Heart Attack Prediction

1.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 is 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. In the data era, with access to all kinds of data, 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.

Several factors like a person's vitals like blood pressure, heart rate, cholesterol and blood sugar, food habits and lifestyle, can be used to detect diseases like heart attack in early stages so that it reduces the risk and helps the person to improve lifestyle and live healthy.

To address the above stated problem, many categorical models are used to interpret and predict with highest accuracy the effect of person's vitals and lifestyle on cardiac health.

2.Data Description

Data is gathered from UIC's Machine Learning Repository with 76 attributes pertaining to health and lifestyle of many patients across the globe.

Out of which we carefully chose a subset of 16 features and 3 additional features were created:

a**ge** - Person's age in years

s**ex** - Sex of the patient (1 = male, 0 = female)

e**xang** - Exercise induced angina (1 = yes; 0 = no)

c**a** - Number of major vessels (0-3)

c**p** - 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

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)

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: down sloping; 1: flat; 2: up sloping

t**hal** - 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 smoke (Implemented K-means clustering on chol, fbs, thalachh columns)

physical activity - Whether a person exercise (Implemented K-means clustering on age, trtbps, chol, fbs, thalachh columns)

diet - Whether a person has low fat diet or high fat diet (Calculated based on chol and age)

3.Analysis Methodology

1. **Data Cleaning and Pre processing** - Performed several data sanitation methods to address the data issues:

- Null Values Imputation
- Outlier Detection/Imputation
- Class imbalance check
- Normalization
- Correlation and Feature Selection

2. **Exploratory Data Analysis** - Data analysis was performed to understand the dependency in theory of features with each other, and each of the features with cardiac disease.
3. **Model Building** - Build 6 classification models to find the one that best suits our data and the one which can predict cardiac arrest for a new set of data with highest accuracy

- Logistic Regression
- Decision tree
- Random forest
- K nearest Neighbor
- Gaussian Naive Bayes
- Support vector Classifier

4. **Prediction** - The dataset was divided for training and test with 80% of data for training. The test dataset was used to test the accuracy of the prediction by each of the models trained.
5. **Results Comparison** - Train accuracy, Test accuracy, Precision, Recall, F1-Score, Sensitivity and Specificity of each of the models were calculated and compared to find the best model with accurate results

4.Data Cleaning and Pre-processing

Data cleaning is the process of removing incorrect, duplicate, or otherwise erroneous data from a dataset.

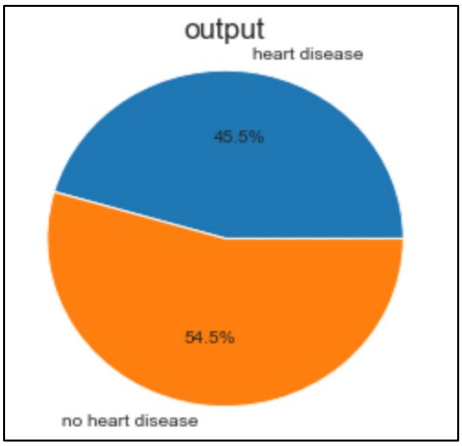
Handling Null Values

Column thall has two records with null values. As thall is a blood disorder and only two records have null values, the records are removed.

	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 Check

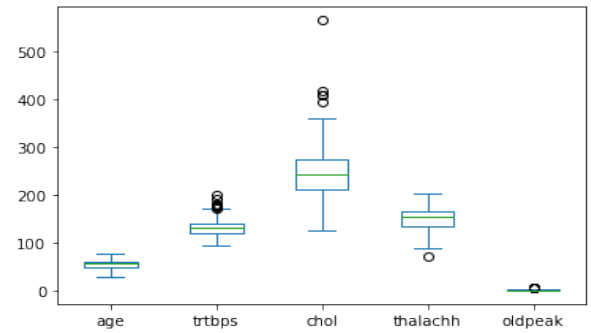
There is no class imbalance in the output column.



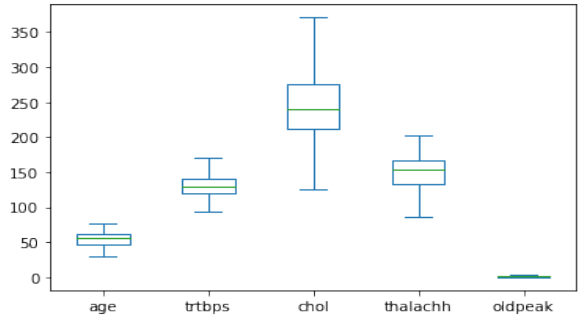
Detecting and Handling Outliers

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. This is implemented for the thalachh column. The upper limit is equated to any outliers that are higher than it and is implemented for trtbps, chol, and oldpeak columns.

