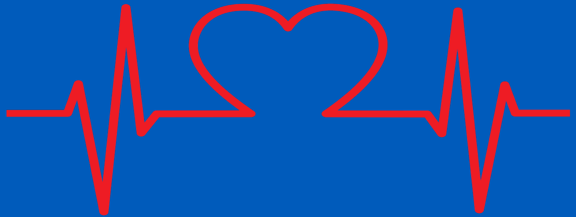
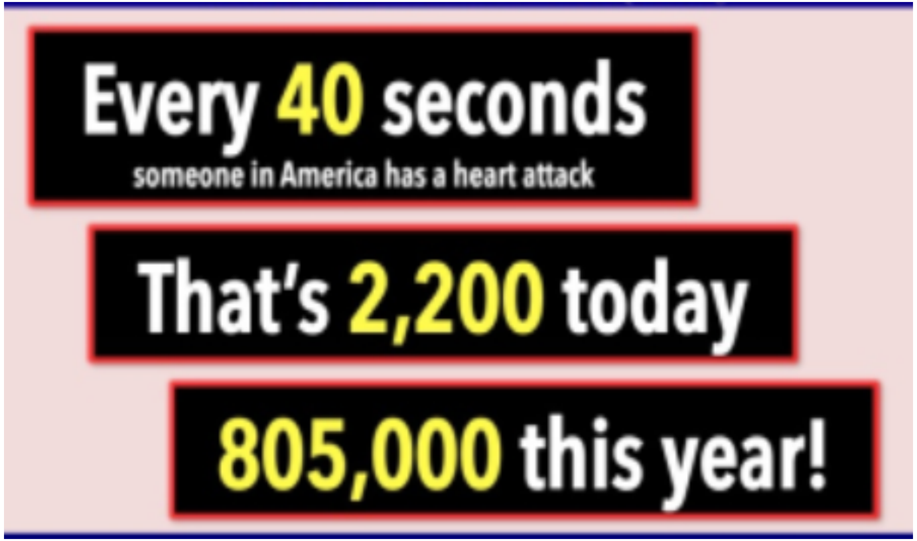
EAS 508
Heart Attack Prediction



University at Buffalo The State University of New York

Problem Statement

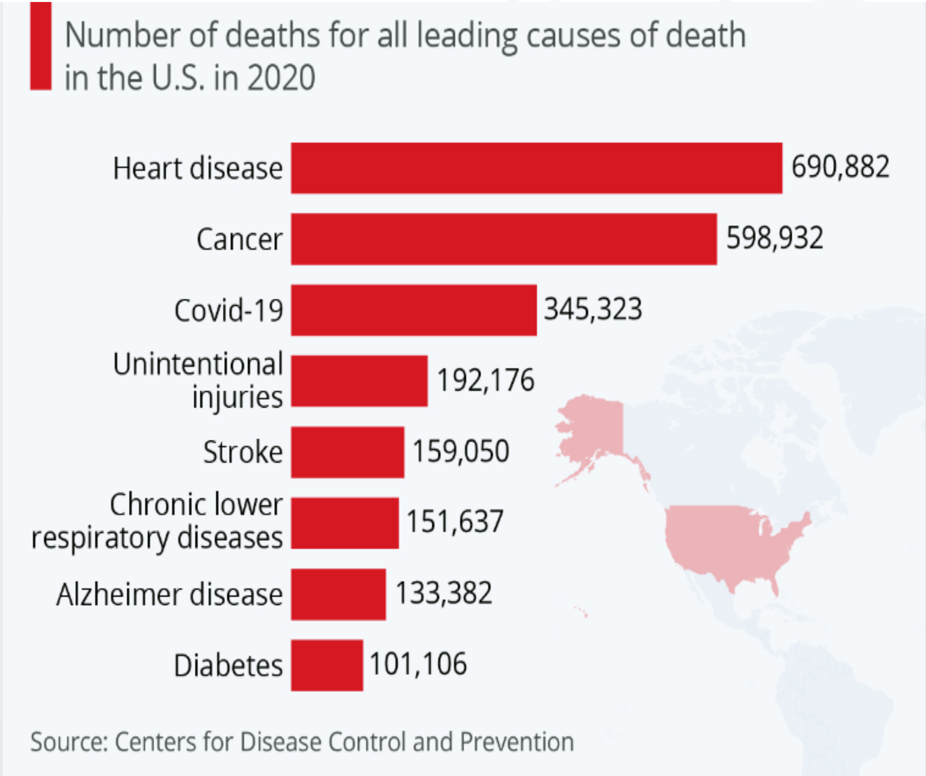
The biggest hurdle with heart disease is detecting it. With early identification of cardiac diseases the mortality rate and overall consequences can be reduced.



Every 40 seconds
someone in America has a heart attack

That's 2,200 today

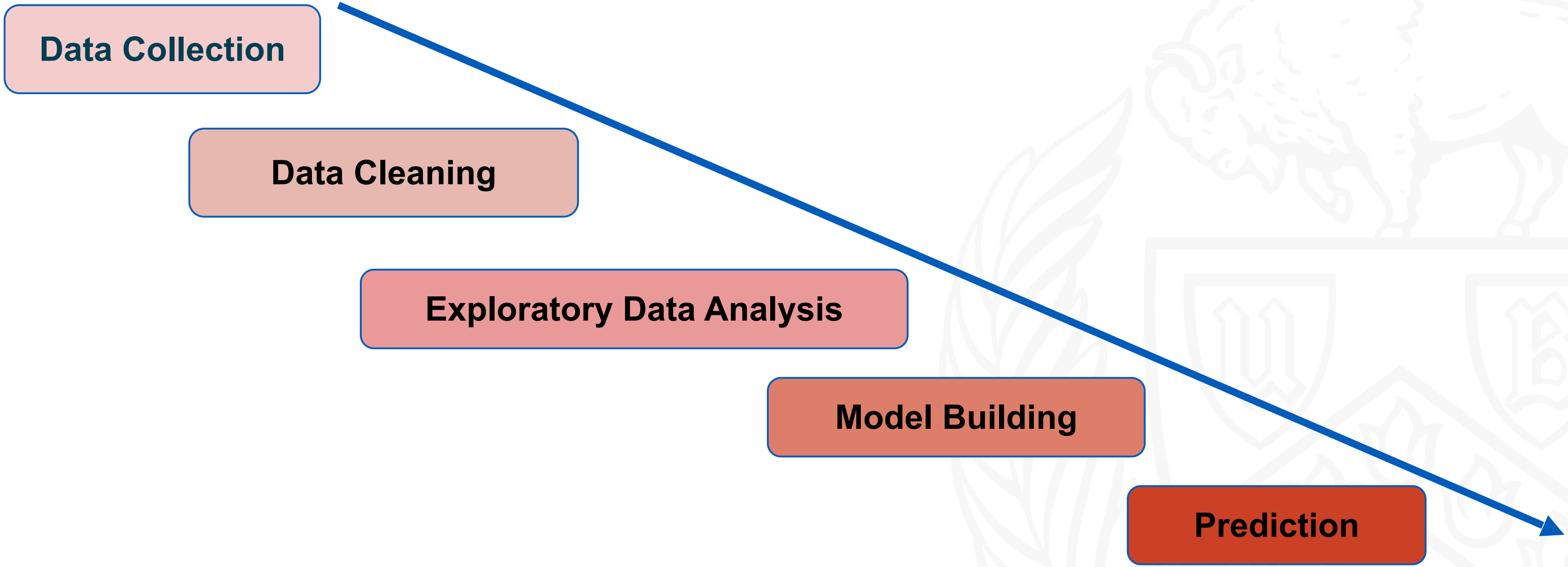
805,000 this year!



