

# **HEALTHIER HEART TODAY... FOR A SAFER TOMORROW**

**Predicting early signs of Heart Diseases**

**IDS 400 : Introduction to Data Science – Python 3.2**

**Janhavi Powale – UIN 664270476**

**Trang Tran – UIN 655813074**

**Danbi Kim- UIN 668022049**

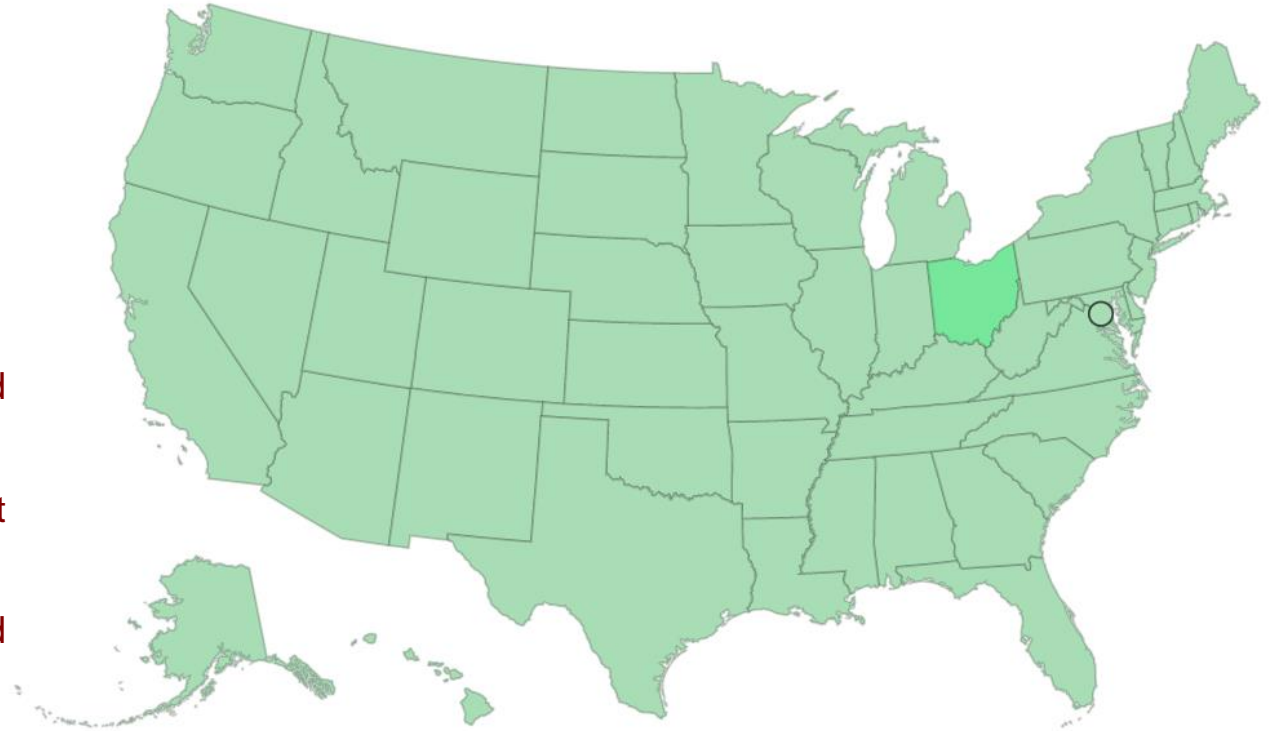
**Ashok Bhatraju –UIN 670248723**

**Karansin Raj – UIN 66797989**

# WHAT IS THE ISSUE WE ARE FACING?

## Heart Diseases in USA...

- Take away 1 life every 37 seconds
- Lead cause of death for Men and Women across all ethnicities
- Are responsible for 25% deaths in USA
- Costs up to \$219 billion on health services, medicines and lost productivity
- Result commonly in Coronary Artery Diseases and Heart Attacks
- Have 3 key risk factors - High blood pressure, high blood cholesterol and smoking



**...a grave concern for 50% Americans**

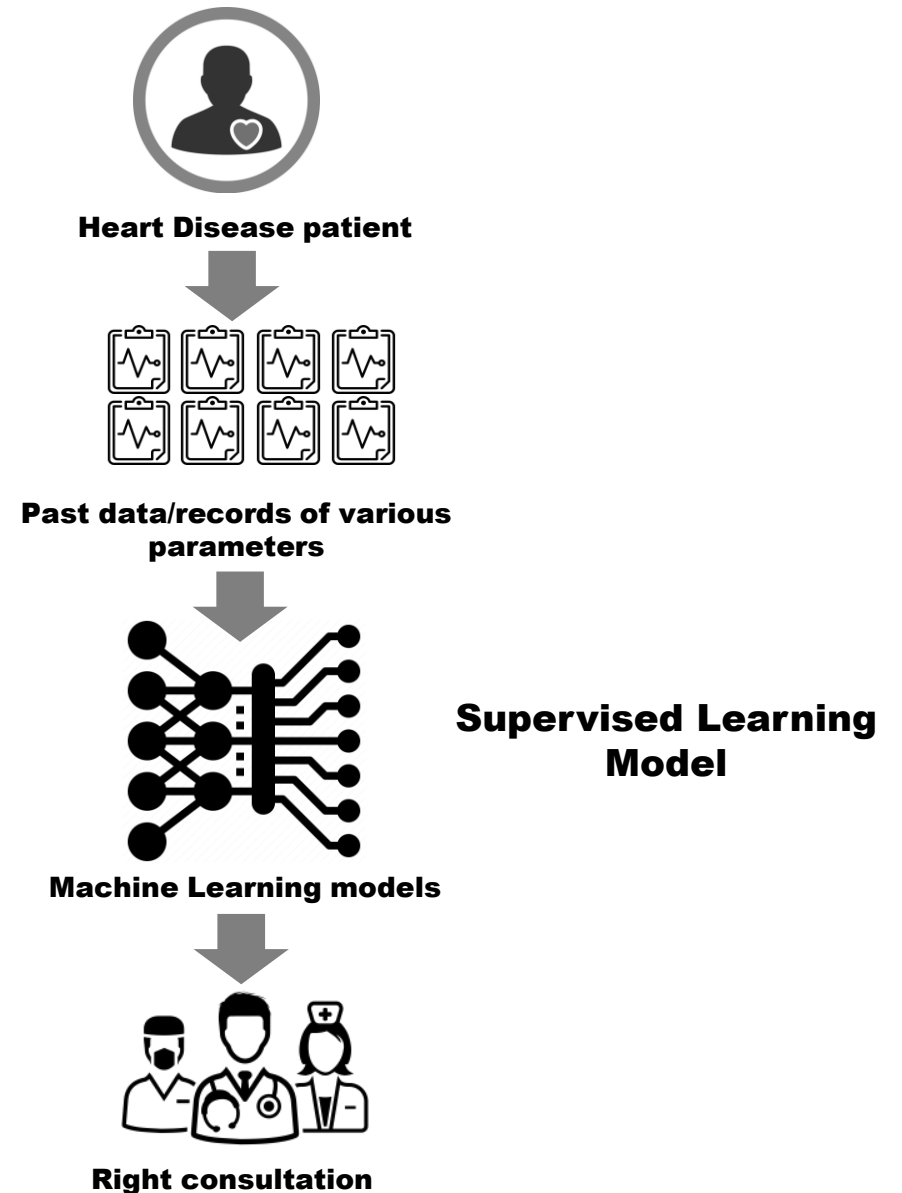
# HOW TO PREEMPT HEART DISEASES?

