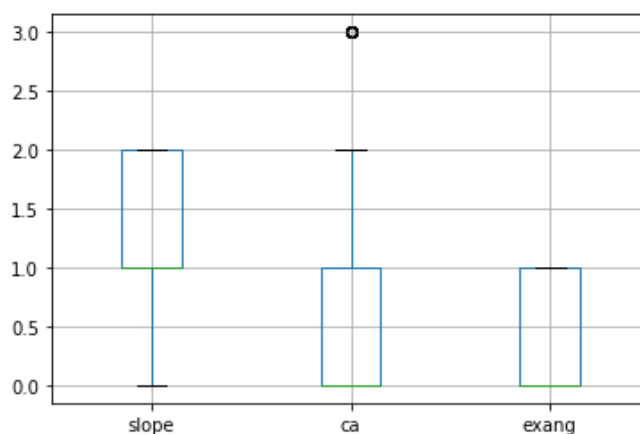
\*) Only 6% of people have fixed defect and 38.7% of people have reversible defect (cannot be counted as normal).



## 9. SLOPE, CA, EXANG :

The slope attributes explain the slope of the peak exercise ST segment which have values :

- 0: down sloping
- 1: flat
- 2: up sloping



CA attribute have range from 0-3 and it is the number of major vessels.

Exang attribute have only two values 1 and 0 which is yes and no respectively, and it is the exercise induced angina.

### OBSERVATIONS:

The boxplot depicts the attributes 'ca', 'slope' and 'exang'. From the graph it is clearly seen that all three attributes have values between possible ranges respective of the ranges given in the dataset. The median is 2 for slope and ca whereas for exang, median is 1, All the data for slope attribute is below the middle quartile.

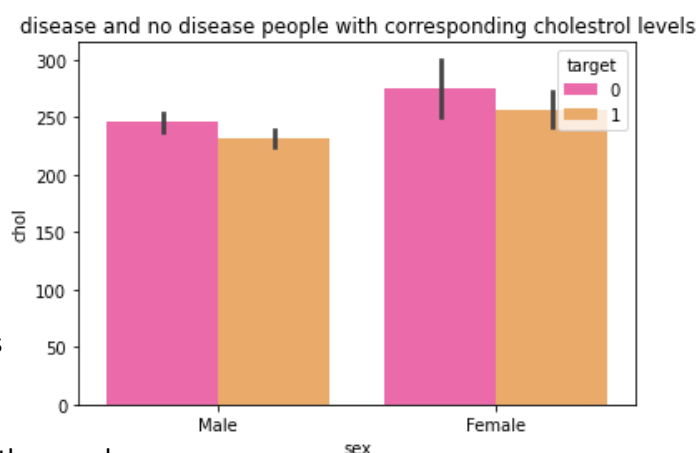
## EXPLORING RELATIONSHIP BETWEEN ALL PAIRS OF COLUMNS

### 1. CHOLESTROL + TARGET + SEX

The graph shows the relationship among the columns chol, target and sex.

#### OBSERVATIONS

- \*) The cholesterol level among females is higher as compared to males.
- \*) The cholesterol level of people with heart disease (target=0) is higher than the people with no heart disease(target=1).
- \*) female with heart diseases have high cholesterol as compared to males.

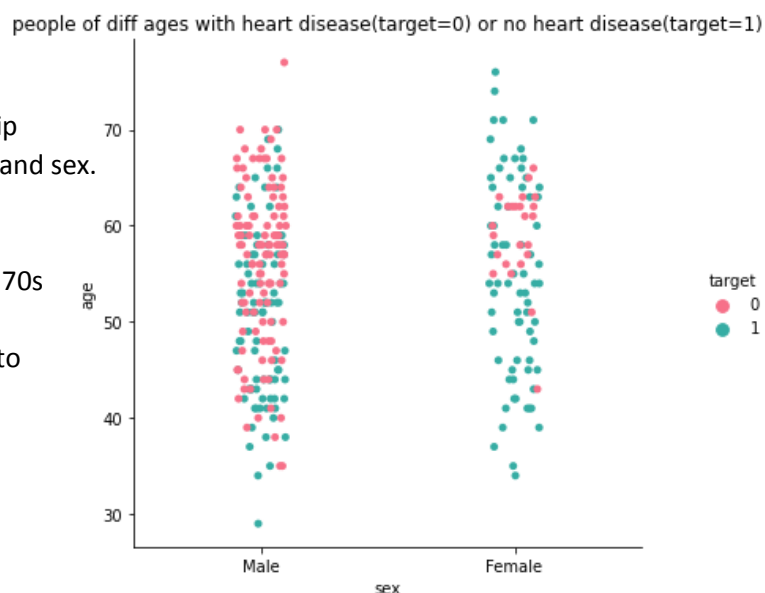


### 2. TARGET + AGE + SEX :

The graph shows the relationship among the columns age, target and sex.