Number of deaths for all leading causes of death in the U.S. in 2020

Source: Centers for Disease Control and Prevention

Process Flow



Data Collection

Data Cleaning

Exploratory Data Analysis

Model Building

Prediction

Data Description

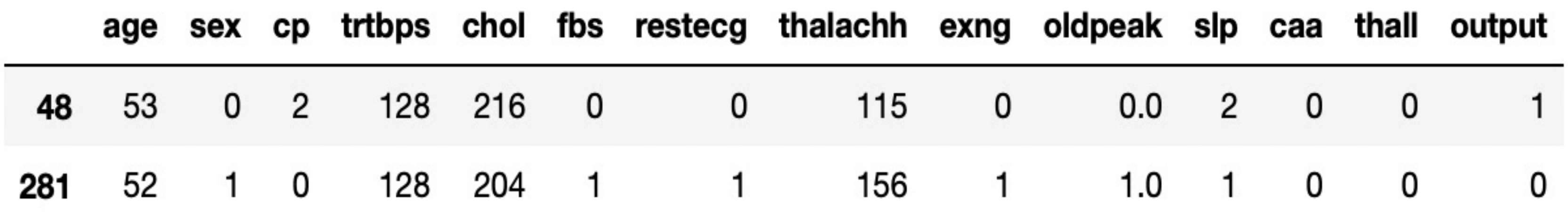
Factors Notation	Description
age	Person’s age in years
sex	Sex of the patient (1 = male, 0 = female)
exang	Exercise induced angina (1 = yes; 0 = no)
ca	Number of major vessels (0-3)
cp	Chest Pain type ➢ Value 0: typical angina ➢ Value 1: atypical angina ➢ Value 2: non-anginal pain ➢ Value 3: asymptomatic
chol	The person’s cholesterol measurement in mg/dl
fbs :	The person’s fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
rest_ecg	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)

Factors Notation	Description
thalach	The person’s maximum heart rate achieved
target	0= Less chance of heart attack 1= more chance of heart attack
oldpeak	ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot)
slope	the slope of the peak exercise ST segment — 0: downsloping; 1: flat; 2:upsloping
thal	A blood disorder called thalassemia ➢ Value 0: NULL (dropped from the dataset previously ➢ 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)
trtbps	The person’s resting blood pressure (mm Hg on admission to the hospital)
ADDITIONAL COLUMNS	**CROSS VERIFIED WITH MEDICAL RESEARCH STUDENT TO UNDERSTAND WHICH FACTORS CAN BE CONSIDERED TO CALCULATE BELOW COLUMNS
smoke habits	Whether a person smokes
physical activity	Whether a person exercises
diet	Whether a person has low fat diet or high fat diet

Data Cleaning

Handling Null Values

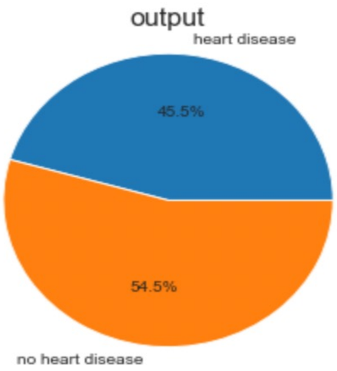
Column thall has two records with null values .



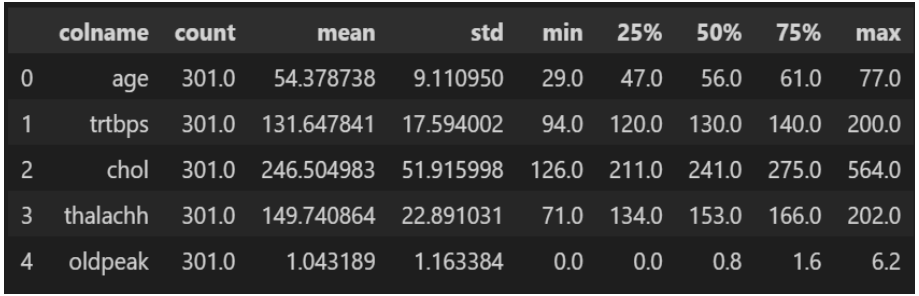
	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
48	53	0	2	128	216	0	0	115	0	0.0	2	0	0	1
281	52	1	0	128	204	1	1	156	1	1.0	1	0	0	0

Class Imbalance

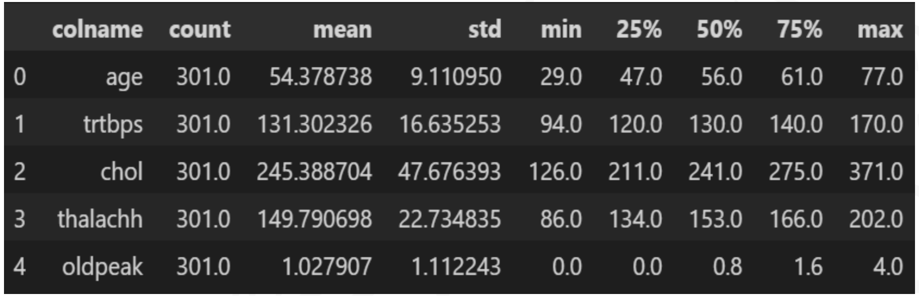
There is no class imbalance in the output column.



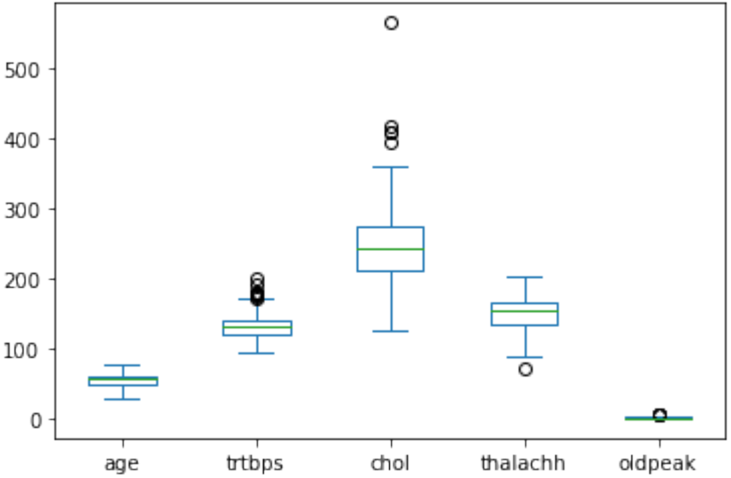
Detecting and handling outliers

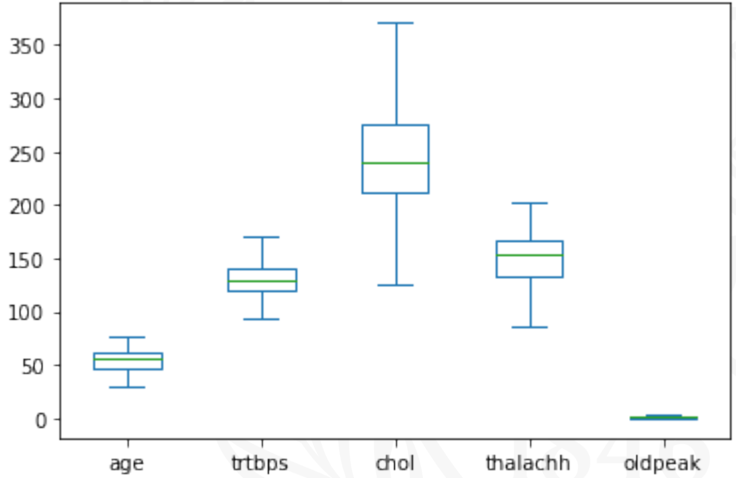


	colname	count	mean	std	min	25%	50%	75%	max
0	age	301.0	54.378738	9.110950	29.0	47.0	56.0	61.0	77.0
1	trtbps	301.0	131.647841	17.594002	94.0	120.0	130.0	140.0	200.0
2	chol	301.0	246.504983	51.915998	126.0	211.0	241.0	275.0	564.0
3	thalachh	301.0	149.740864	22.891031	71.0	134.0	153.0	166.0	202.0
4	oldpeak	301.0	1.043189	1.163384	0.0	0.0	0.8	1.6	6.2