## Challenges:

- Medical systems are often challenged with distribution of senior clinicians, higher rate of mis-diagnosis , longer training period of staff
- Mis-diagnosis can lead to higher expensive for end consumers, a Type II can be even worse

## Solution:

- Machine learning is more accurate than human medical professionals in predicting vulnerability among patients suspected of heart diseases
- Supervised learning algorithms in machine learning can train the past data to classify individuals who are post prone to heart diseases or not
- With this goal, we can get the right medical advise at the right time with no human intervention
- It can improve accuracy of operations and diagnostic efficiencies




# ABOUT CLEVELAND's DATASET



- Dataset from UCI – ML Repository
- 76 attributes but published experiments use only 14 , 303 rows
- Individuals are grouped in 5 types – 0 for absence for disease. 1,2,3 and 4 for presence of a disease
- The variables consist of five continuous and eight discrete attribute
- Discrete values are majorly categorical variables. Eg Thalassemia, Sex, Chest-pain type etc
- There are 6 missing values – They in the form of ‘?’ It was tough to spot because it did not show in missing values

Sr No	Column Name	Meaning
1	Age	Shows age of the individual
2	Sex	Shows binary gender of the individual 1= male , 0 =Female
3	Chest-pain type	Shows type of chest pain of 4 types
4	Resting blood pressure	Shows resting BP in mmHg units
5	Serum Cholestrol	Shows serum cholesterol in mg/dl
6	Fasting blood sugar	Compates sugar to baseline of 120mg/dl >120 = 1, <120 mg=0
7	Resting ECG	Shows resting ECG of 3 types 0=normal, 2= LVH, 3= ST-T abnormal
8	Maximum heart rate achieved	Shows Maximum heart rate achieved
9	Exercise induced angina	Binary values ,1= yes, 0=No
10	ST depression induced by exercise relative to rest	Done during stress test
11	Peak exercise ST segment	1= upslope, 2=flat and = downslope
12	No of major vessels (0–3) colored by flouroscopy	Shows value of major
13	Thalassemia	Shows 3 types of Thalassemia
14	Diagnosis of heart disease	Response variable ,0= Absence 1,2,3,4= Presence of heart disease

# PROJECT APPROACH



## Exploratory Data Analysis

- This will be primarily analysis and data visualization on different attributes – Age, Sex etc to see effect on target variable
- We have used seaborn, matplotlib, Sklearn
- Checking class balance
- Imputing missing values with median values
- Checking correlation plots

## Data Transformation

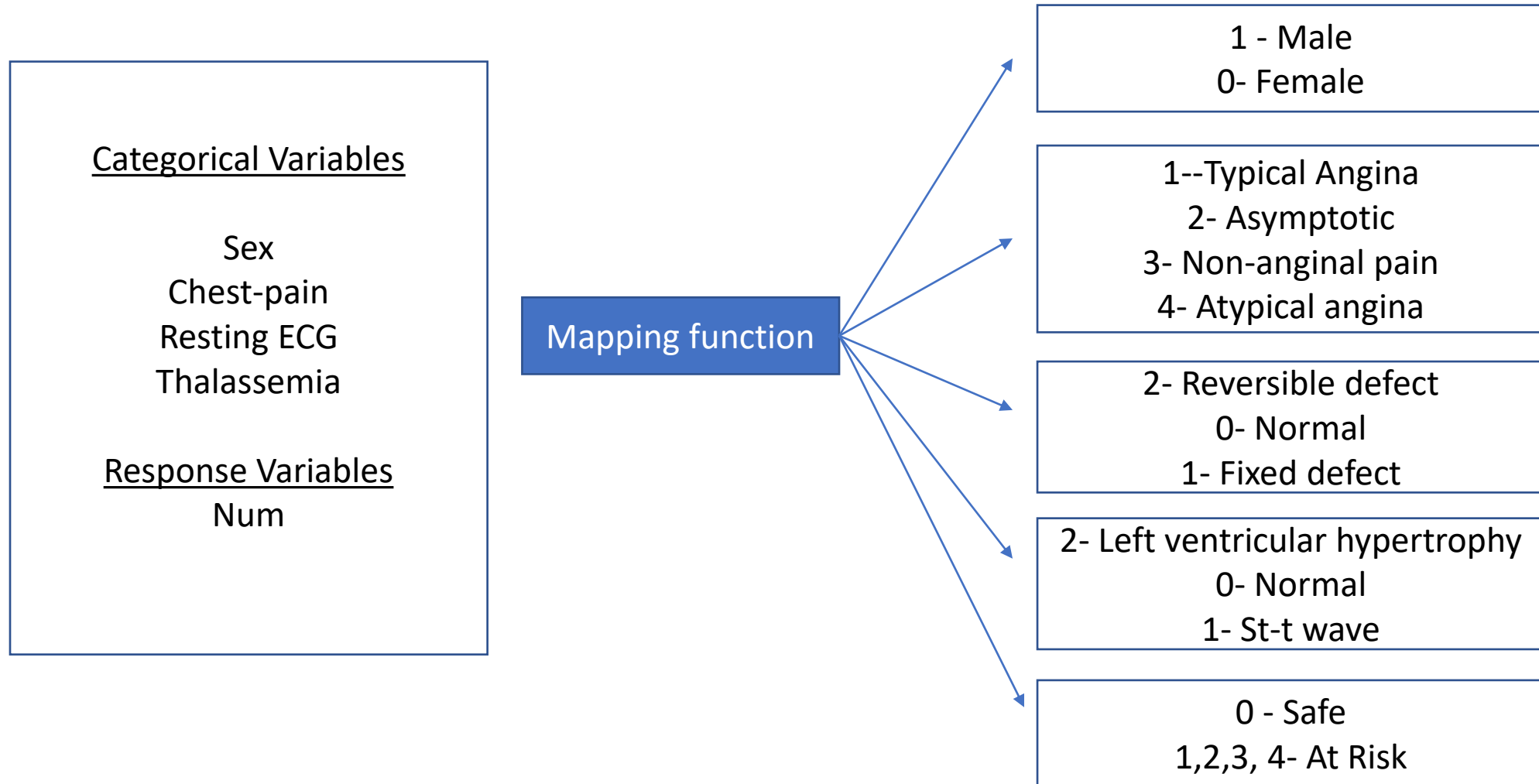
- Primarily used dummies to convert categorical values to one-hot coding
- Transformed skewness with log transformation to follow normal distribution with histogram plots
- We used libraries such as scipy

