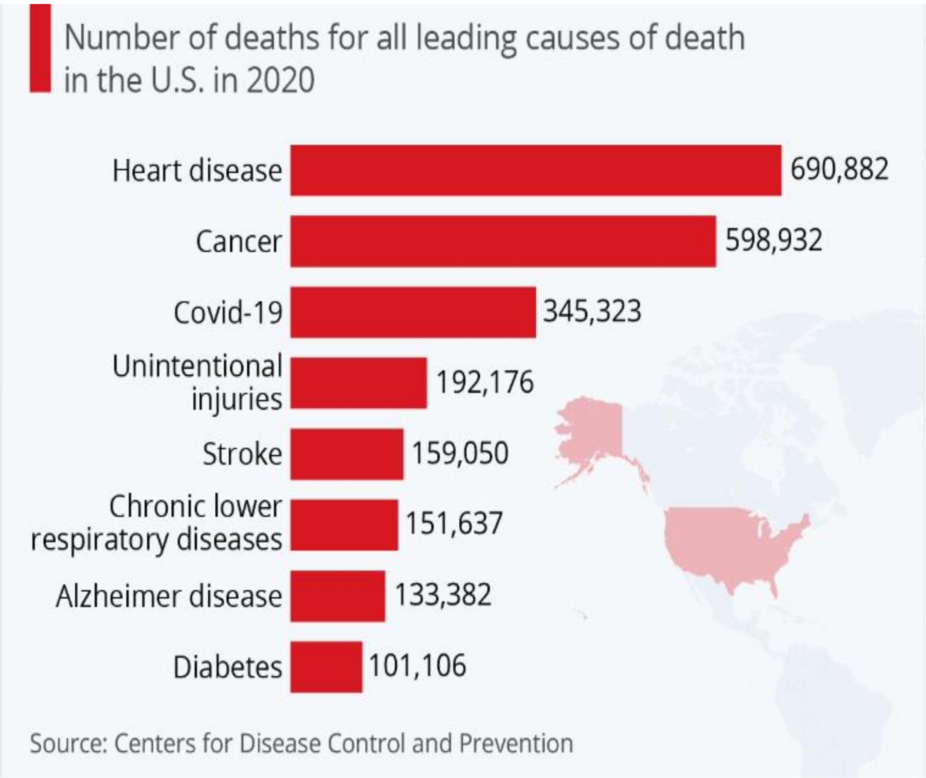
HEART DISEASE PREDICTION SYSTEM

Introduction

In the United States, someone has a heart attack every 40 seconds. Every year, about 805,000 people in the United States have a heart attack. Of these, 605,000 are a first heart attack. The biggest hurdle with heart disease is detecting it. Although there are equipment that can detect heart disease, they are either highly priced or ineffective at calculating the likelihood of heart disease in humans. The mortality rate and overall consequences can be reduced by early identification of cardiac diseases. Since there is a lot of data available nowadays, we can use a variety of machine learning methods to search for hidden patterns. In medical data, the hidden patterns might be used for health diagnosis.

An estimated 6.5 million Americans age 65 and older are living with Alzheimer's in 2022. Seventy-three percent are age 75 or older. About 1 in 9 age 65 and older (10.7%) has Alzheimer's.It is challenging to tell if someone has Alzheimer's disease using the current approach. Only the clinical history and knowledge of any genetic conditions can be used to identify the disease. Also, it can occur that the doctor fails to identify the illness. Although there is no cure for Alzheimer's disease (AD) at this time, early detection can help people live better lives and reduce the severity of the condition. Therefore, a proper diagnosis is crucial, especially in the early stages of AD.

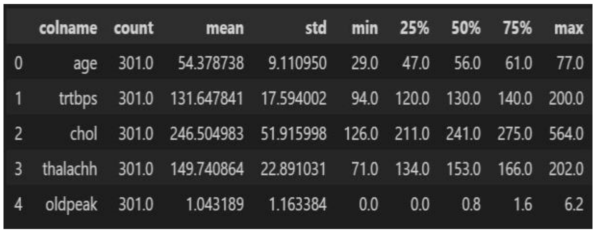
Several factors like a person's vitals like Blood Pressure, Heart rate, Cholesterol and blood sugar, food habits and life style, age can be used to detect diseases like heart attack and Alzheimer's in early stages so that it reduces the risk and helps the person to improve life style and live healthy.



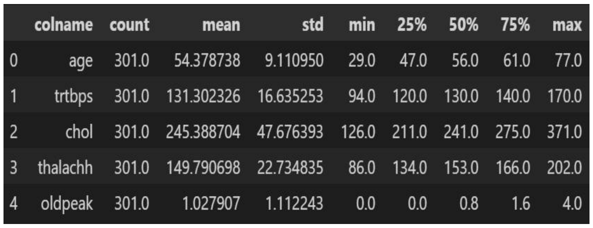
Source: Centers for Disease Control and Prevention

Detecting And Handling Outliers

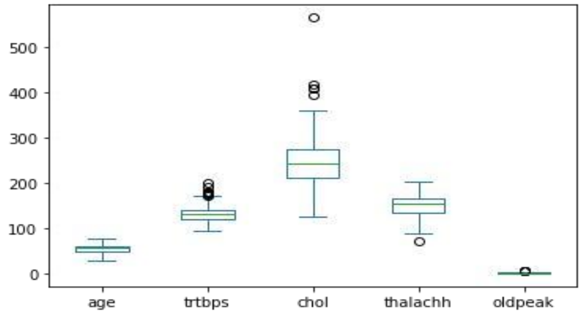
Finding and dealing with outliers is the next phase in the data cleansing process. To lower the data's variability and preserve proper correlation, outliers are removed. In this case, outliers were found using box plots, and they were eliminated using the inter-quartile method.