	colname	count	mean	std	min	25%	50%	75%	max
0	age	301.0	54.378738	9.110950	29.0	47.0	56.0	61.0	77.0
1	trtbps	301.0	131.302326	16.635253	94.0	120.0	130.0	140.0	170.0
2	chol	301.0	245.388704	47.676393	126.0	211.0	241.0	275.0	371.0
3	thalachh	301.0	149.790698	22.734835	86.0	134.0	153.0	166.0	202.0
4	oldpeak	301.0	1.027907	1.112243	0.0	0.0	0.8	1.6	4.0



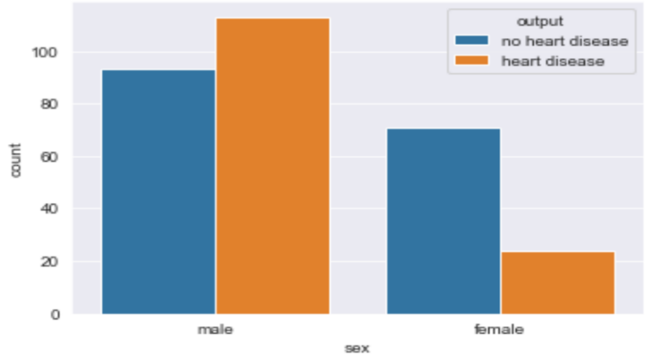


Data scaling

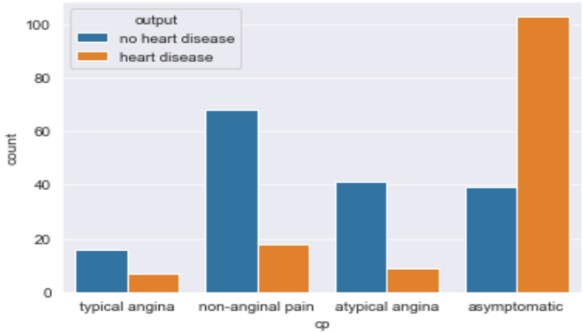
- Standard scaling has been done for the below numerical columns
  - ➢ Age
  - ➢ Trtbps
  - ➢ Chol
  - ➢ Thalachh
  - ➢ Oldpeak