Creation of NEW Columns:

Instead of just building our model based on current data, new three columns have been created using ***K means clustering method*** with number of clusters two to find whether these factors actually affect the heart disease or not.

a. Diet
b. Smoke
c. Physical Activity

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

Scaling:

Scaling is simply transforming your data such that it fits a common scale. The data itself doesn't change, only the range of the data is modified when scaled.

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

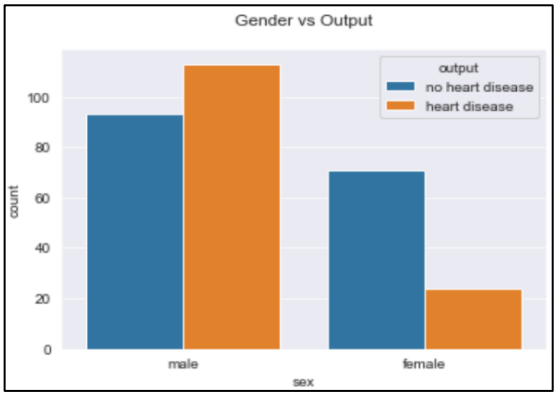
	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output	diet	physical activity	smoke
0	0.947829	1	3	0.824784	-0.260283	1	0	0.009222	0	1.145623	0	0	1	1	1	1	0
1	-1.910633	1	2	-0.078417	0.096882	0	1	1.639390	0	2.226320	0	0	2	1	1	0	0
2	-1.470869	0	1	-0.078417	-0.869563	0	0	0.978511	0	0.335100	2	0	2	1	1	0	0
3	0.178243	1	1	-0.680552	-0.197254	0	1	1.242863	0	-0.205249	2	0	2	1	1	0	0
4	0.288184	0	0	-0.680552	2.281887	0	1	0.581984	1	-0.385365	2	0	2	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
296	0.288184	0	0	0.523717	-0.092205	0	1	-1.180361	1	-0.745597	1	0	3	0	1	1	1
297	-1.031106	1	3	-1.282686	0.391017	0	1	-0.783833	0	0.154984	1	0	3	0	1	0	1
298	1.497533	1	0	0.764570	-1.100669	1	1	-0.387306	0	2.136262	1	2	3	0	0	1	0
299	0.288184	1	0	-0.078417	-2.403269	0	1	-1.532830	1	0.154984	1	1	3	0	0	1	0
300	0.288184	0	1	-0.078417	-0.197254	0	0	1.066628	0	-0.925714	1	1	2	0	1	0	0

5.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.

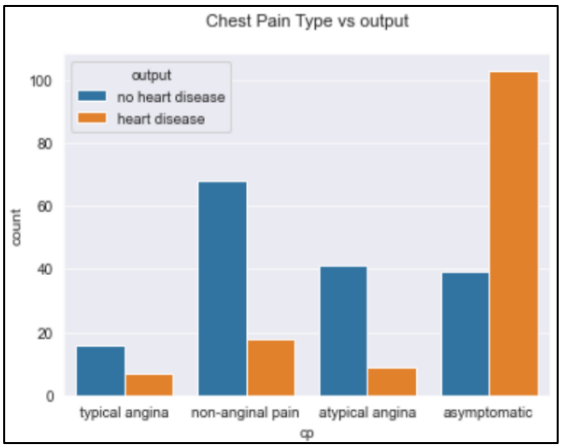
i. Gender vs output

➢ From the graph it is visible that Female has low chance of heart attack compared to male.



ii. Chest pain type vs output

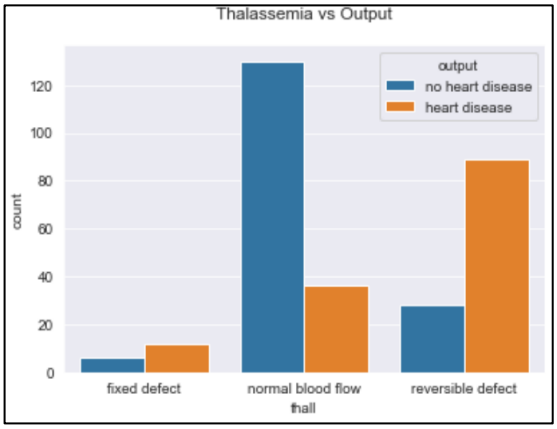
➢ This is the most contributing factor among all the other aspects because people who experience asymptomatic chest pain are more likely to suffer a heart attack than people who experience other types of chest pain