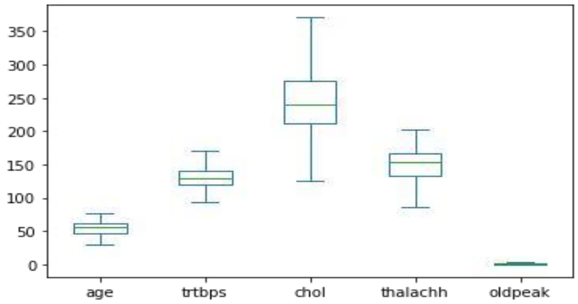


	colname	count	mean	std	min	25%	50%	75%	max
0	age	301.0	54.378738	9.110950	29.0	47.0	56.0	61.0	77.0
1	trtbps	301.0	131.647841	17.594002	94.0	120.0	130.0	140.0	200.0
2	chol	301.0	246.504983	51.915998	126.0	211.0	241.0	275.0	564.0
3	thalachh	301.0	149.740864	22.891031	71.0	134.0	153.0	166.0	202.0
4	oldpeak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





Any outlier below the lower limit is equated to lower limit. Eg: For the thalachh column, this is done.The upper limit is equated to any outliers that are higher than it. For trtbps, chol, and oldpeak columns, this is done.

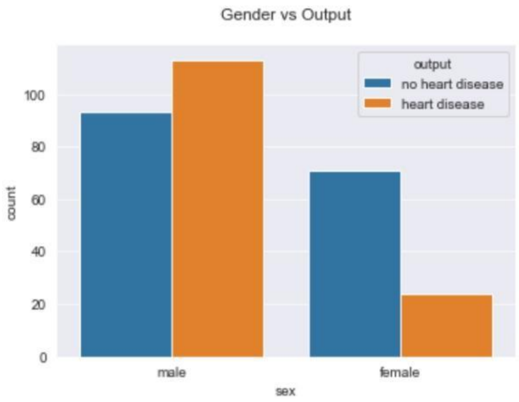
Scaling:

To limit the range and create models, standard scaling is used to bring the data to a single platform. The numerical columns below have been scaled: Trtbps Age Oldpeak Chol Thalachh

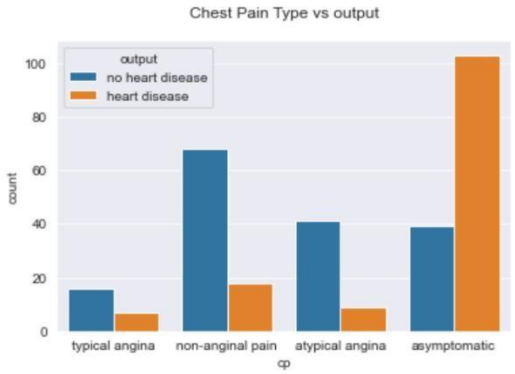
Exploratory Data Analysis:

Data cleansing is completed first, and then exploratory data analysis is performed. It is crucial to correctly evaluate which indicators are most helpful in detecting heart attacks rather than just charting the data. It matters whether each aspect actually has an impact or not. Various graphs have been plotted below.

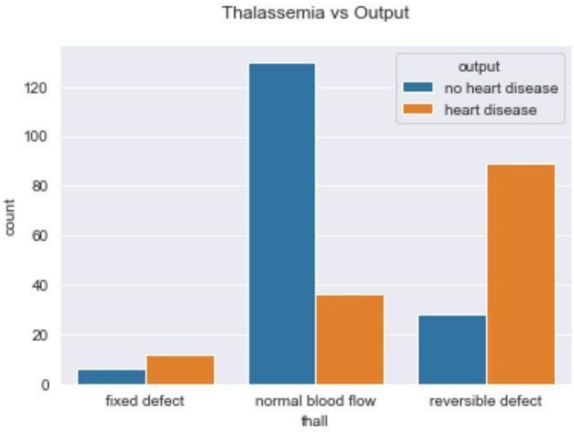
1. ***Gender vs output***
   - From the graph it is visible that Female has low chance of heart attack compared to male.



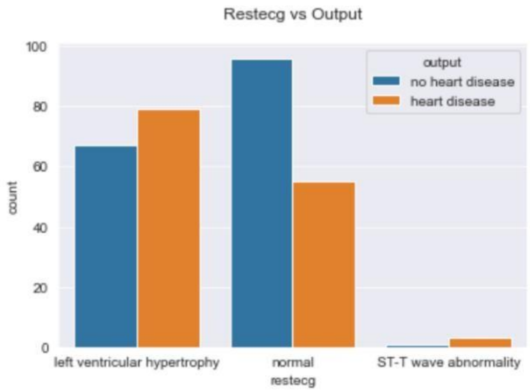
2. ***Chest pain type vs output***
   - This is the most contributing factor among all the other aspects because people who experience asymptomatic chest pain are more likely to suffer a heart attack than people who experience other types of chest pain.



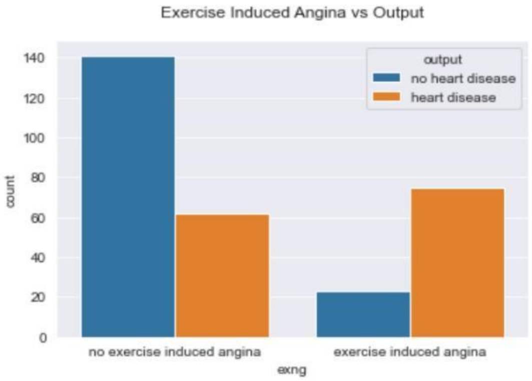
3. ***Thalassemia vs output***
   - Thalassemia is a kind of blood disorder due to insufficient hemoglobin
   - People who have reversible blood flow and fixed defect are more prone to more heart attack.