Exploratory Data Analysis

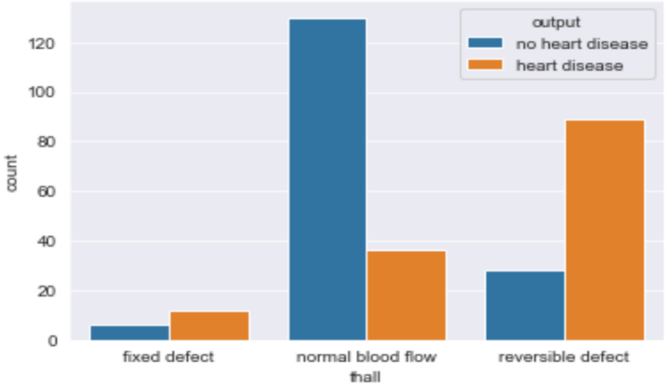
Gender vs Output



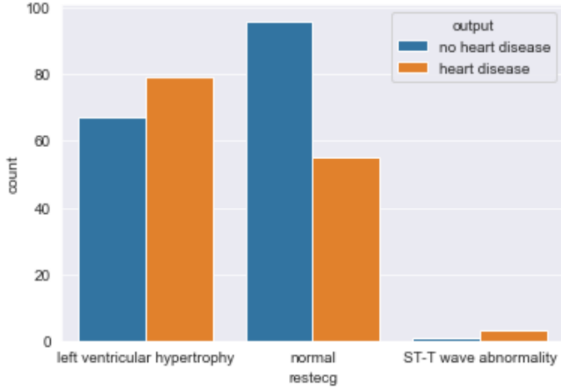
Chest Pain Type vs output

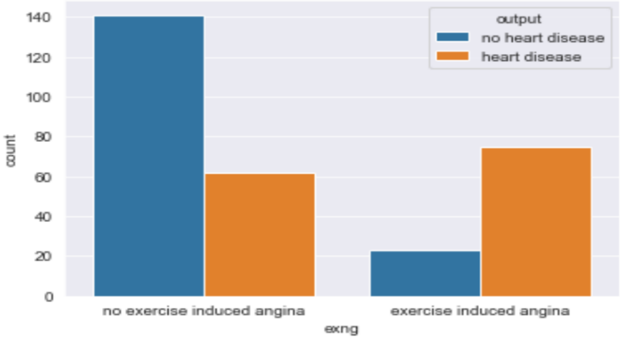


Thalassemia vs Output

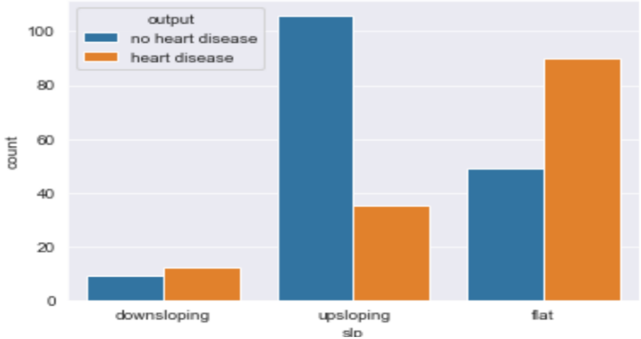


Restecg vs Output

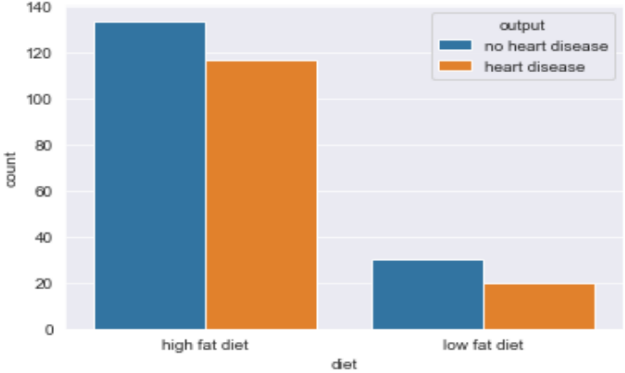




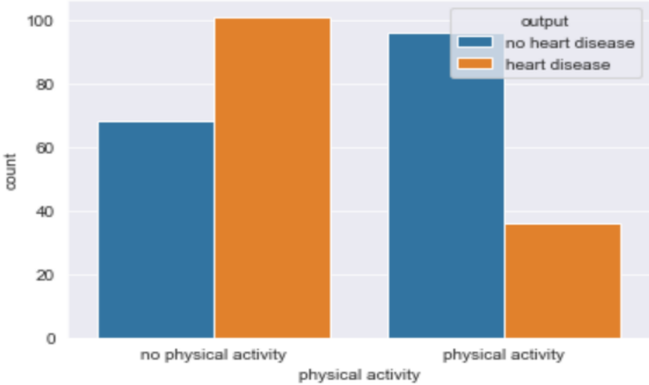
Exercise Induced Angina vs Output



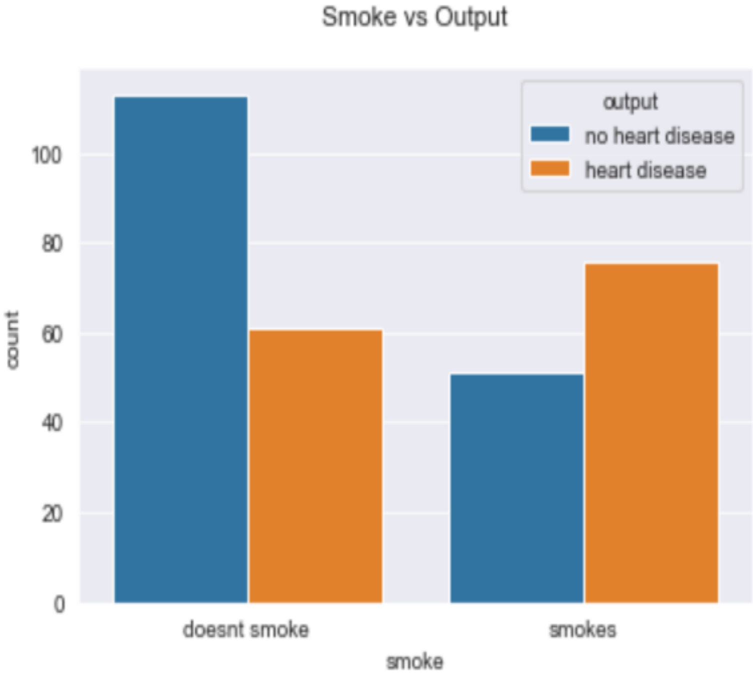
Slope vs Output

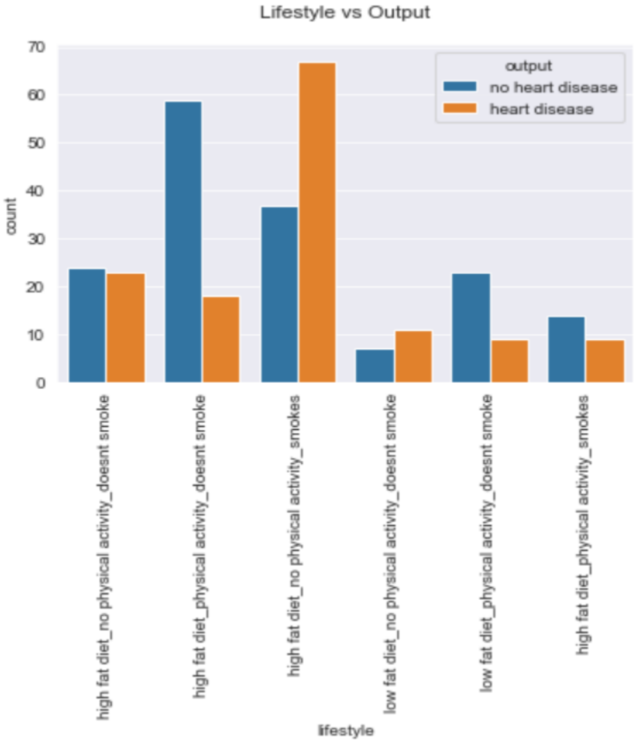


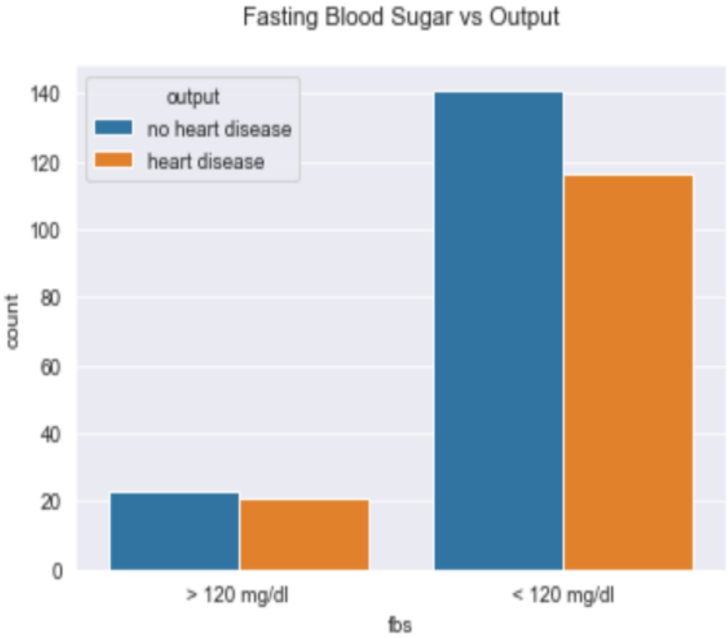
Diet vs Output

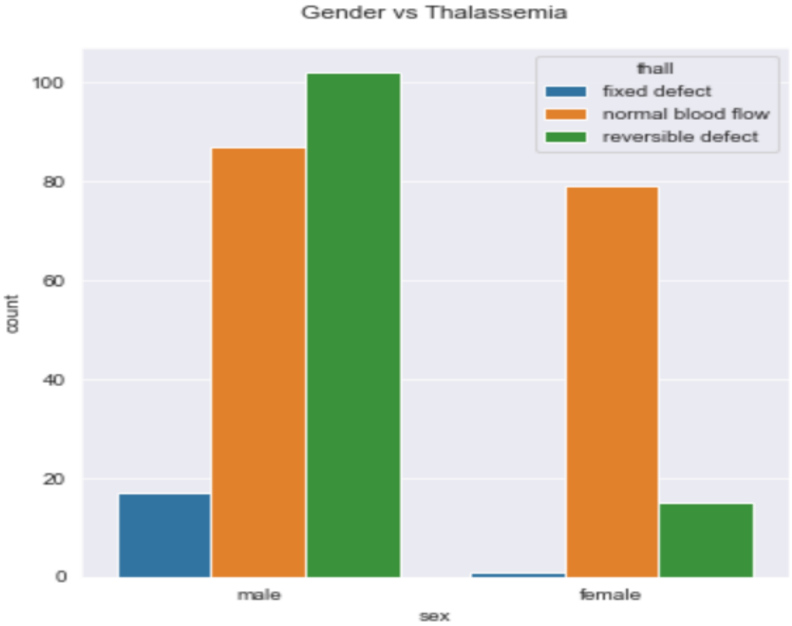


Physical Activity vs Output

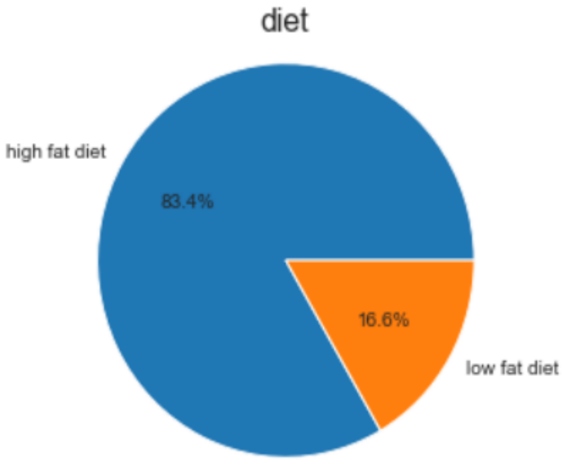

Smoke vs Output
output
no heart disease
heart disease
count
0
20
40
60
80
100
doesnt smoke
smokes
smoke