## Developing ML Models

- Train-Test data split approach
- 3 models to predict if a person has heart disease or not
  - Logistic Regression
  - KNN
  - Decision Trees
- Conducted hyper parameter tuning and used cross-validation to determine the best K value for KNN

# EXPLORATORY DATA ANALYSIS

## DATA MAPPING

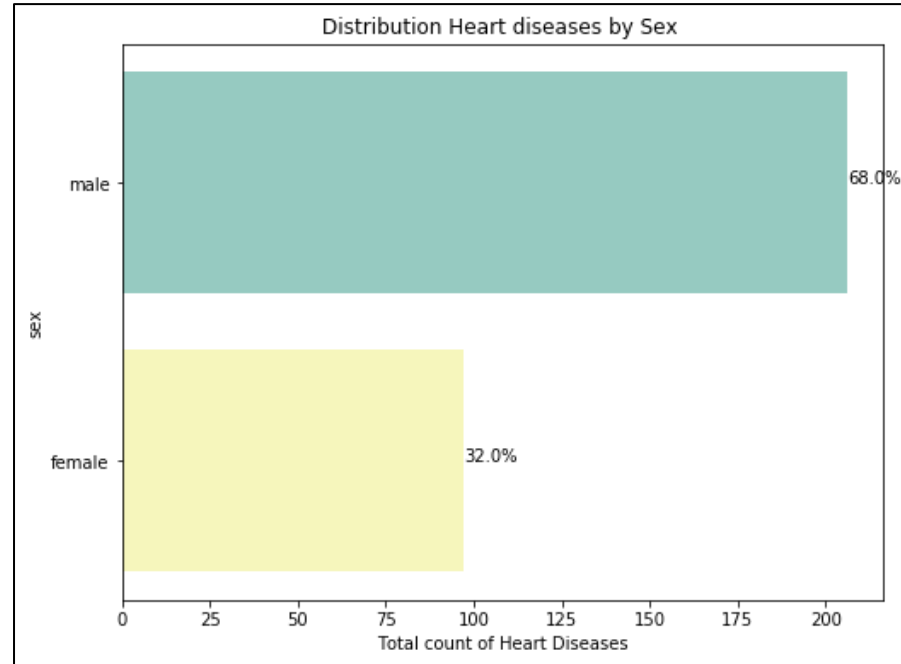
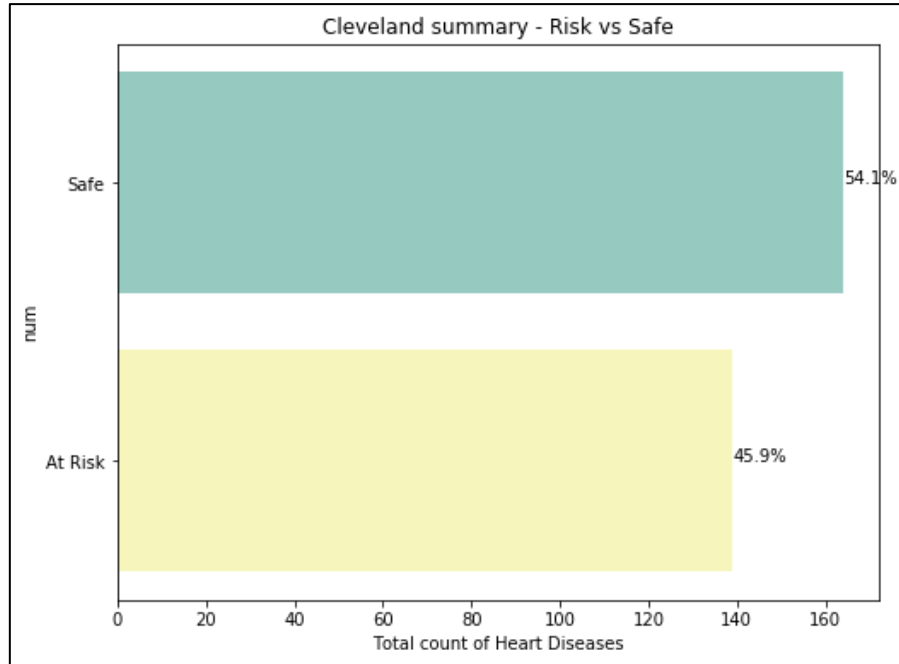


# EXPLORATORY DATA ANALYSIS

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	num
count	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00
mean	54.44	0.68	3.16	131.69	246.69	0.15	0.99	149.61	0.33	1.04	1.60	0.94
std	9.04	0.47	0.96	17.60	51.78	0.36	0.99	22.88	0.47	1.16	0.62	1.23
min	29.00	0.00	1.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	1.00	0.00
25%	48.00	0.00	3.00	120.00	211.00	0.00	0.00	133.50	0.00	0.00	1.00	0.00
50%	56.00	1.00	3.00	130.00	241.00	0.00	1.00	153.00	0.00	0.80	2.00	0.00
75%	61.00	1.00	4.00	140.00	275.00	0.00	2.00	166.00	1.00	1.60	2.00	2.00
max	77.00	1.00	4.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	3.00	4.00

- Summary statistics reveals that data is from 29 yrs to 77 yrs
- Cholesterol ranges from 126 to 564.
- Resting blood pressure value ranges from 94 to 200, which is normal only if one is much older.