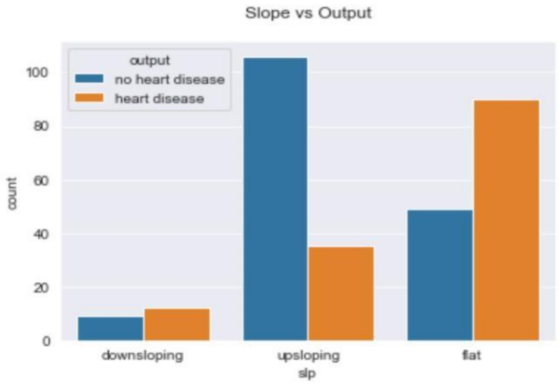
4. ***Restecg vs output***
   - People who have left ventricular hypertrophy are more prone to heart attack



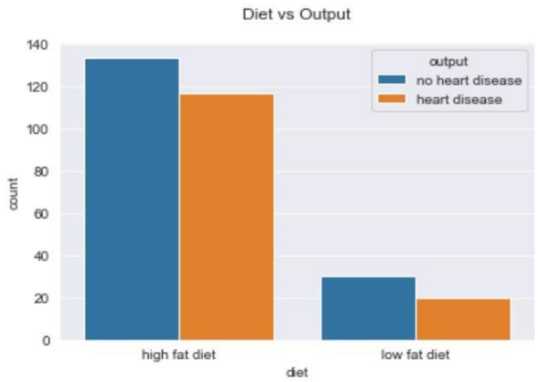
5. ***Exercise induced angina vs output***
   - ➢ People who experience discomfort while exercising are more likely to suffer a heart attack. People who engage in strenuous exercise, such as jogging or working out, frequently exhibit this.



6. ***Slope vs output***
   - ➢ Slope is defined as the slope of a line formed from a graph of the electrocardiogram (ECG) that contains both peaks and depressions. The heart does not beat frequently if the slope is flat.
   - ➢ People with flat types of slope are more susceptible to heart attacks since their hearts don't contract and expand as much..

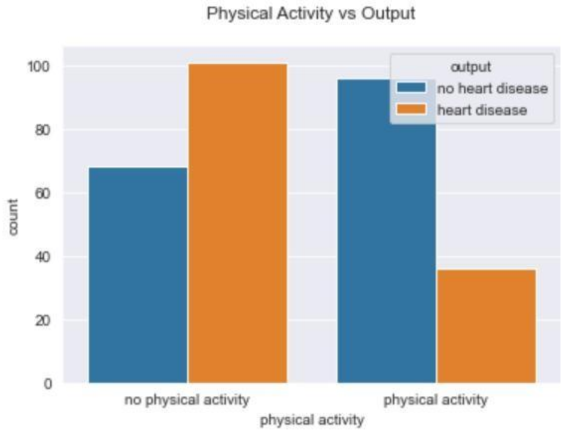


7. ***Diet vs output***
   - ➢ The graph demonstrates that nutrition has little effect on output, making it a minor contributing factor.



8. Physical activity vs output

People who don't exercise regularly are more likely to experience a heart attack than those who do.



9. Smoke vs Output

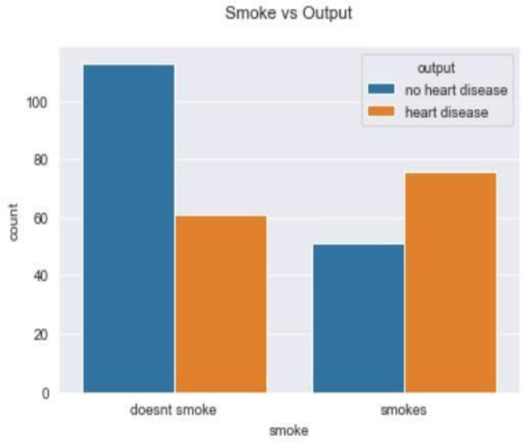


Figure 3.1

People who smoke are more prone to heart disease compared to the people who doesn't smoke.

10. Lifestyle vs Output

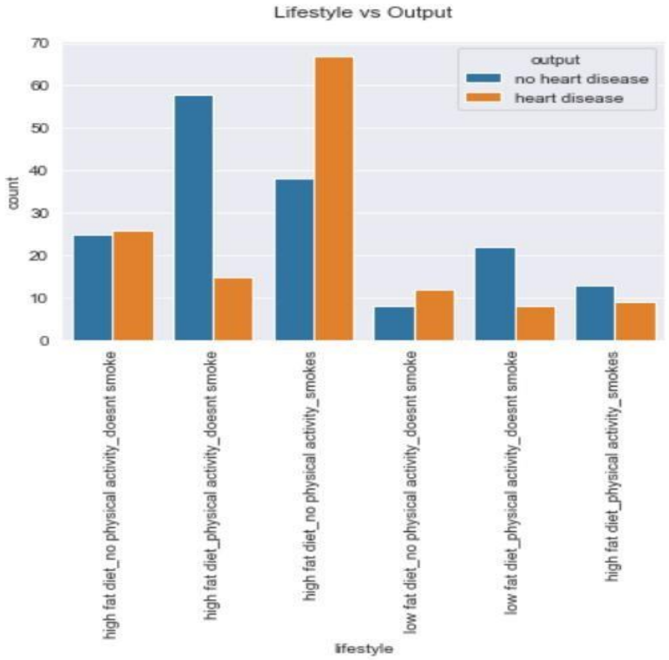


Figure 3.2

In comparison to People who consume a high-fat diet or a low-fat diet, engage in physical activity, refrain from smoking, heart attacks are more likely to happen in people who smoke, do no exercise, and consume a high-fat diet.