#### OBSERVATIONS :

- \*) Males between ages 50s and 70s are more likely to have heart disease(target=0) as compared to females.



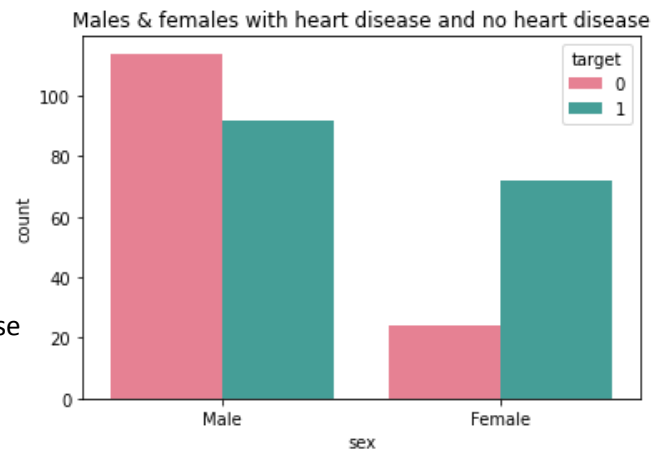
- \*) Number of cases of heart disease among females are more in the ages between 50 -65.
- \*) Number of cases of heart disease are less in the age gap of 25 – 50 among males.

### 3. SEX + TARGET :

The graph shows relationship among the columns sex and target.

#### OBSERVATIONS:

- \*) Heart disease among males contributes for total for approximately 120 cases whereas females heart disease cases are nearly 40.
- \*) More number of males are healthier(target=1) than females.

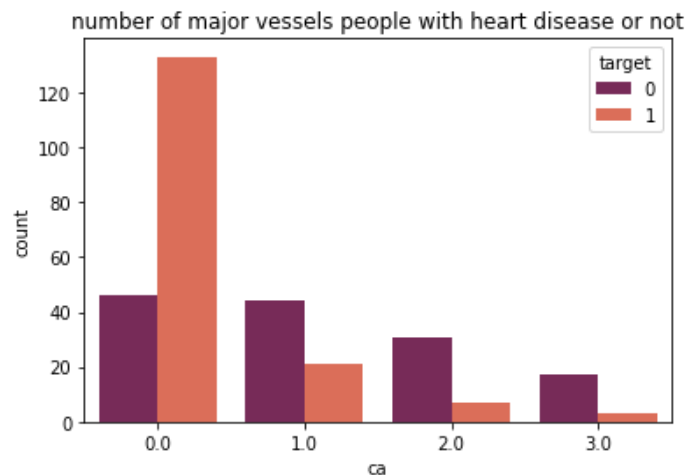


### 4. TARGET + CA :

The graph shows relationship among the columns ca and target.

#### OBSERVATIONS :

- \*) There are no major vessels in the case where people have no heart disease (target=1) .
- \*) The people with heart disease (target=0) have number of major vessels ranging between 0-3 .
- \*) Three major vessels are less likely to be found in case of people with heart disease.

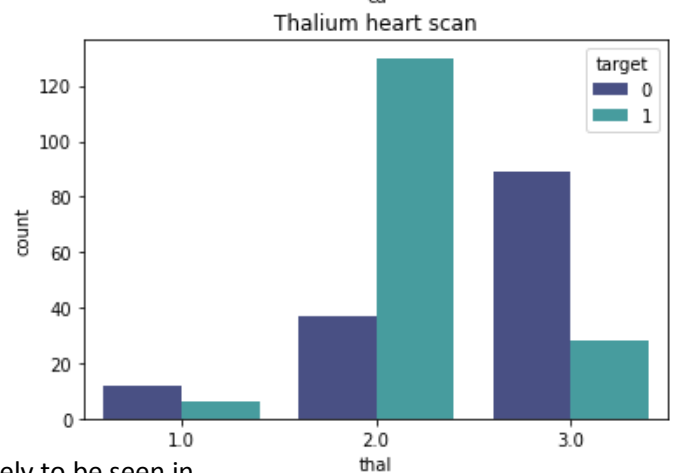


### 5. THAL + TARGET :

The graph shows relationship among the columns thal (a blood disorder) and target.