iii. Thalassemia vs output

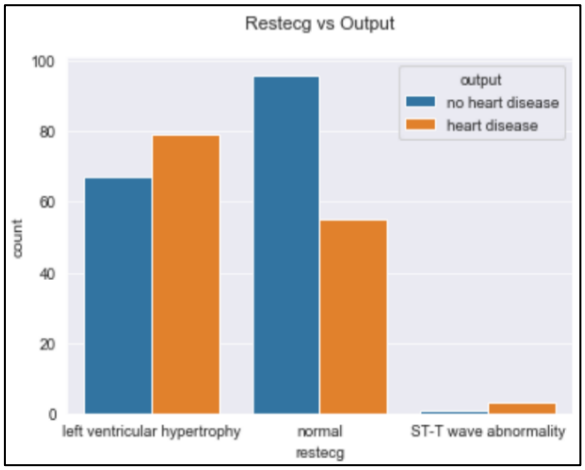
➢ Thalassemia is a kind of blood disorder due to insufficient hemoglobin.

➢ People who have reversible blood flow and fixed defect are more prone to more heart attack.



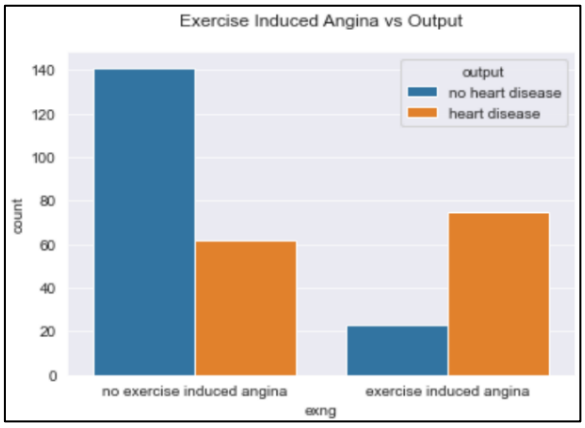
iv. Restecg vs output

➢ People who have left ventricular hypertrophy are more prone to heart attack



v. Exercise induced angina vs output

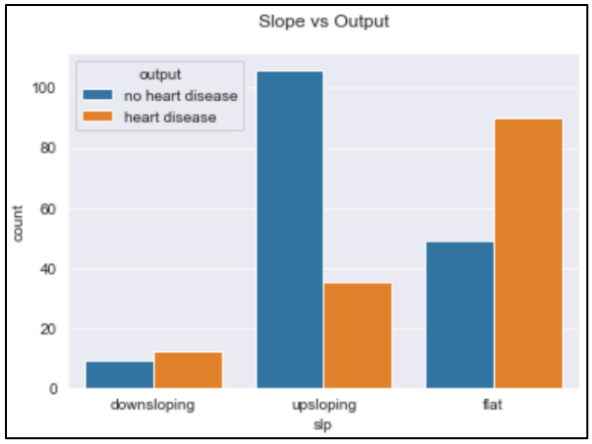
➢ People who experience discomfort while exercising are more likely to suffer a heart attack. People who engage in strenuous exercise, such as jogging or working out, frequently exhibit this.



vi. Slope vs output

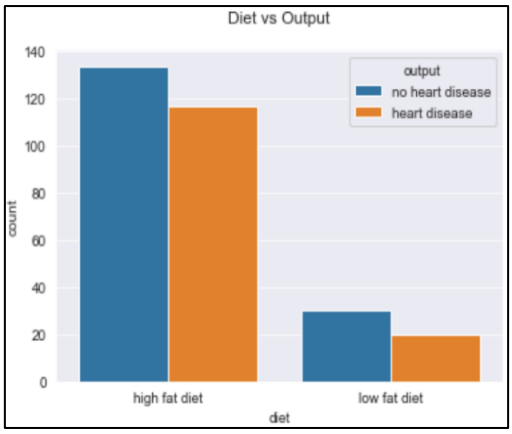
➢ Slope is defined as the slope of a line formed from a graph of the electrocardiogram (ECG) that contains both peaks and depressions. The heart does not beat frequently if the slope is flat.

➢ People with flat types of slopes are more susceptible to heart attacks since their hearts don't contract and expand as much.



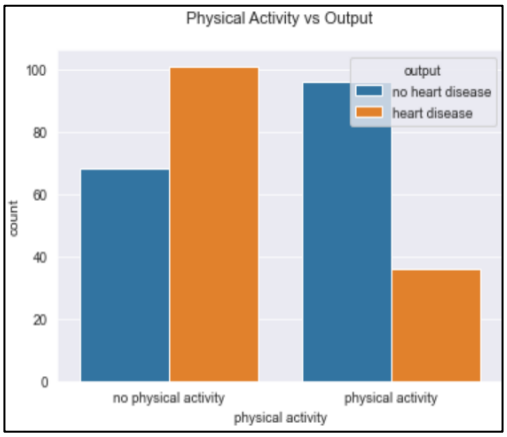
vii. Diet vs output

- The graph demonstrates that nutrition has little effect on output, making it a minor contributing factor.