11. Fasting Blood Sugar vs Output, Gender vs Thalassemia :

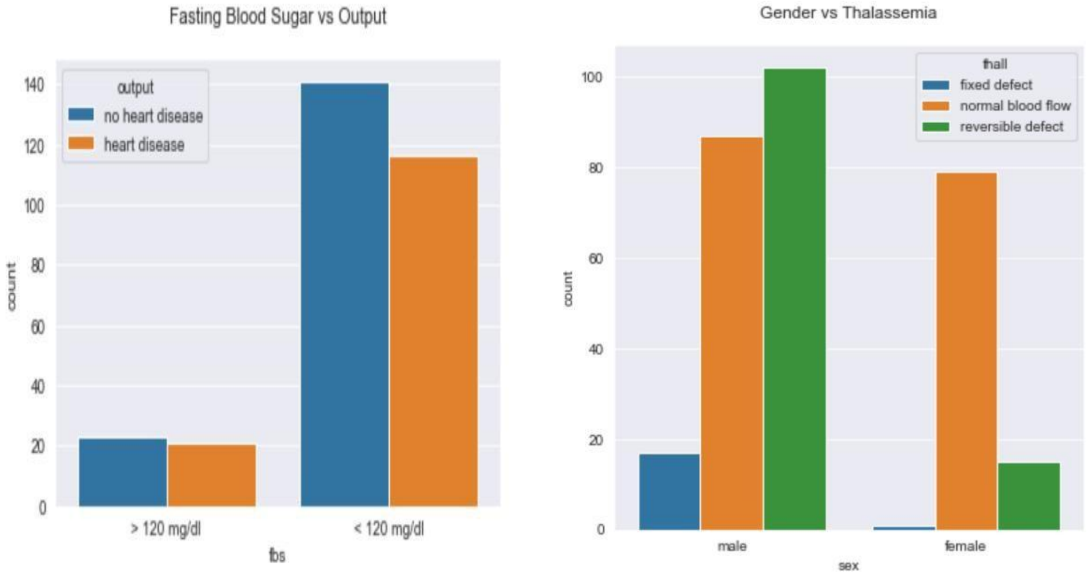


Figure 3.3

- **Fasting Blood Sugar vs Output Graph :** Heart health is not significantly impacted by fasting blood sugar levels. But over time, diabetes has a greater possibility of affecting heart health because it thickens blood, which has an impact on blood flow.
- **Gender vs Thalassemia Graph :** A reversible defect(abnormality) that increases the risk of heart disease exists in the majority of men. There is a modest risk of heart disease in women because the majority of them have normal blood pressure.

<u>Statistical Interpretation of Diet, Smoke habits and Physical Activity</u>

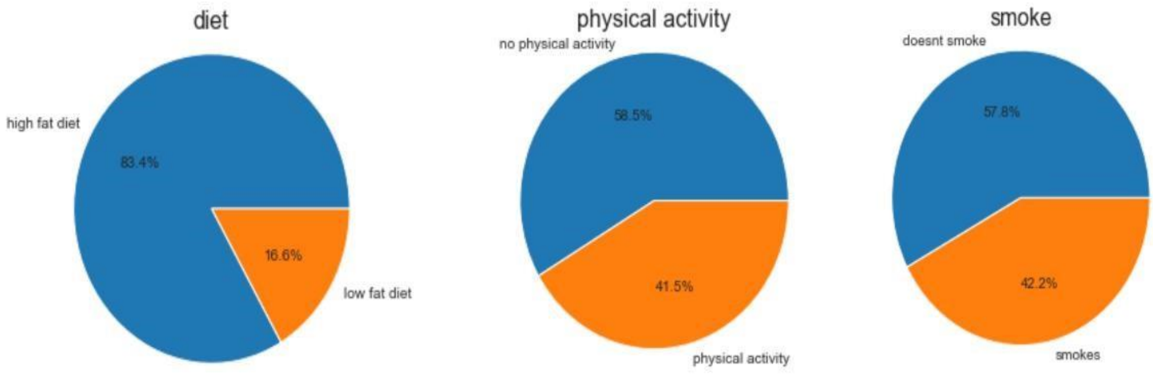


Figure 3.4

Around 83.4% of the population consumes a diet heavy in fat. In addition, the majority of people (58.5%) do not exercise regularly. Non-smokers make up to the majority of the population (57.2%).

Statistical Distribution of Age, Cholesterol, Heart Rate, Blood Pressure on Heart Diseased Patients