#### OBSERVATIONS :

- \*) People with no heart disease(target=1) contributes for the major section of normal blood flow i.e. thal value =2.
- \*) Fixed defect (thal value =1) is less likely to be seen in people with heart disease and no heart disease., but it is present among nearly 20 people.

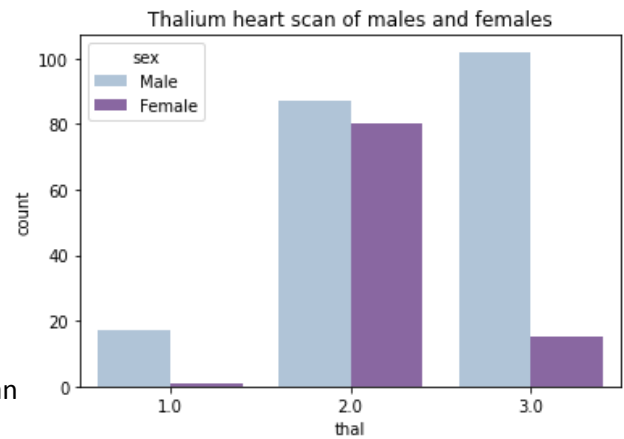


#### 6. THAL + SEX :

The graph shows relationship among the columns thal (a blood disorder) and sex.

##### OBSERVATIONS :

- \*) Normal thal (value=2) is higher in males than in females.
- \*) More than 100 males have higher number of reversible defect (thal value =3)
- \*) Females are less likely to fall under the category of fixed defect (thal value =1) than males.

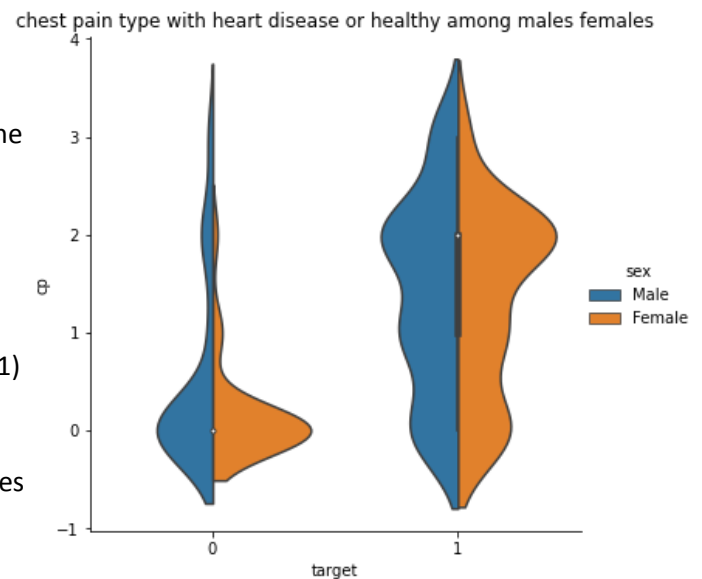


#### 7. CP + TARGET + SEX :

The graph shows relationship among the columns cp (chest pain type), sex and target.

##### OBSERVATIONS :

- \*) It is clear that males and females with no heart disease (target=1) have atypical angina (cp value=1) and non-anginal pain (cp value=2) types.
- \*) Approximately equal number of males and females with heart disease (target=0) falls under the category of asymptomatic chest pain type (value=0).

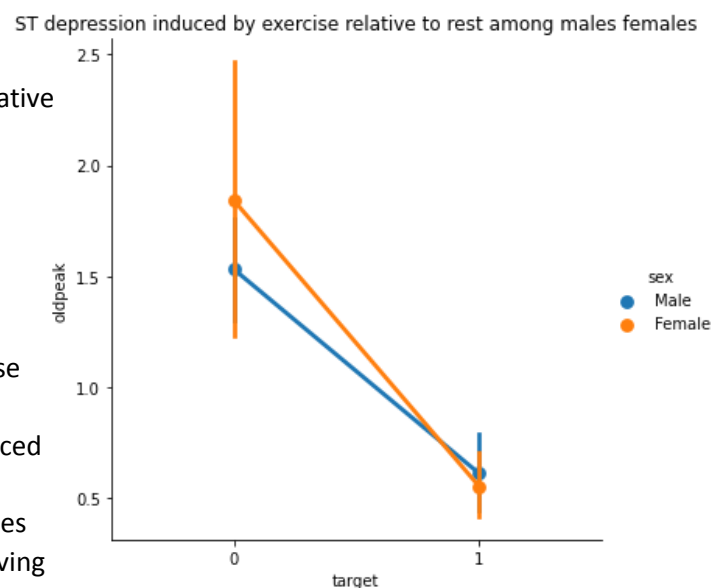


#### 8. OLDPEAK + TARGET + SEX :

The graph shows relationship among the columns oldpeak (ST depression induced by exercise relative to rest), sex and target.