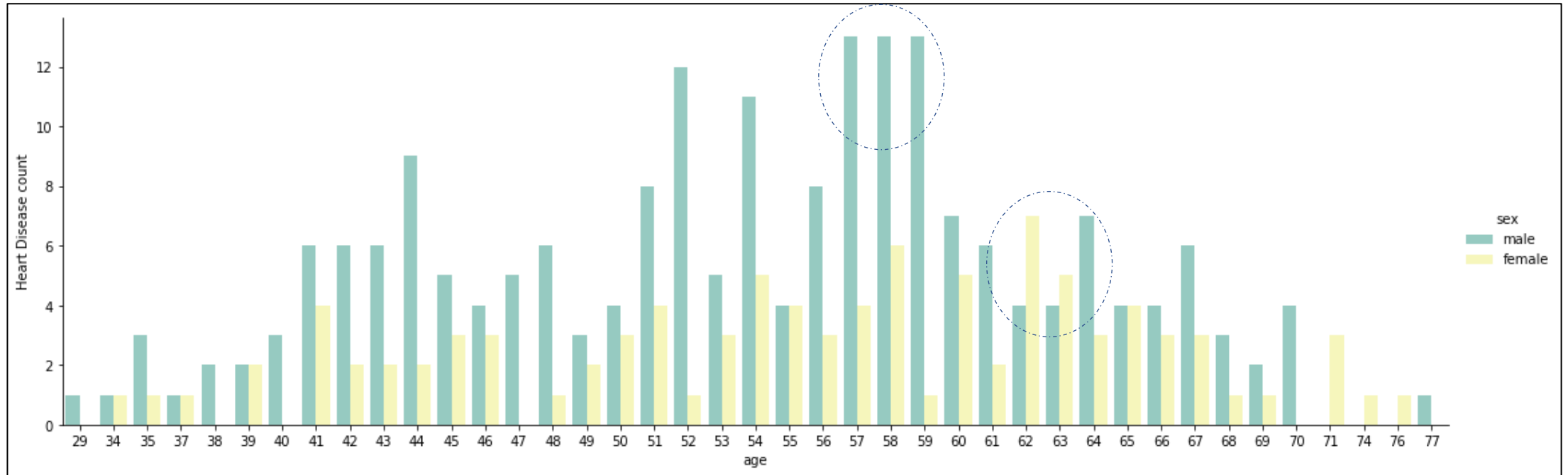
# EXPLORATORY DATA ANALYSIS



- A simple histogram chart reveals that At Risk = 46% , Safe = 54% , thus classes are balanced
- Another bar chart tells that Males are almost twice at risk than females for a heart disease . Males = 68% , Female = 32%
- This is a very alarming and immediately calls for caution for Men. There could be several reasons – improper eating , irregular exercises or overtly alcoholic habits

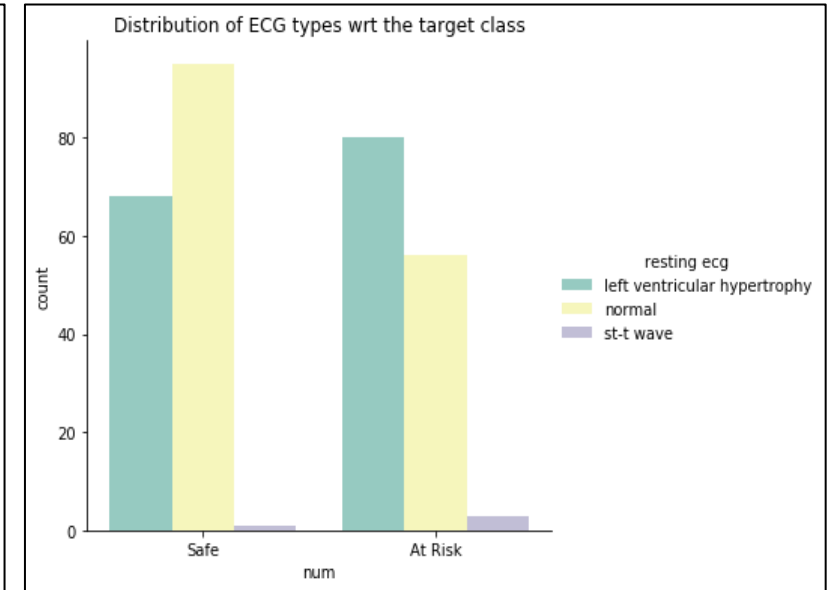
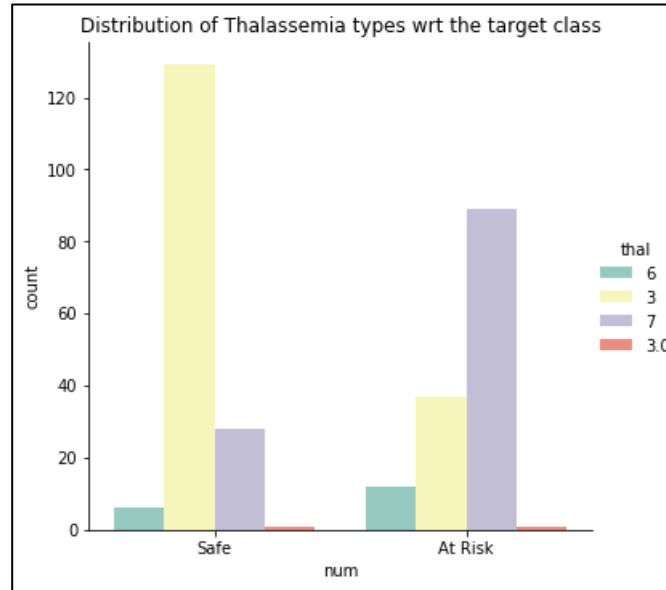
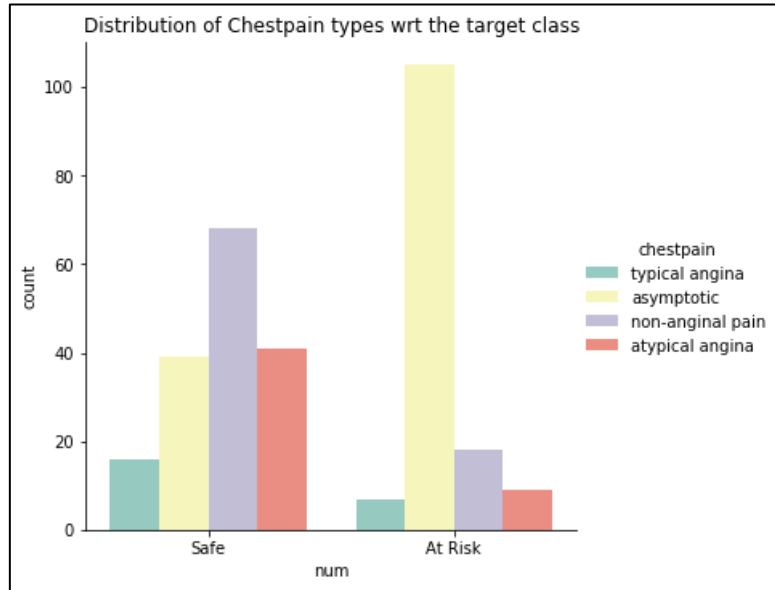