1. Age, Cholesterol :

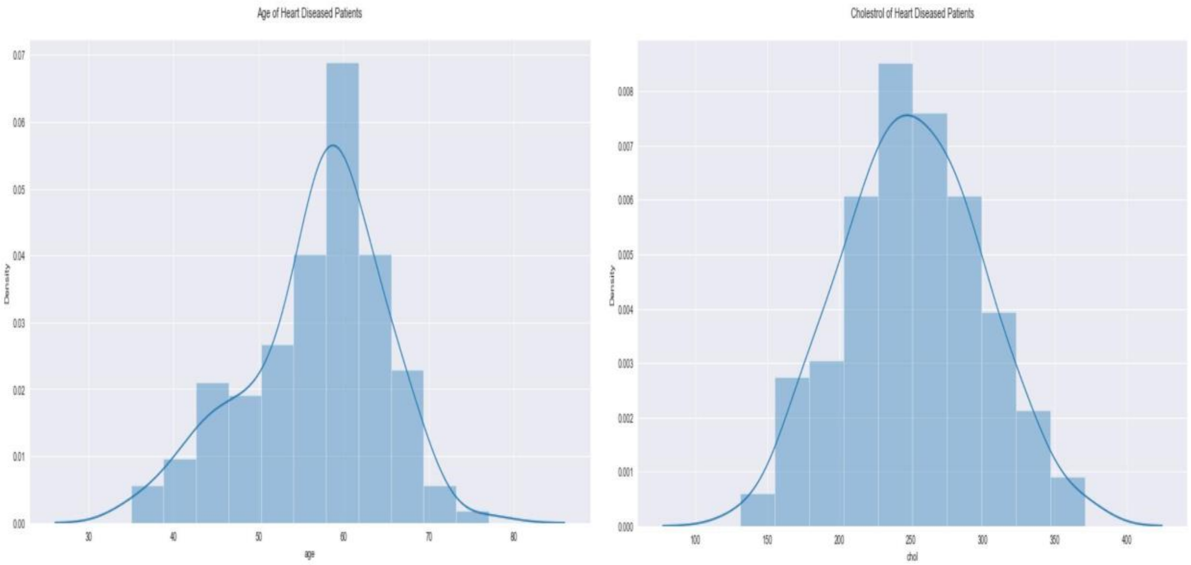
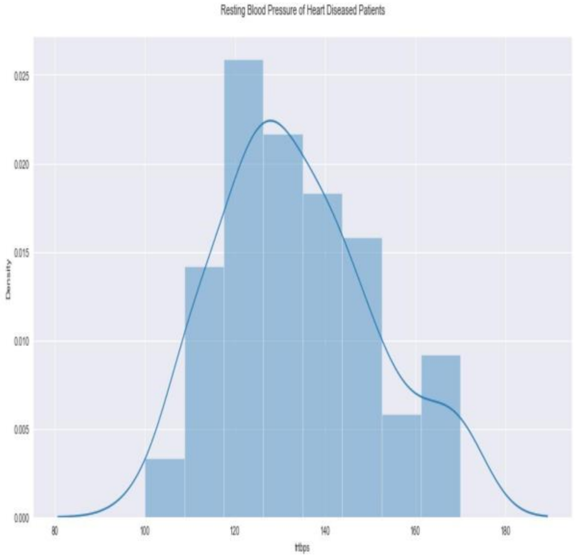


Figure 3.5

- ➢ Heart Disease is very common in the seniors, which is composed of age group centred at 60 and above and common among adults which belong to the age group of 40 to 60 (range) and is clustered between 55 to 65 age group. But it's rare among the age group of 19 to 40 and very rare among the age group of 0 to 18.
- ➢ The second graph shows the Cholesterol distribution of heart diseased patients, where the Cholesterol level varies from 150 to 350, centred around 250, and is clustered between 200 to 300.

2. Resting Blood Pressure, Maximum Heart Rate :



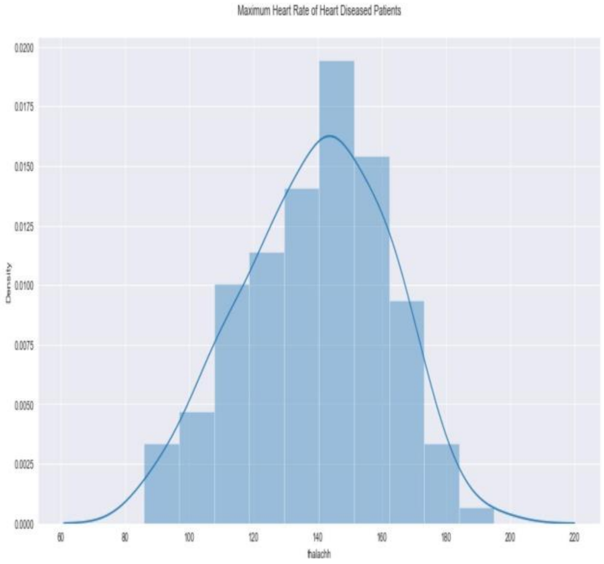


Figure 3.6

- ➢ The graph of the Resting Blood pressure distribution of heart diseased patients depicts the Resting Blood pressure level values ranging from 110 to 170, where mean is around 130, and is clustered between 120 to 140.
- ➢ The graph of the Maximum Heart Rate distribution of heart diseased patients shows the Maximum Heart Rate levels ranging from 100 to 180, the mean is around 145, and is clustered between 120 to 170.

Feature Selection :