##### OBSERVATIONS :

- \*) Females with heart disease (target=0) is larger in count in terms of depression in contrast to the females with no heart disease (target=1).
- \*) Males have less depression induced by exercise relative to rest in comparison of females, in both cases of having heart disease and not having heart disease.





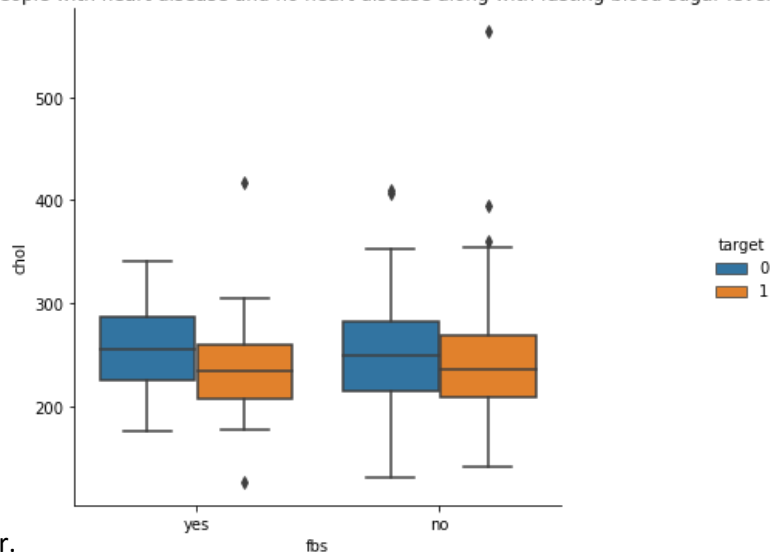
## 9. CHOLESTROL + BLOOD SUGAR + TARGET :

The graph shows relationship among the columns chol (cholesterol level) , fbs (fasting blood sugar level) and target.

### OBSERVATIONS:

\*) People who have healthy heart(target=1) also have the fasting blood sugar.

\*) People with heart disease are more likely to have fasting blood sugar.



## 10. EXANG + THAL :

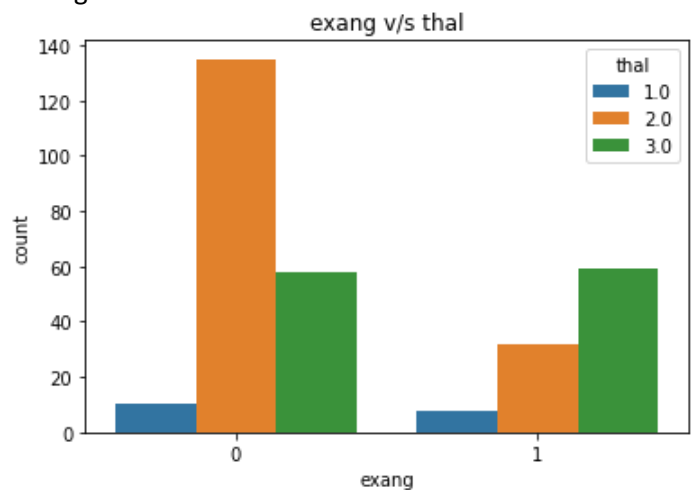
The graph shows the relationship among the attributes exang(exercise induced angina) and thal(a blood disorder).

### OBSERVATIONS :

\*) People having no angina (0 = no) are more likely to have normal blood flow (thal value=2) in comparison of people having angina.

\*) People with reversible defect flow (thal value =3) along with having angina and not having angina are same.

\*) Also, people with fixed defect (thal value=0) are same irrespective of having angina or not having angina.



## What is the number of males and females having heart disease and not having heart disease?

This question has been posed to explore the dataset more deeply. To find the answer we will use graphical visualisations and other appropriate methods.

Firstly, we will know the exact number of males and females in each category of having heart disease and not having heart disease, using the command “`pd.crosstab(df['sex'], df['target'])`”. This gives the result as follows :

Here , target 0 = heart disease and target 1 = No heart disease.