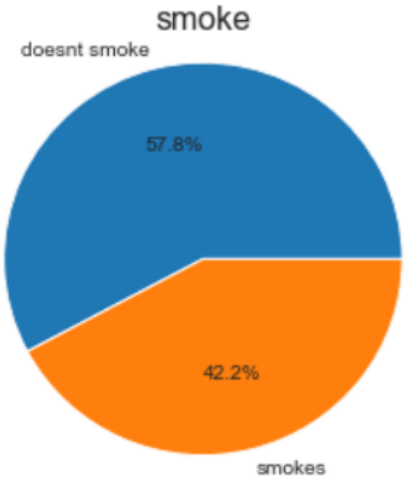

Lifestyle vs Output
output
no heart disease
heart disease
count
0
10
20
30
40
50
60
70
high fat diet_no physical activity_doesnt smoke
high fat diet_physical activity_doesnt smoke
high fat diet_no physical activity_smokes
low fat diet_no physical activity_doesnt smoke
low fat diet_physical activity_doesnt smoke
high fat diet_physical activity_smokes
lifestyle

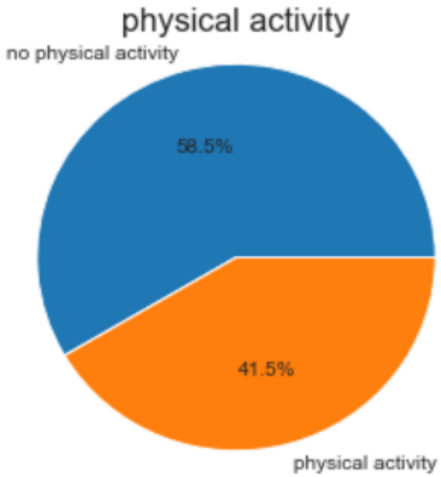

Fasting Blood Sugar vs Output
output
no heart disease
heart disease
count
0
20
40
60
80
100
120
140
> 120 mg/dl
< 120 mg/dl
fbs


Gender vs Thalassemia
thall
fixed defect
normal blood flow
reversible defect
count
0
20
40
60
80
100
male
female
sex

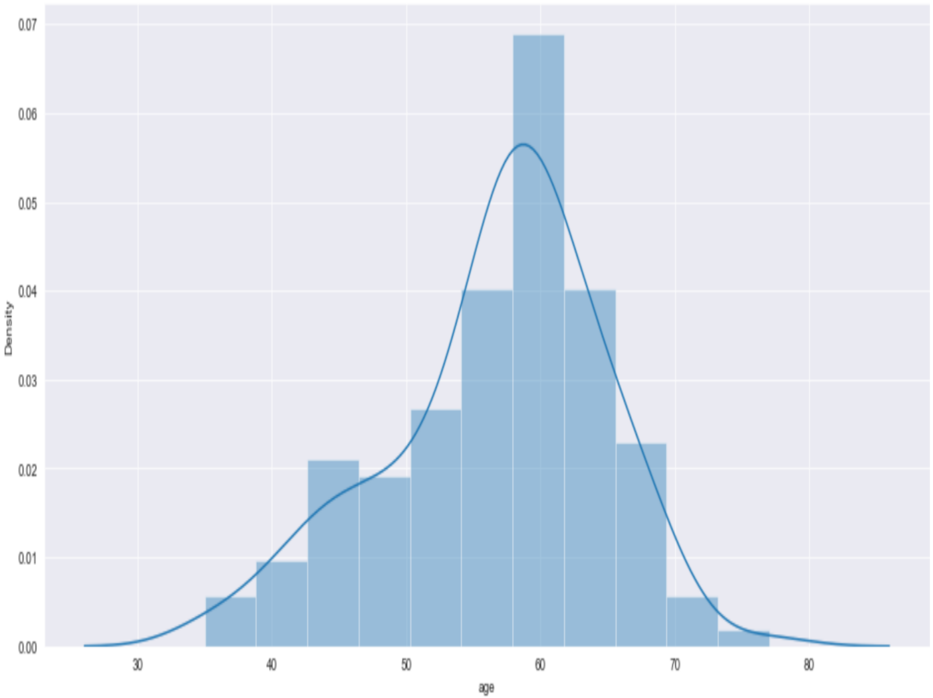
Statistical Interpretation of Diet, Smoke habits and Physical Activity