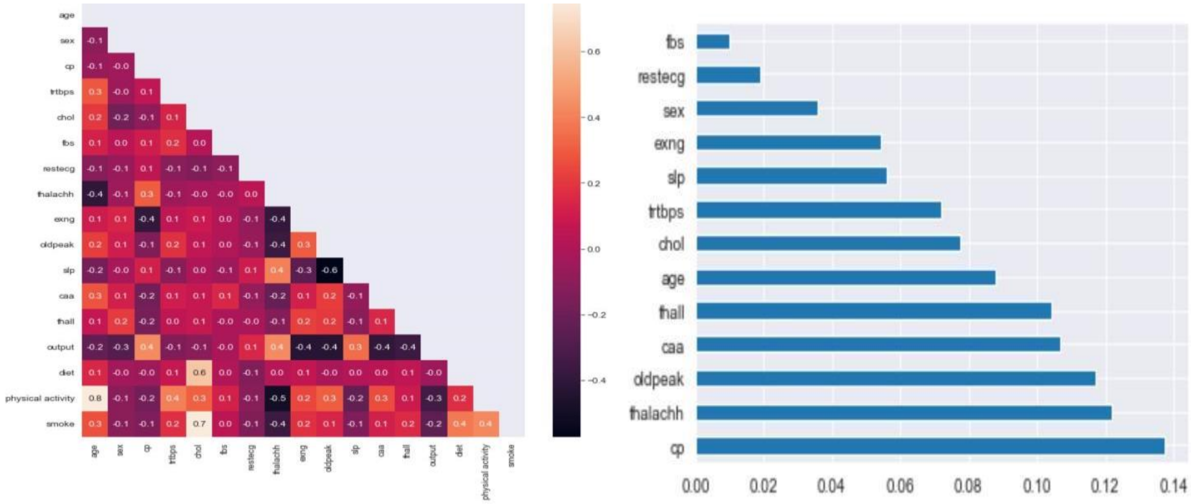


Figure 3.7

- ➢ Correlation shows whether the characteristics are related to each other or to the target variable. Correlation can be positive (increase in one value, the value of the objective variable increases) or negative (increase in one value, the value of the target variable decreased)
- ➢ From the Correlation matrix of heatmap, there is no strong correlation between the attributes, indicating no multi-colinearity issues in the data.
- ➢ We can gain the significance of each feature of our dataset by using the Model Characteristics property. Feature value gives us a score for every function of our results, the higher the score the more significant or appropriate the performance variable is. Feature importance is the built-in class that comes with Tree Based Classifiers to extract the top features for the dataset.
- ➢ Features from cp (Chest Pain) to sex are chosen based on the Random Forest Classifier's feature importance since they collectively account for 90% of Feature importance.

Model Building

As there are no challenges in the data like multicollinearity or imbalanced target variable, all of the classification models are implemented to check which model works best for the taken dataset.

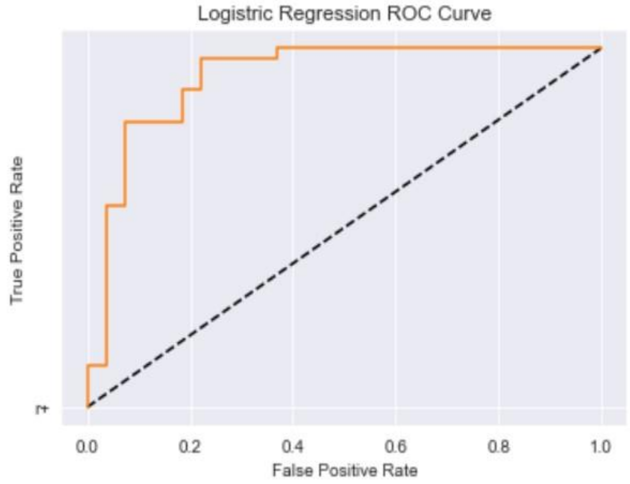
Models implemented:

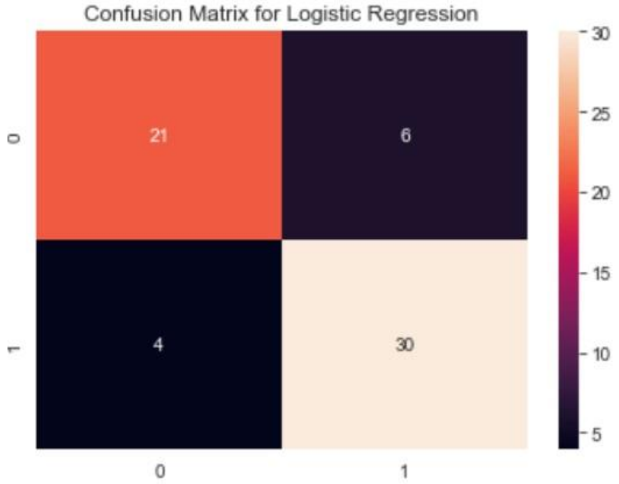
1. Logistic Regression
2. Decision Tree
3. Random Forest Classifier
4. K-nearest Neighbors
5. Gaussian Naive Bayes
6. Support Vector Classifier