For more clarification , the below stacked bar plot has been shown to conclude key observations.

target	0	1
sex		
Female	24	72
Male	114	92

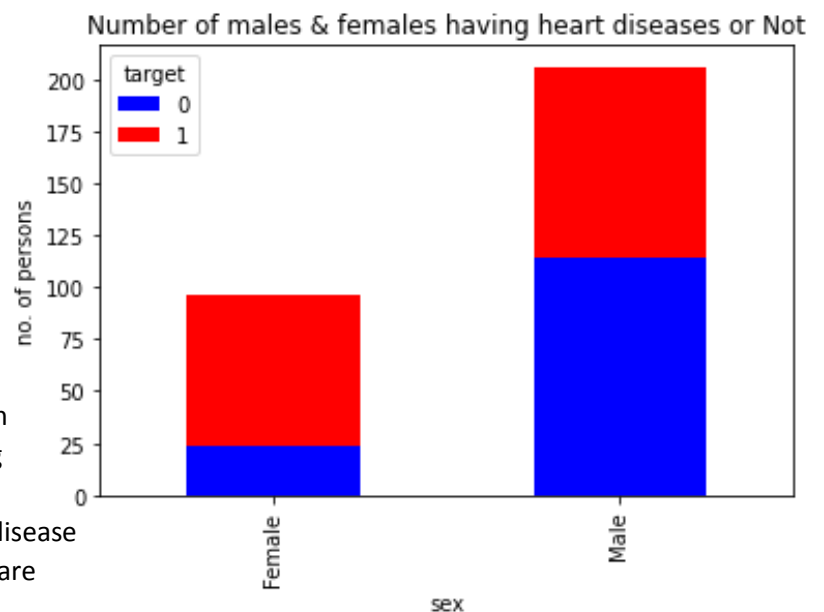
#### OBSERVATIONS:

\*) Females are less likely to fall under the category of having heart disease. The exact number of females with healthy heart is 72 whereas the females with heart disease are only 24 in number.

\*) Males are more in number in considering the factor of having heart disease. To be exact, 114 males are suffered from heart disease whereas 92 among total males are those with a healthy heart.

\*) According to this dataset, 92 males and 72 females have no heart disease, which means men exceeds total 20 in number in terms of having a healthy heart.

\*) There are 114 males and 24 females who have heart disease. This concludes that males are more in number who have heart disease than females.



### TASK 3 : DATA MODELLING

#### 1. DATA PRE-PROCESSING :

\*) The categorical columns who has string values has been converted to convert all the attributes of the data set into numerical terms.

1. The attribute sex ,which initially had values 'male' and 'female' are converted as 1 and 0 respectively.

2. The attribute fbs ,which initially had values 'yes' and 'no' are converted as 1 and 0 respectively.

3. The command for converting these two attributes is as follows:

```
df['sex'].replace( 'Female',0 , inplace= True)
```

```
df['sex'].replace( 'Male',1, inplace= True)
```

```
df['fbs'].replace( 'yes',1, inplace= True)
```

```
df['fbs'].replace( 'no',0, inplace= True)
```

\*) The target attribute has value 0 as heart disease and 1 as no heart disease. I interchanged the values as 0 for no heart disease and 1 as heart disease, for better interpretation.

\*) Standardization (a.k.a. Z-score normalisation) has been used to re-scale some features to have the mean value of 0 and the standard deviation of 1.

The features which undergo standardization are ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']. The command used to do this step is as follows :

```
columns_to_scale=['age', 'trestbps', 'chol', 'thalach', 'oldpeak']
```

```
df[columns_to_scale] = StandardScaler.fit_transform(df[columns_to_scale])
```

2. **SPLITTING THE DATA** : My aim is to detect that whether the person have heart disease or a healthy heart . For that, the feature “target” has been selected using the command given below:

```
x= df.drop(['target'],axis=1)
```

```
y= df['target']
```

After this ,the data has been split into three suites.

Suite1: 50% for training and 50% for testing

Suite2: 60% for training and 40% for testing

Suite3: 80% for training and 20% for testing

3. **DATA MODELING AND EXECUTION** : I have chosen Logistic Regression and Decision Tree classifier algorithms for experiments. Both algorithms have given varying results in all the three suites. The best accuracy is obtained by using Logistic Regression algorithm , which is exactly 90.2% accuracy in the third suite of dataset.

**LOGISTIC REGRESSION** : It is a supervised learning that computes the probabilities of a categorical dependent variable. The dependent variable is a binary variable that contains data coded as 1(yes, success, etc.) or no (no, failure, etc.).

**DECISION TREE CLASSIFIER** : It is a supervised machine learning method. Decision trees are popular tool in decision analysis.