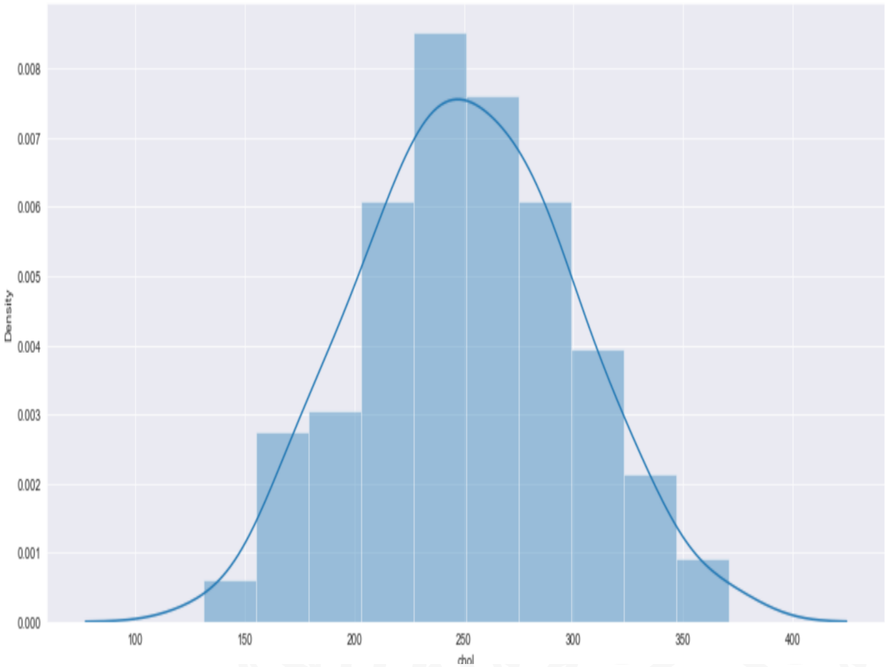


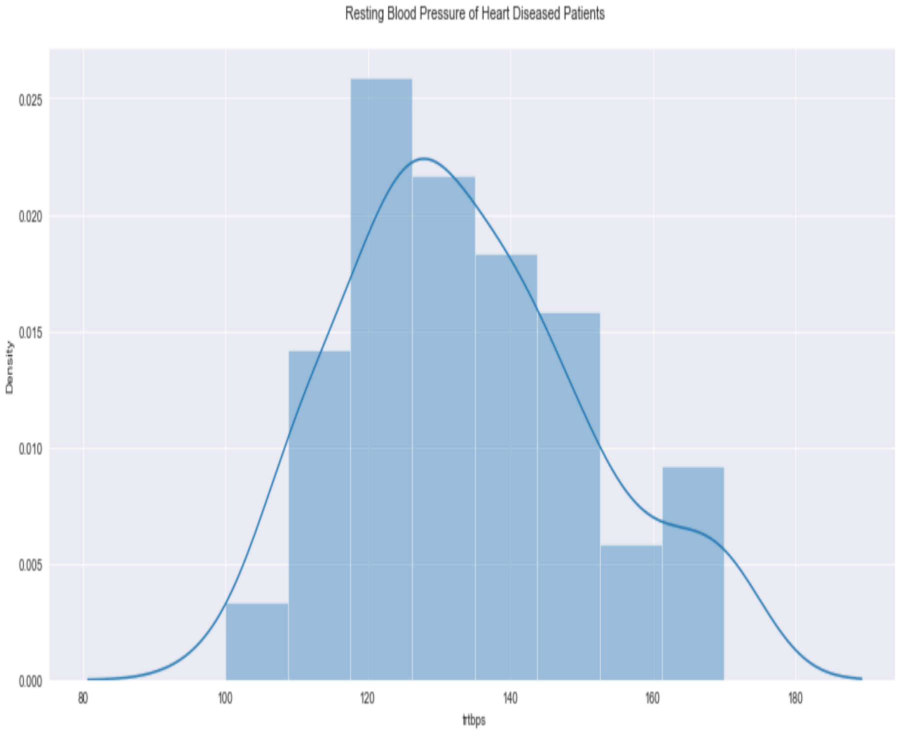


Age of Heart Diseased Patients

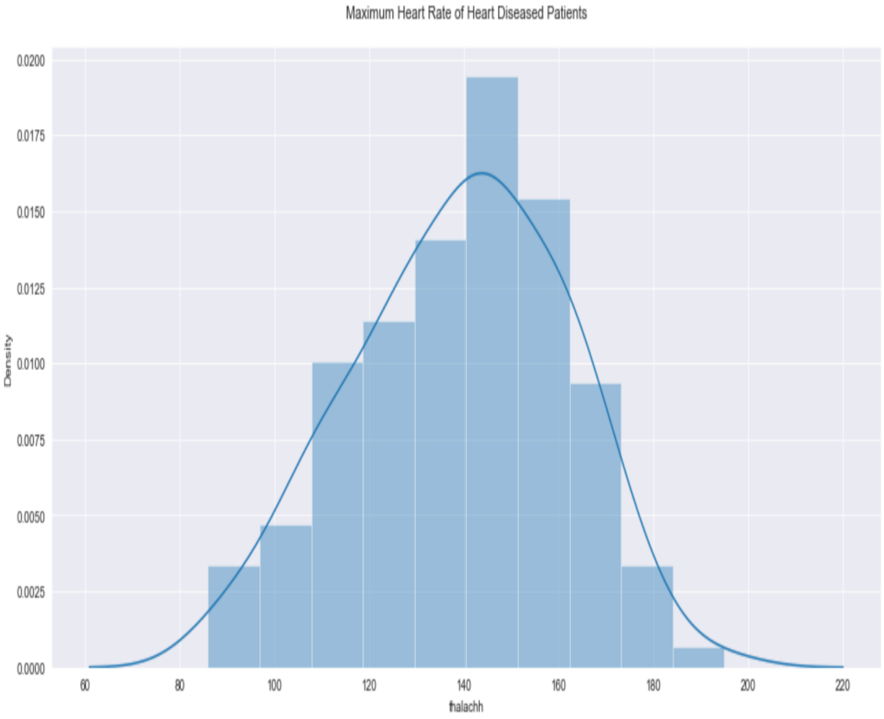


Cholestrol of Heart Diseased Patients



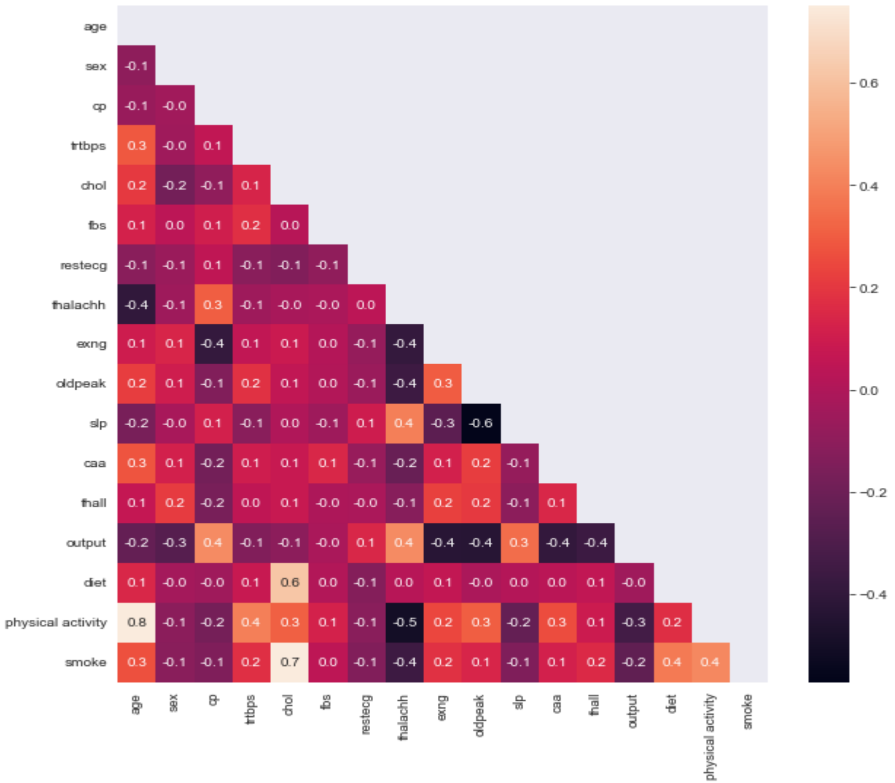


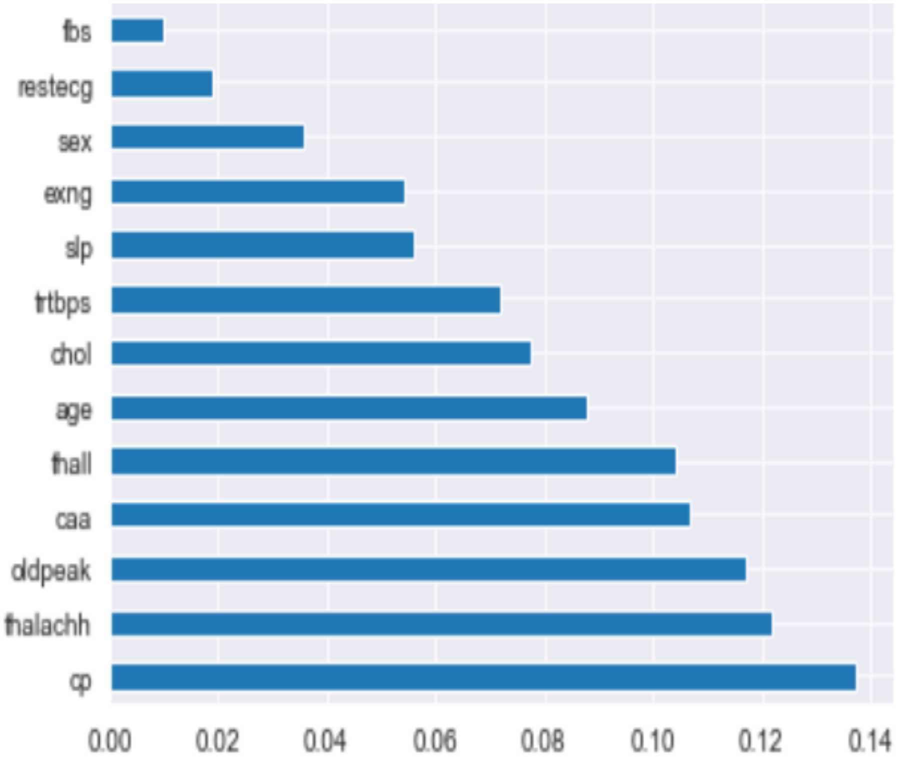
Resting Blood Pressure of Heart Diseased Patients