Results

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

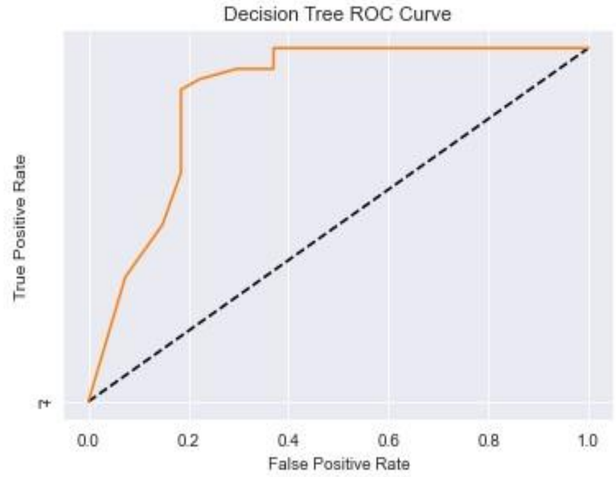


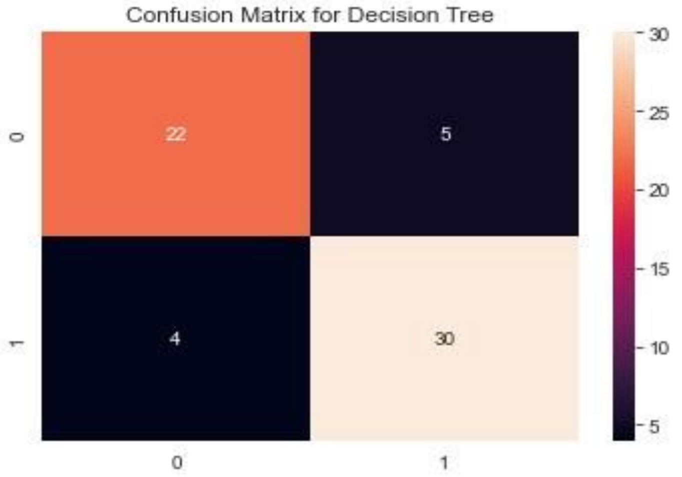


Decision Tree

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

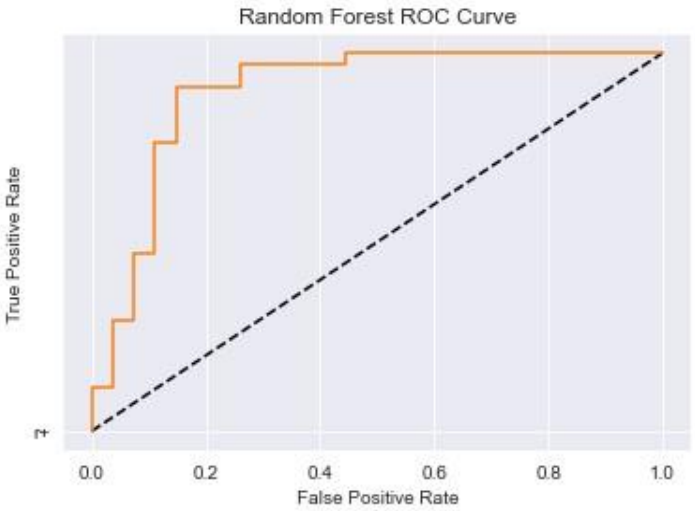


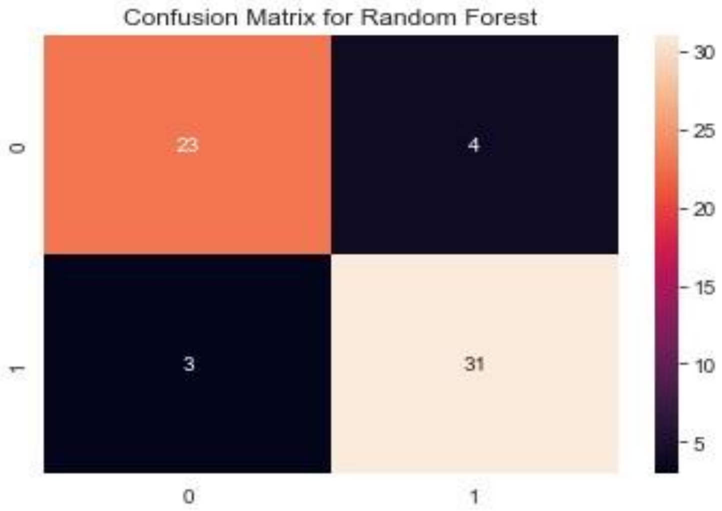


Random Forest Classifier

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

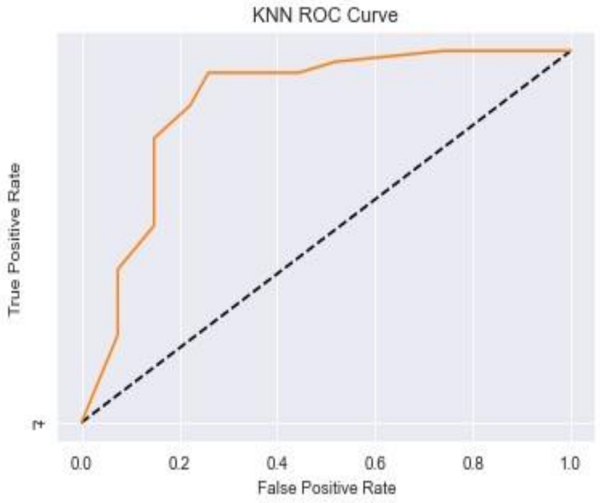


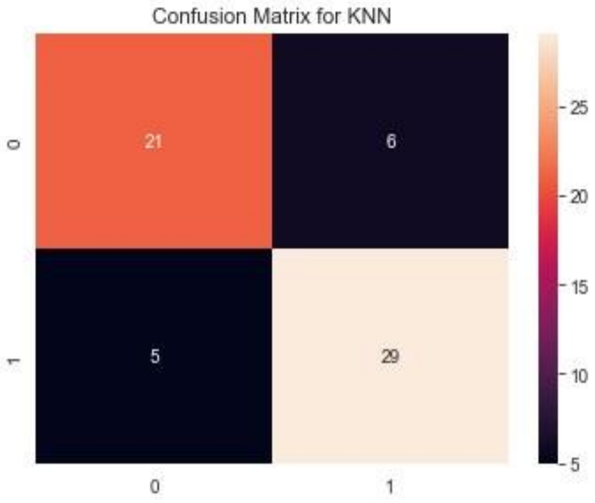


K-nearest Neighbors

Parameters used: n_neighbors = 10, n_jobs = -1

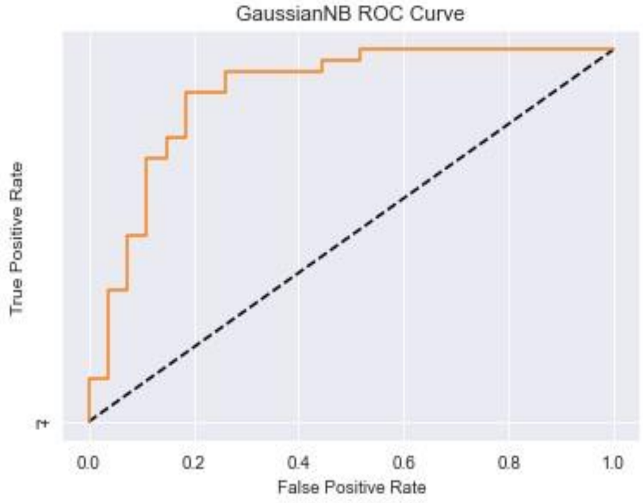
Train Accuracy	**Test Accuracy**	**Precision**	**Recall**	**F1- score**	**Sensitivity**	**Specificity**
82.91	81.96	82.85	85.29	84.05	85.29	77.77