**NOTE:** The accuracy results using this algorithm fluctuates in every run in python, by range of 2-4 % as the random state is picked up by the model automatically in every suite. We cannot assign random\_state=0 to get the same result in every run because I have to compare the performance of the model in every suite with different ratios of training and testing set.

**Explanation for all the three suites has been given below using two models:**

**\*) SUITE 1 : : 50% for training and 50% for testing**

The size of the dataset used in this suite comes out to be as:

x\_train 1963

x\_test 1976

y\_train 151

y\_test 152

The size of training and testing set are approximately same because 50% of data is used for training and 50% is used for testing.

**Logistic Regression :**

In this model, model1 is developed based on 50% of training data and then it is used to predict the response of the data. To check the accuracy , the prediction is made on y\_test and the confusion matrix comes out as 75 cases true positive that means 75 cases are rightly predicted by the algorithm ,54 cases are true negative, 6 cases are false positive and 17 are false negative that means they are falsely predicted cases by algorithm . **The testing accuracy in this case is 84.86%.**

Classification report for this model is as follows:

	precision	recall	f1-score	support
0	0.82	0.93	0.87	81
1	0.90	0.76	0.82	71
accuracy			0.85	152
macro avg	0.86	0.84	0.85	152
weighted avg	0.85	0.85	0.85	152

**Decision Tree classifier:** Similarly, as we did the prediction on suite 1 using Logistic Regression algorithm. The prediction is made on this model as well. **The accuracy comes out to be 74.34%**, which is less than the model 1 (using Logistic Regression). The testing accuracy using confusion matrix is calculated as follows:

**TP=cmatrix[0][0]**

**TN=cmatrix[1][1]**

**FN=cmatrix[1][0]**

**FP=cmatrix[0][1]**

**print('Testing Accuracy:', (TP+TN)/(TP+TN+FN+FP))**

The result is **0.743421052631579**, which means 74.34%. Classification report of this model is as follows:

	precision	recall	f1-score	support
0	0.76	0.77	0.76	81
1	0.73	0.72	0.72	71
accuracy			0.74	152
macro avg	0.74	0.74	0.74	152
weighted avg	0.74	0.74	0.74	152

**CONCLUSIONS FOR SUITE 1 :** The precision in model 1 using Logistic Regression algorithm for 0 and 1 is 0.82 and 0.92 resp., which means 82% and 92% cases are rightly predicted by the algorithm. While, in model 2, using Decision tree classifier algorithm, the precision is 76% and 73%, which is low as compared to the model 1.

#### **\*) SUITE 2 : 60% for training and 40% for testing**

The size of the dataset used in this suite comes out to be as:

x\_train 2353

x\_test 1586

y\_train 181

y\_test 122

The size of training and testing set differs because 60% of data is used for training and 40% is used for testing.

#### **Logistic Regression :**

In this model, model4 is developed based on 60% of training data and then it is used to predict the response of the data. To check the accuracy, the prediction is made on y\_test and the confusion matrix comes out as 61 cases true positive that means 61 cases are rightly predicted by the algorithm, 46 cases are true negative, 4 cases

are false positive and 11 are false negative that means they are falsely predicted cases by algorithm .

**The testing accuracy in this case is 87.7%.**

Classification report for this model is as follows:

	precision	recall	f1-score	support
0	0.85	0.94	0.89	65
1	0.92	0.81	0.86	57
accuracy			0.88	122
macro avg	0.88	0.87	0.88	122
weighted avg	0.88	0.88	0.88	122

**Decision Tree classifier:** Similarly, as we did the prediction on suite 2 using Logistic Regression algorithm. The prediction is made on this model as well. **The accuracy comes out to be 70.5% , which less than the model 4(using Logistic Regression).**

Classification report for this model is as follows:

	precision	recall	f1-score	support
0	0.75	0.68	0.71	65
1	0.67	0.74	0.70	57
accuracy			0.70	122
macro avg	0.71	0.71	0.70	122
weighted avg	0.71	0.70	0.71	122

**CONCLUSIONS FOR SUITE 2 :** The precision in model 4 using Logistic Regression algorithm for 0 and 1 is 0.85 and 0.92 resp., which ,means 85% and 92% cases are rightly predicted by the algorithm. While , in model 5, using Decision tree classifier algorithm , the precision is 75% and 67% , which is low as compared to the model 4.

#### **\*) SUITE 3 : 80% for training and 20% for testing**

The size of the dataset used in this suite comes out to be as:

x\_train 3146

x\_test 793

y\_train 242

y\_test 61

The size of training and testing set differs on a big margin because 80% of data is used for training and 20% is used for testing.