Maximum Heart Rate of Heart Diseased Patients

Feature Selection

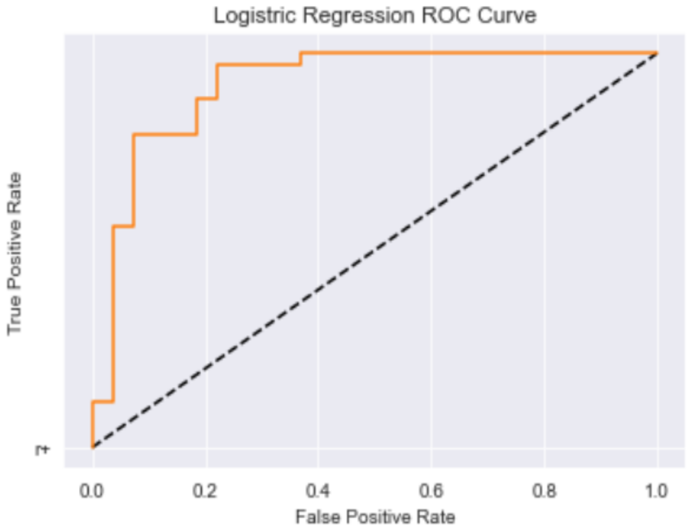


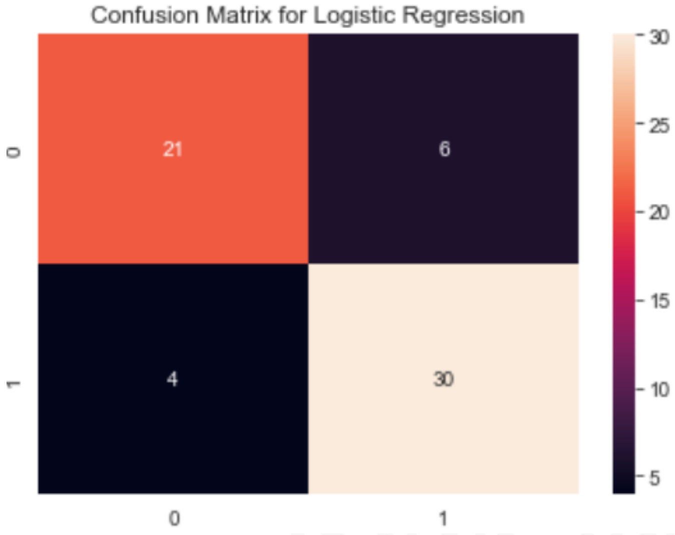


Model Building

Logistic Regression

Train Accuracy	Test Accuracy	Precision	Recall	F1- score	Sensitivity	Specificity
85.83	83.60	83.33	88.23	85.71	88.23	77.77

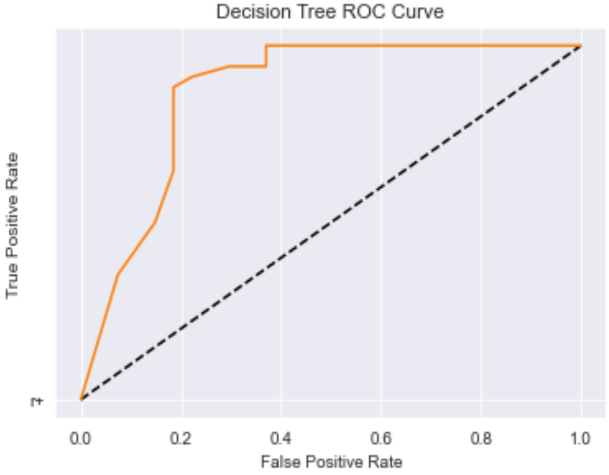


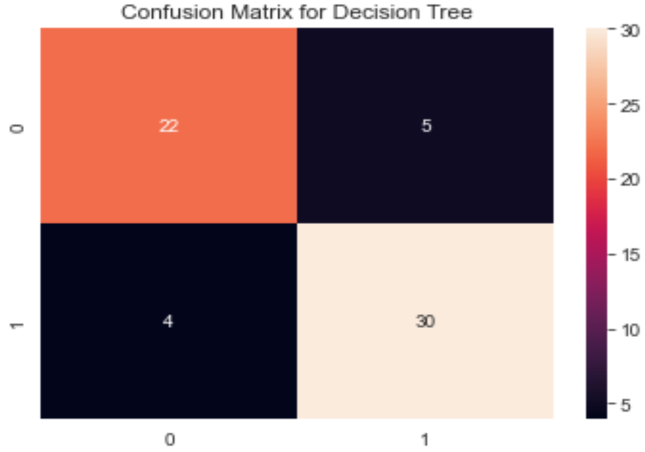


Decision Tree

Parameters used: min_samples_split=25,random_state = 42

Train Accuracy	Test Accuracy	Precision	Recall	F1- score	Sensitivity	Specificity
87.50	85.24	85.71	88.23	86.95	88.23	81.48

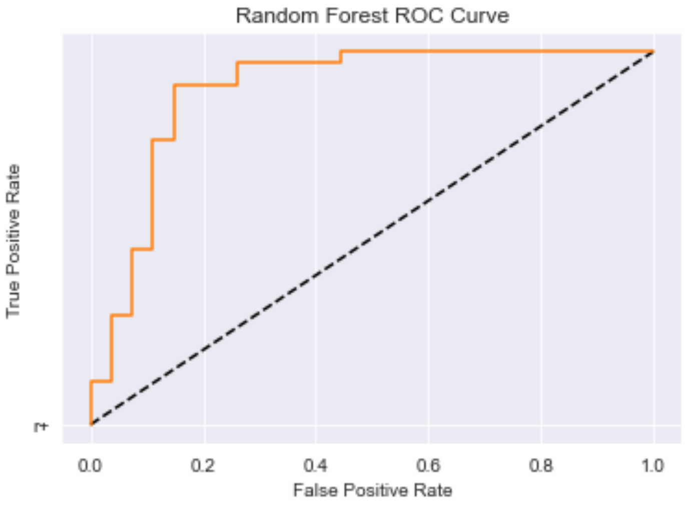


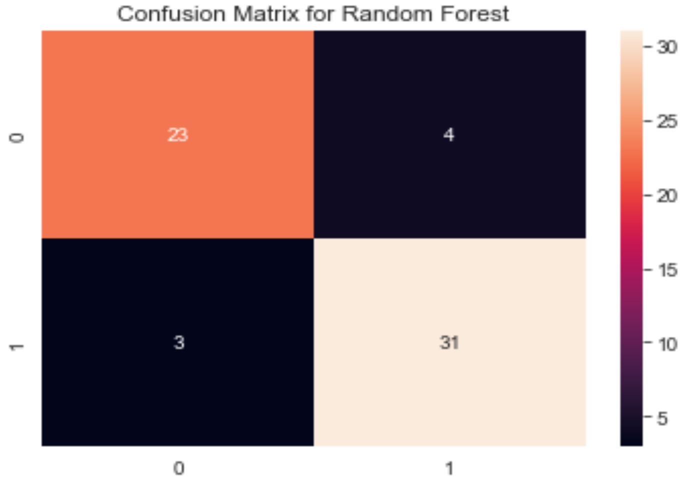


Random Forest

Parameters used: n_estimators = 65,min_samples_split=25,random_state = 42

Train Accuracy	Test Accuracy	Precision	Recall	F1- score	Sensitivity	Specificity
89.58	88.52	88.57	91.17	89.85	91.17	85.18