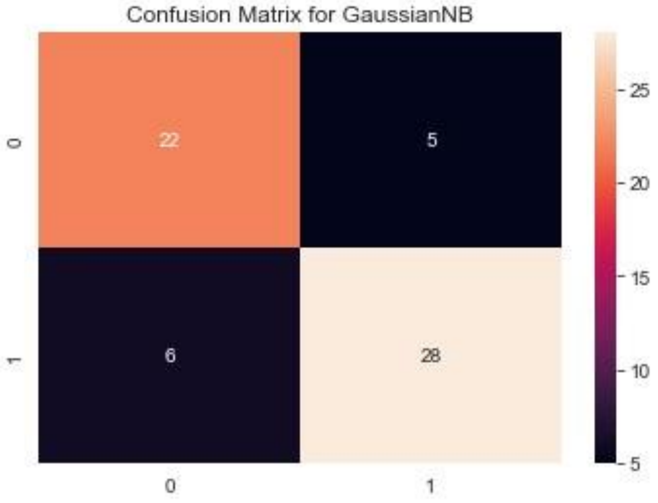




Gaussian Naive Bayes

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

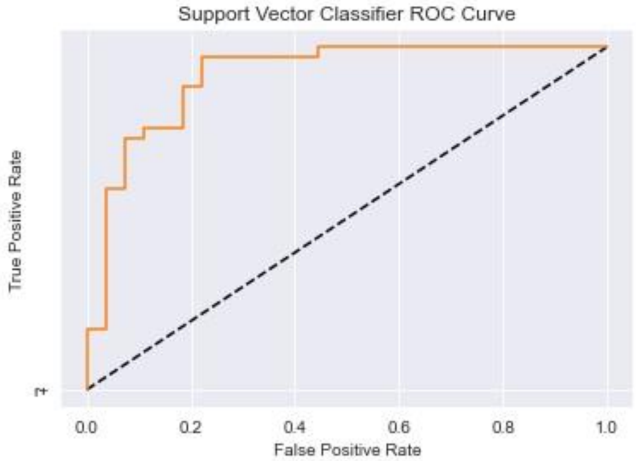


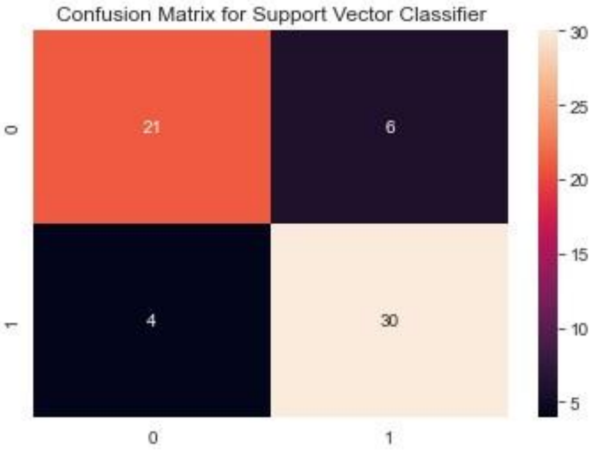


Support Vector Classifier

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

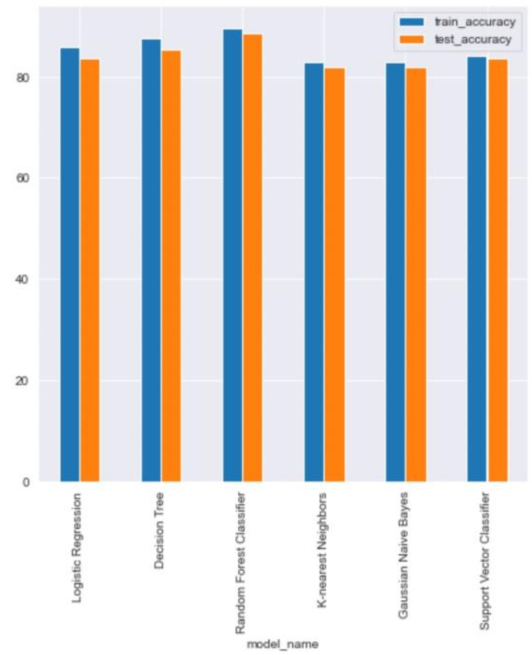
Train Accuracy	Test Accuracy	Precision	Recall	F1- score	Sensitivity	Specificity
84.16	83.60	83.33	88.23	85.71	88.23	77.77

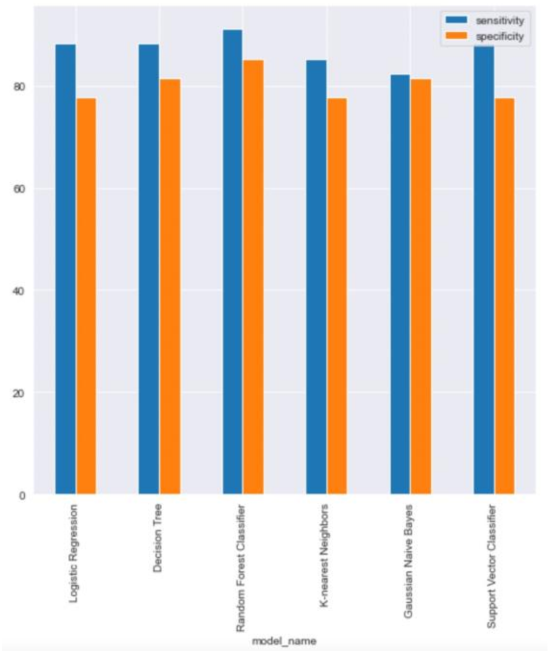




Comparison

From the results, we can observe that Random Forest has better train and test accuracy as well as better specificity and sensitivity when compared to other models. The model is more sensitive than specific i.e., the model may predict person with no heart disease as heart disease which is way better than predicting person who has heart disease as no heart disease. The most contributing features are **chest pain** and **maximum heart rate achieved**.





Conclusion

Using 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.