# EXPLORATORY DATA ANALYSIS



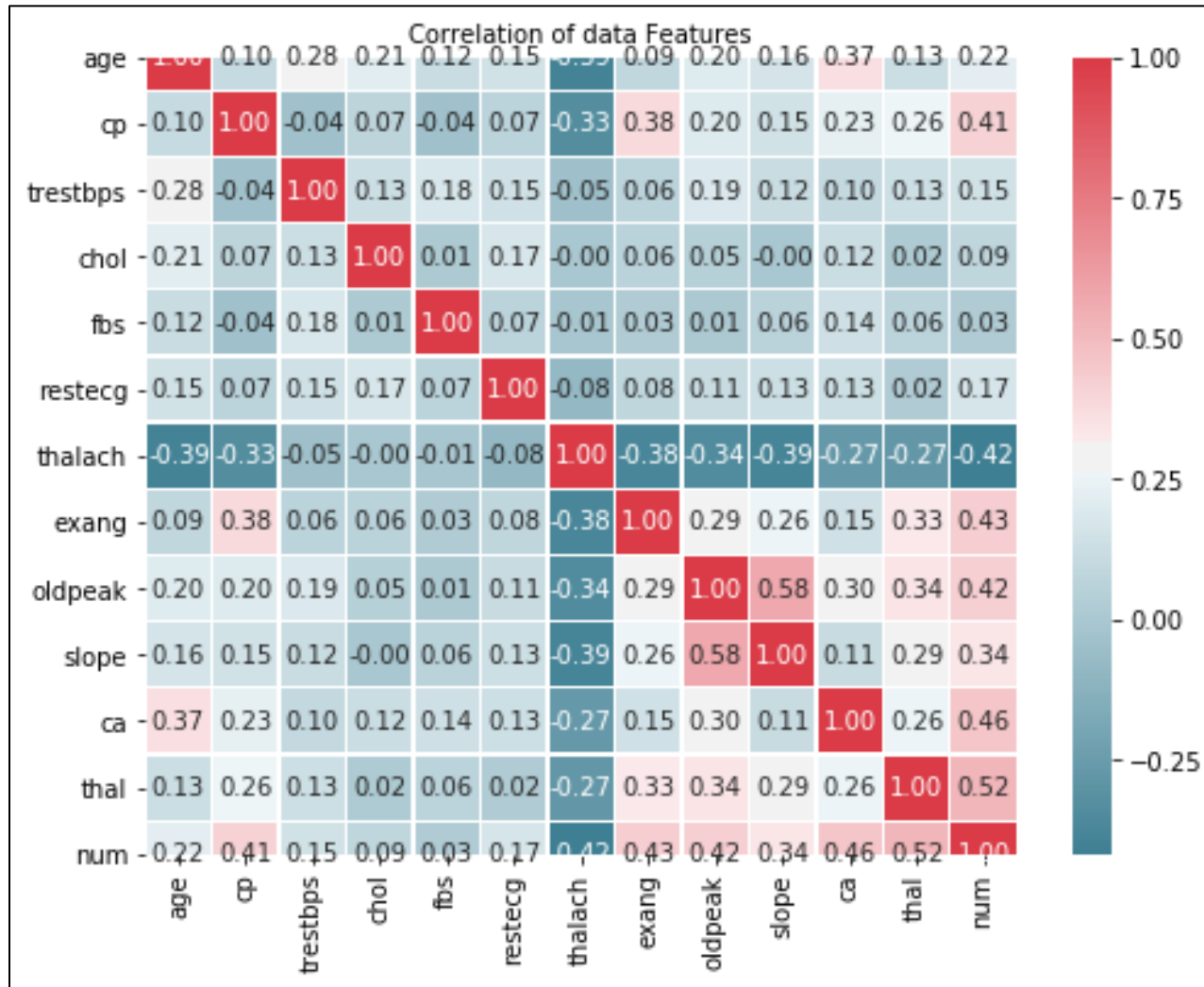
- Looking at plot for Ages vs Sex , it is clear that
  - 56 to 60 is the most vulnerable age for a heart disease.
  - Females are most vulnerable at ages 62-63
  - Males even younger at 55-56

# EXPLORATORY DATA ANALYSIS



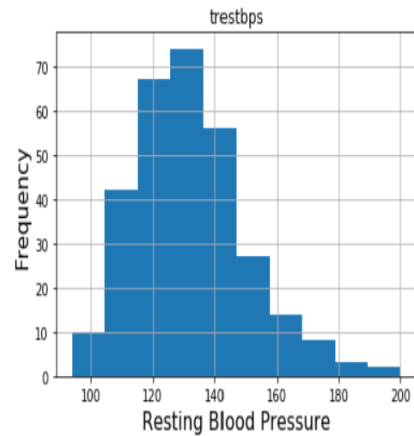
- However, thalassemia shows that having type 7 thalassemia has higher chances of being at risk
- Likewise, having asymptotic chest-pain seems as the most common reason behind causing heart diseases
- Left ventricular hypertrophy meaning damage to the left ventricle of the heart increases the chances of the heart attack.