#### **Logistic Regression :**

In this model, model6 is developed based on 80% of training data and then it is used to predict the response of the data. To check the accuracy , the prediction is made on y\_test and the confusion matrix comes out as 34 cases true positive that means 34 cases are rightly predicted by the algorithm ,21 cases are true negative, 1 case is false positive and 5 are false negative that means they are falsely predicted cases by algorithm .

**The testing accuracy in this case is 90.1%.**

Classification report for this model is as follows:

	precision	recall	f1-score	support
0	0.87	0.97	0.92	35
1	0.95	0.81	0.88	26
accuracy			0.90	61
macro avg	0.91	0.89	0.90	61
weighted avg	0.91	0.90	0.90	61

**Decision Tree classifier:** Again, as we did the prediction on suite 3 using Logistic Regression algorithm. The prediction is made on this model as well. **The accuracy comes out to be 77.1%**, which is less than the model 4 (using Logistic Regression). Classification report for this model is as follows:

	precision	recall	f1-score	support
0	0.84	0.74	0.79	35
1	0.70	0.81	0.75	26
accuracy			0.77	61
macro avg	0.77	0.78	0.77	61
weighted avg	0.78	0.77	0.77	61

**CONCLUSIONS FOR SUITE 3 :** The precision in model 6 using Logistic Regression algorithm for 0 and 1 is 0.85 and 0.92 resp., which means 85% and 92% cases are rightly predicted by the algorithm. While, in model 7, using Decision tree classifier algorithm, the precision is 84% and 70%, which is low as compared to the model 6.

### COMPARISON OF PERFORMANCE OF TWO MODELS IN EACH SUITE :

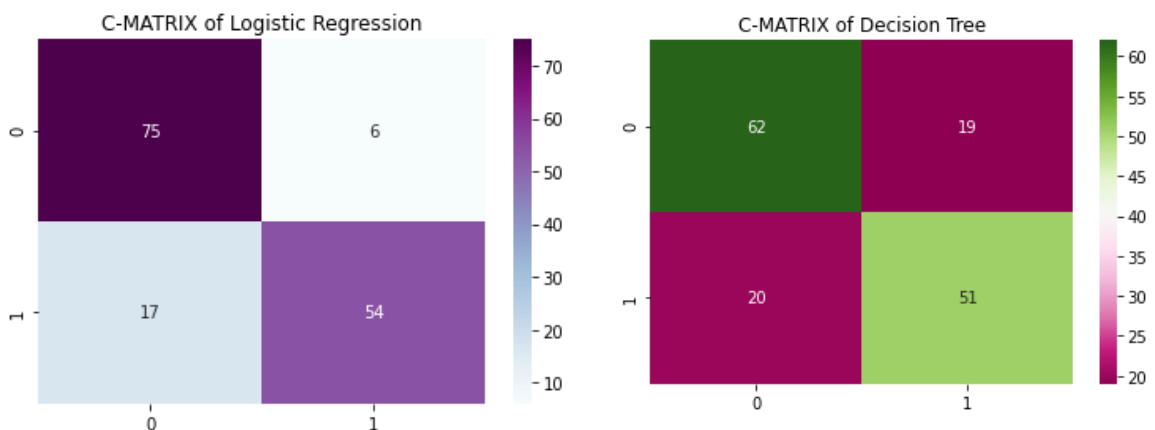
The accuracy of each model in every suite is described below in a table:

	SUITE 1	SUITE 2	SUITE 3
<b>LOGISTIC REG.</b>	<b>84.86%</b>	<b>87.7%</b>	<b>90.1%</b>
<b>DECISION TREE</b>	<b>74.34%</b>	<b>70.5%</b>	<b>77.1%</b>