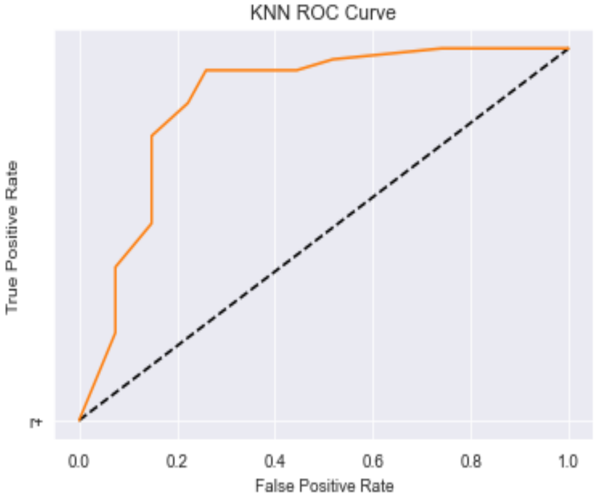




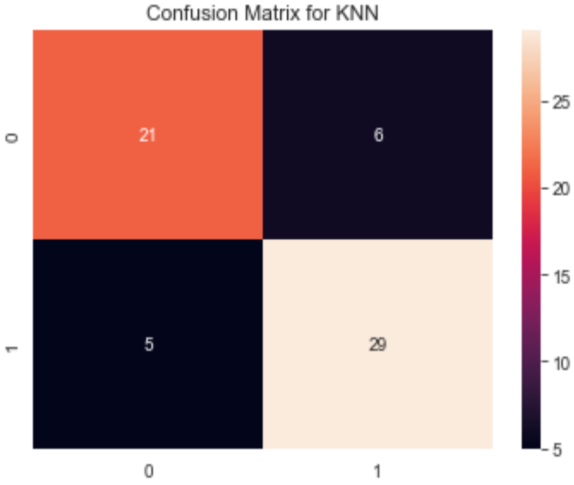
K Nearest Neighbors

Parameters used: n_neighbors=10,n_jobs=-1

Train Accuracy	Test Accuracy	Precision	Recall	F1- score	Sensitivity	Specificity
82.91	81.96	82.85	85.29	84.05	85.29	77.77



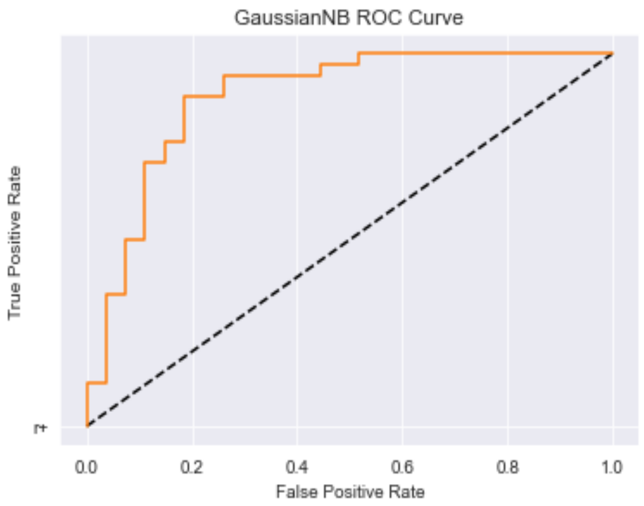
KNN ROC Curve

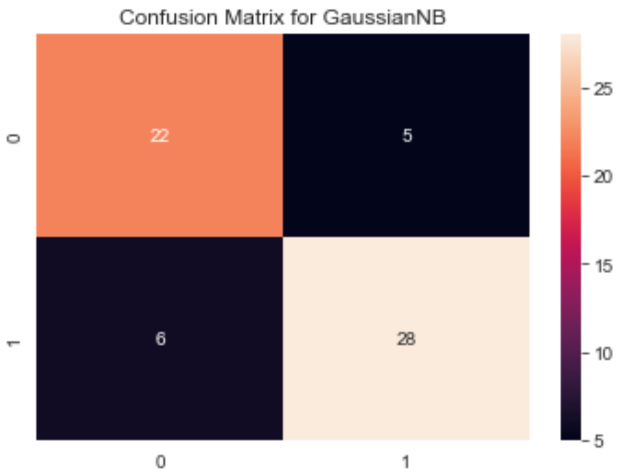


Confusion Matrix for KNN

Gaussian Naive Bayes

Train Accuracy	Test Accuracy	Precision	Recall	F1- score	Sensitivity	Specificity
82.91	81.96	84.84	82.35	83.58	82.35	81.48

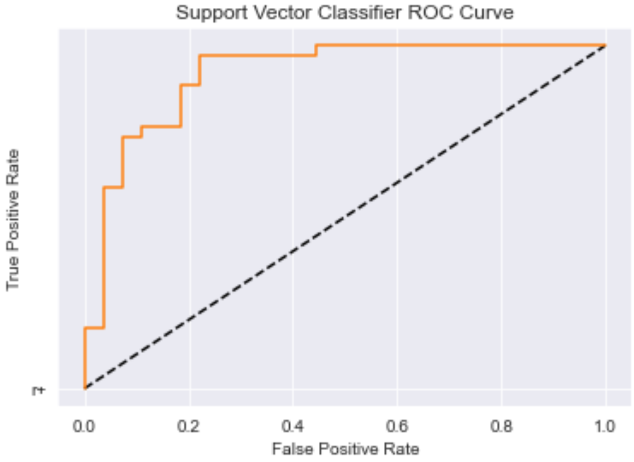


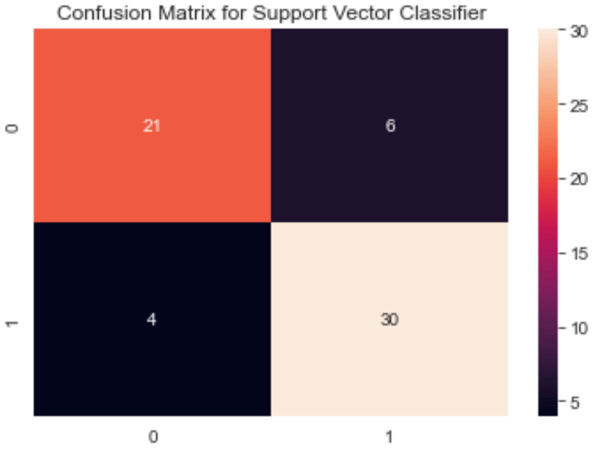


Support Vector Classifier

Parameters used: kernel='linear', C=1,random_state=42,probability=True

Train Accuracy	Test Accuracy	Precision	Recall	F1- score	Sensitivity	Specificity
84.16	83.60	83.33	88.23	85.71	88.23	77.77





Comparison

Model Name	Logistic Regression	Decision tree	Random forest	K nearest Neighbor	Naive Bayes	Support vector Classifier
Train Accuracy%	85.83	87.50	89.58	82.91	82.91	84.16
Test Accuracy%	83.60	85.24	88.52	81.96	81.96	83.60
Precision%	83.33	85.71	88.57	82.85	84.84	83.33
Recall %	88.23	88.23	91.17	85.29	82.35	88.23
F1- score%	85.71	86.95	89.85	84.05	83.58	85.71
Sensitivity%	88.23	88.23	91.17	85.29	82.35	88.23
Specificity%	77.77	81.48	85.18	77.77	81.48	77.77

Conclusion

- Comparing the prediction results of all the models employed, **Random Forest model** has highest accuracy for prediction of unseen data i.e., **88.52%.** The model is more sensitive than specific. The most contributing features are **chest pain** and **maximum heart rate achieved.**

- Using our heart attack prediction model, given any person’s medical data, it is easy to almost accurately predict the risk of heart attack at early stages. Through the diagnostic and predicted result, one can be treated with apt medication and follow healthy lifestyle to prevent from getting cardiovascular diseases.

Future Aspects

- We desire to apply AI to exhibit a connection between different cardiovascular illnesses.
- We can add the treatment and Medicine recommendation.

Questions?