# EXPLORATORY DATA ANALYSIS

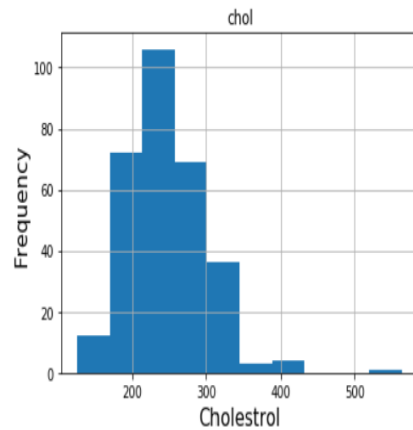


- Correlation plot does not show very high values with the target variables and attributes with each other
- Thus, there is no issue of multicollinearity or confounding variables.
- The highest correlation coefficient is seen with 'Thal' or thalassemia and ca or no of major vessels filled by fluoroscopy, the higher no of vessel, higher the chance of a heart disease

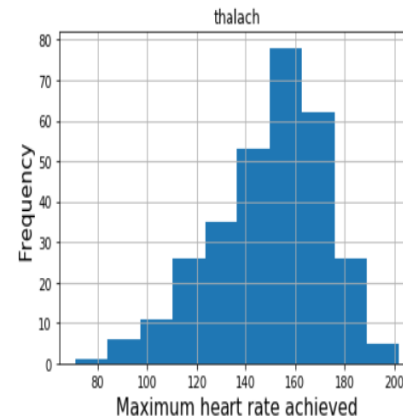
# DATA TRANSFORMATION



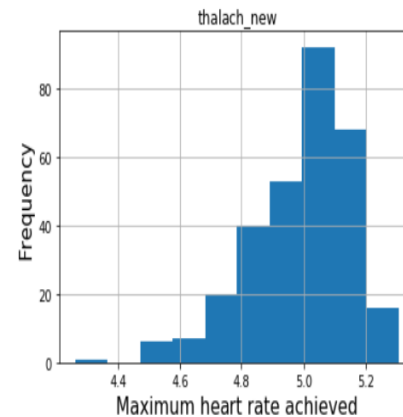
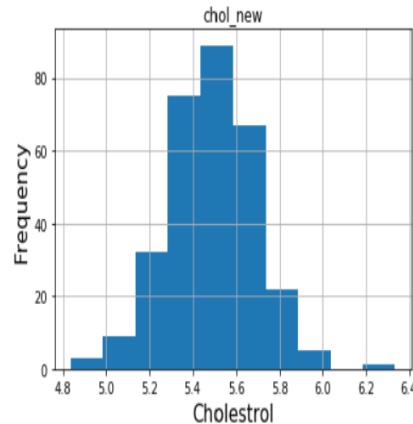
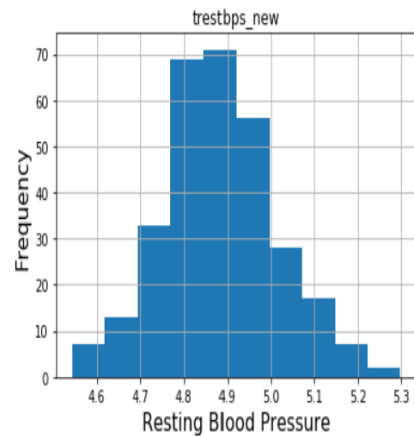
<Figure size 1224x720 with 0 Axes>



<Figure size 1224x720 with 0 Axes>



<Figure size 1224x720 with 0 Axes>



We did skewness test for 3 quantitate variables – resting blood pressure , cholesterol and fasting blood sugar

- All the variables showed strong skewness both positive and negative.

- Skewness hampers prediction and it is required to correct it

- We applied log transformation to all three and achieved near normal distributions.

- The along side figures show before-after histogram plots

- All the categorical variables were converted to one-hot coding eg

# MACHINE LEARNING MODELS

## DECISION TREES

Decision Tree is complex version of “if-then” algorithms. Based on the number of different predictors, this algorithm will generate various “if-then” rules and will split the data in top to bottom direction. The accuracy and model performance depends on amount of split and the purity of those splits.

```
[385] #create decision tree object
      clf=DecisionTreeClassifier(criterion= 'gini', max_depth=4,min_samples_split=12,random_state=0)

      #Train decision tree classifier
      clf=clf.fit(Xtrain,ytrain)
```