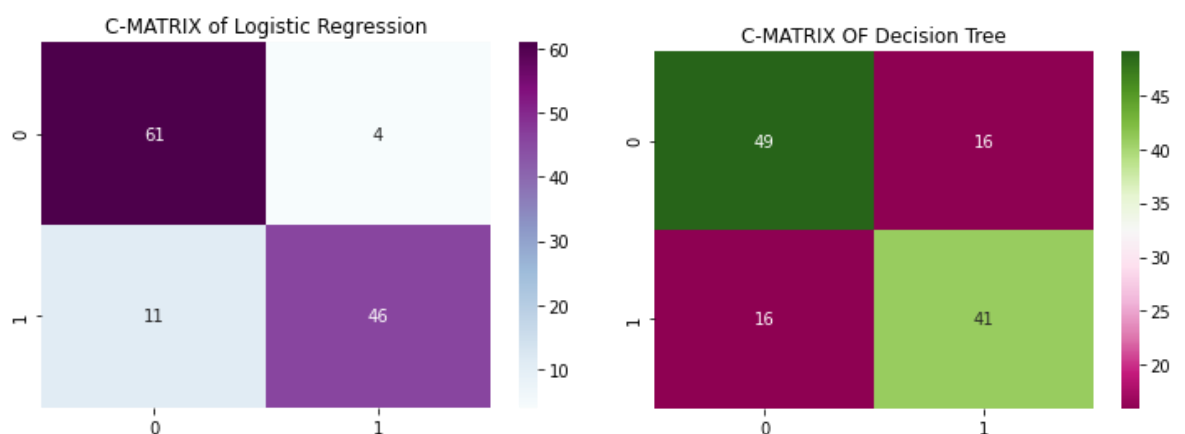
It can be concluded that Logistic regression algorithm has given the best accuracy in each suite and the best accuracy is calculated as 90.1% in the third suite, where 80% of data was used for training and 20% for testing.

### COMPARISON USING GRAPHICAL VISUALISATIONS :

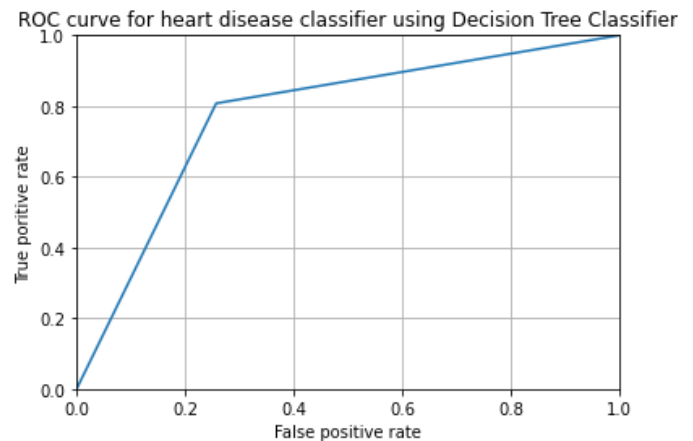
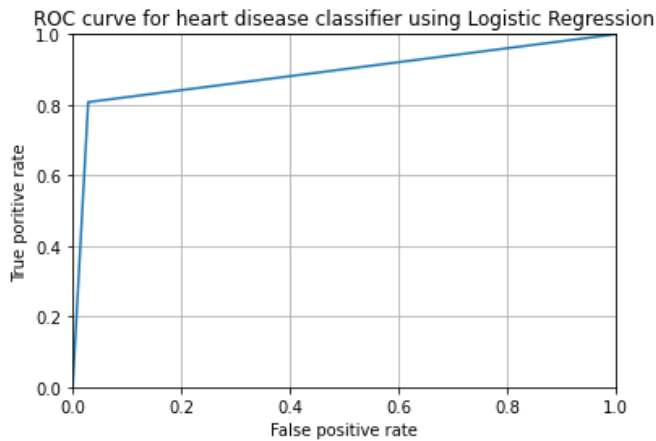
\*) **SUITE 1 :** The heatmaps for both the models in this suite are given below.



\*) **SUITE 2 :** The heatmaps for both the models in this suite are given below.



**\*) SUITE 3 :** The ROC curves for two model in this suite is given below.



### **DISCUSSIONS AND CONCLUSIONS :**

**\*)** In the dataset ,there are many important features to be considered ,which plays role in the person's life to have a healthy heart. These important features are chest pain type, fasting blood sugar level, thal value( blood disorder ) , cholesterol, ECG results and thalach (max heart rate achieved).

**\*)** Exploration of the dataset and the relationship among features, tells us that men are more likely to be suffered to have heart disease than women.

**\*)** Most of the people with heart disease , also have fasting blood sugar level >120 md/dl.

**\*)** Logistic regression algorithm gives the best accuracy of 90.1% to detect that whether the person have heart disease or have healthy heart. The reason it outperforms decision tree classifier algorithm is that decision tree bisects the space into smaller and smaller region.

---

### **REFERENCES :**

**\*)** <https://www.kaggle.com/datasets?datasetsOnly=true>

**\*)** <https://www.kaggle.com/ronitf/heart-disease-uci>

**\*)** <https://www.kaggle.com/rajeshjnv/heart-disease-exploration-ml-prediction>

---