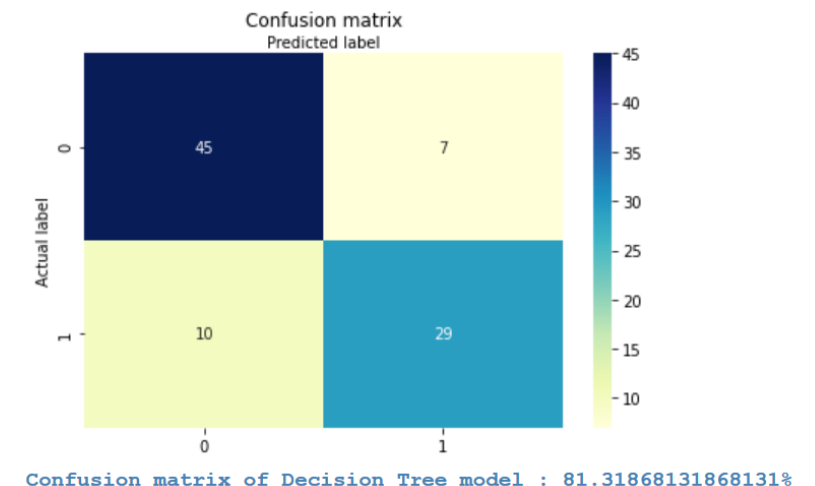
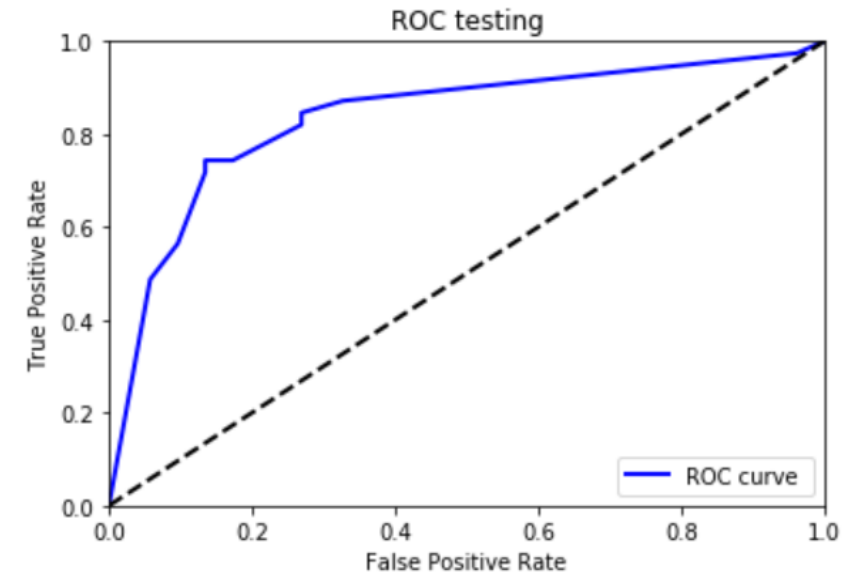
As discussed previously, we’ve used ‘Gini Index’ as the purity measure, max depth of 4, and minimum sample split of 12 for creating the DT.

The DT is known for its best performance when de-correlation data is fed. That is the main result which has resulted into best AUC of 95% for this algorithm.

Further, this model maintains the accuracy of 81% during the true class prediction.

However, for our case Recall is rather more important parameter for model goodness measure. Considering the higher model complexity with DT we’re getting only 74.35% Recall.

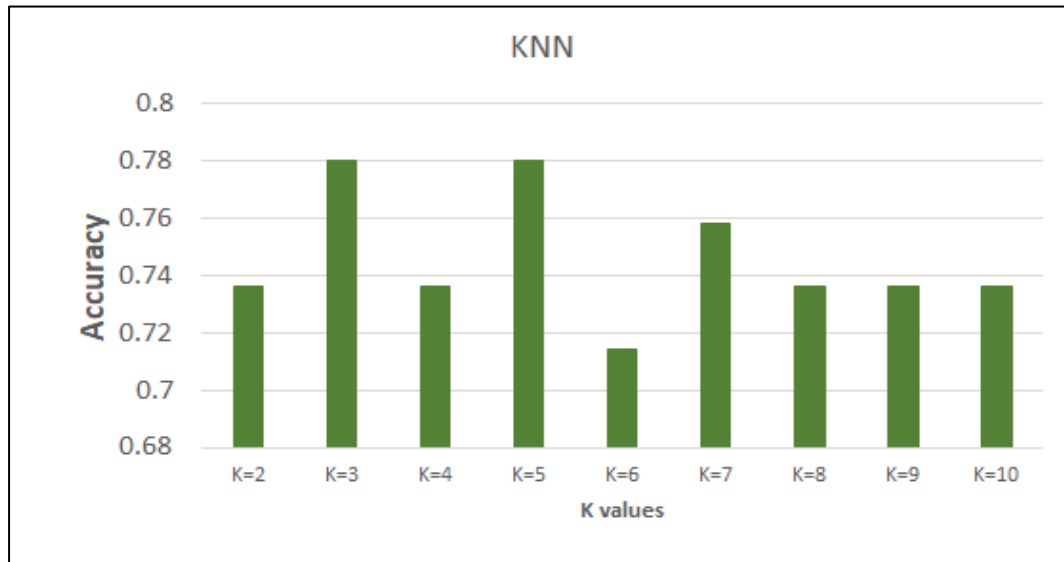
AUC: 0.8360453648915188



# MACHINE LEARNING MODELS

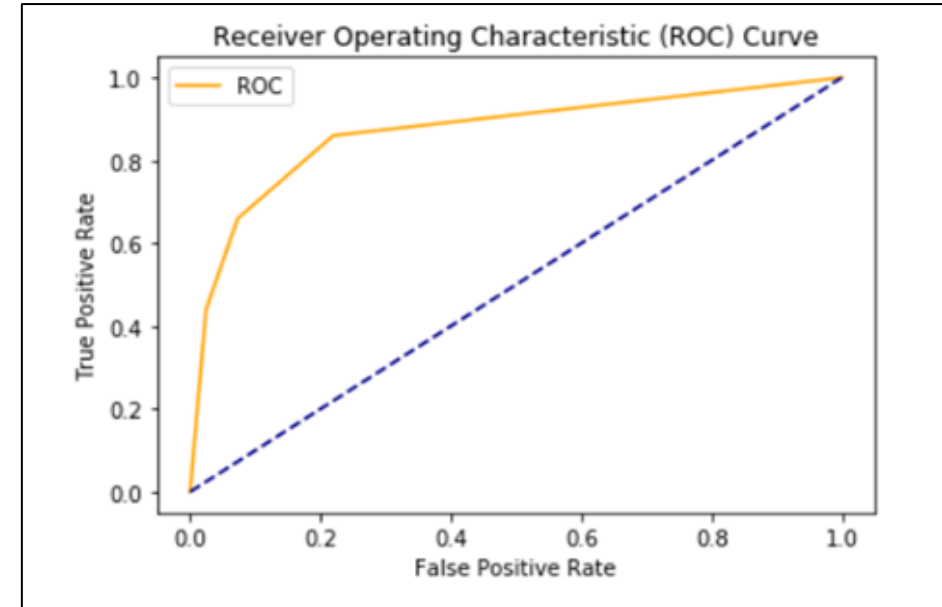
## KNN

K nearest neighbour divides the existing data into clusters based on the K values. It groups the existing data into a particular cluster based on the similarities among the data. We can use these clusters to predict the outcome based on distance from these clusters.



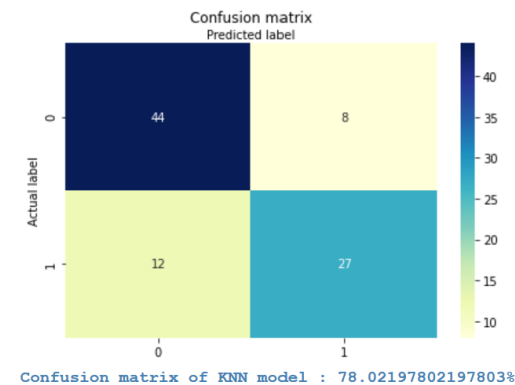
We used multiple K values to check which value gives us the best accuracy. In our case K = 3 and 5 gives the best possible accuracy for this model. **Accuracy = 0.78**

Recall is 69% for class 1



ROC curve explains how good our model is. The higher the area under the curve, the better the model.

The AUC for the ROC curve is 81%



# MACHINE LEARNING MODELS

## LOGISTIC REGRESSION

Considering the reduced model complexity and higher model interpretability with the categorical target variable, Logistic Regression approach was also utilized to check the model performance for our data.

**This approach has allowed us to get the most optimum model with best model accuracy of 86% in true class prediction along with the best possible Recall of 86.36%.**

Also the AUC for this model is far above acceptable range.

```
[413] #Building a ml model

from sklearn.datasets import make_classification

x=hdc.iloc[:, 0: 13].values
y=hdc.iloc[:,13].values

#testing
x, y = make_classification(n_samples=1000, n_classes=2, random_state=1)

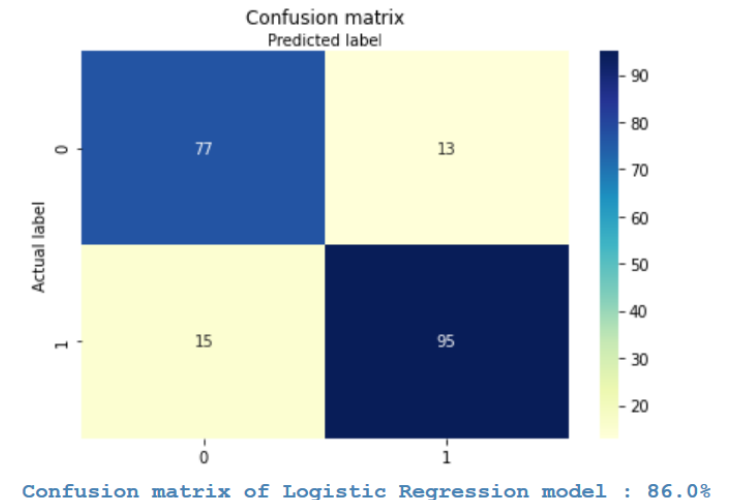
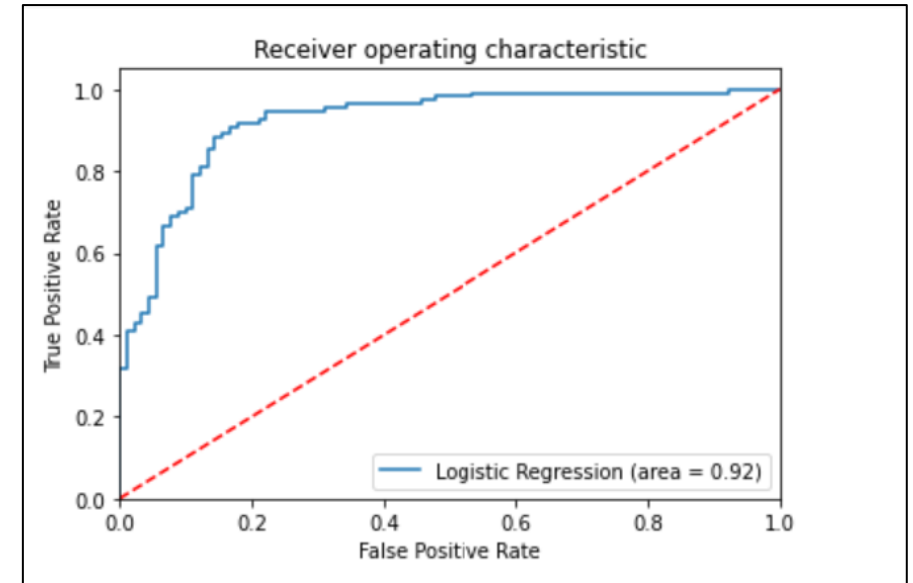
from sklearn.model_selection import train_test_split

x_trn, x_tst, y_trn, y_tst = train_test_split(
    x, y, test_size=0.2, random_state=1)

from sklearn.preprocessing import StandardScaler

sc_x= StandardScaler()
x_trn = sc_x.fit_transform(x_trn)
x_tst = sc_x.transform(x_tst)

y_pred = classifier.predict(x_tst)
y_pred_trn = classifier.predict(x_trn)
```



# COMPARISON AND RESULT

Model	Recall Percentage
Decision Tree	74%
KNN	69%
Log Regression	86%

Model	AUC
Decision Tree	84%
KNN	81%
Log Regression	92%

Model	Accuracy
Decision Tree	81%
KNN	78%
Log Regression	86%

From above comparison we've found that the best model for our case is Logistic-Regression. This result is based on the "Recall" measurement of model prediction, which is the model performance to predict positive class perfectly (class that is having heart disease).

There are two major benefits of using LR over other 3 machine learning model, viz:

1. Model Interpretability: Amongst all the applied models, LR is the one with the highest statistical interpretability.
2. Lowest Model Complexity: Which will facilitate in lower probability of Overfitting.



**THANK YOU**