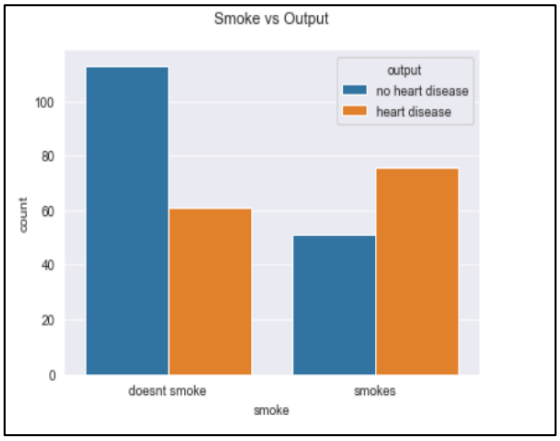
viii.Physical activity vs output

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



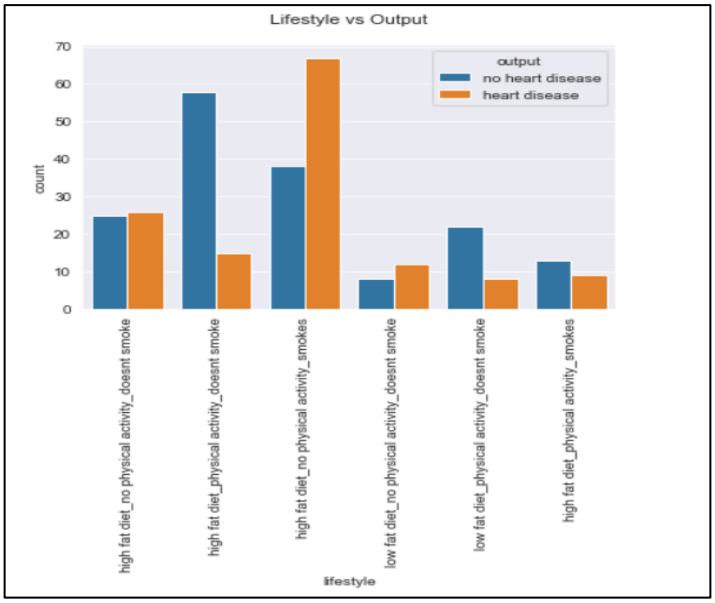
ix. Smoke vs Output

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

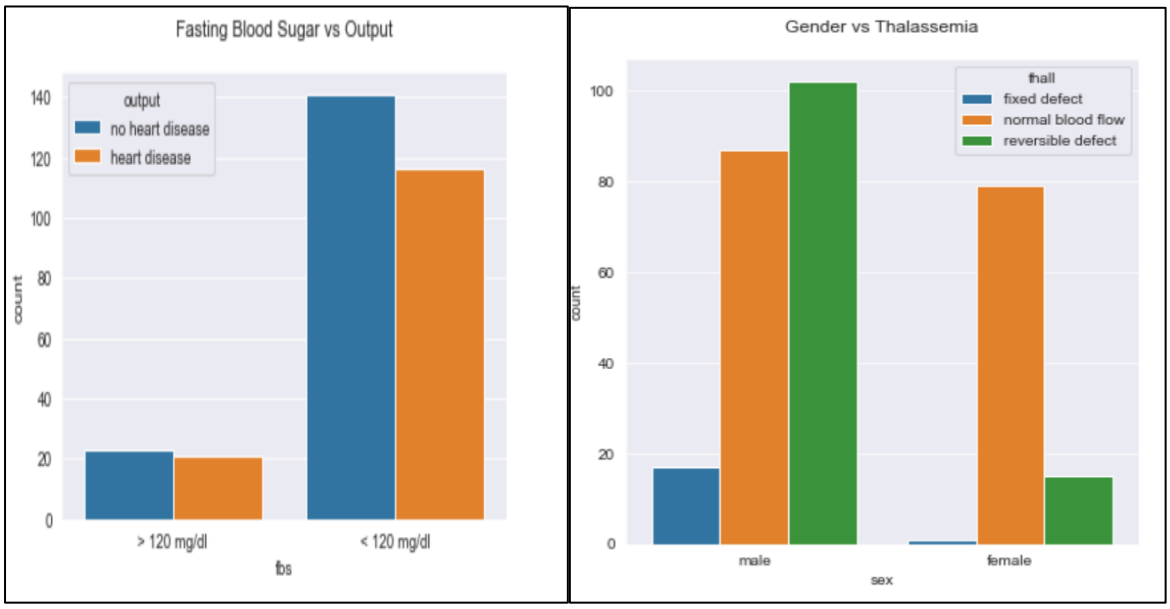


x. Lifestyle vs Output

➢ 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.



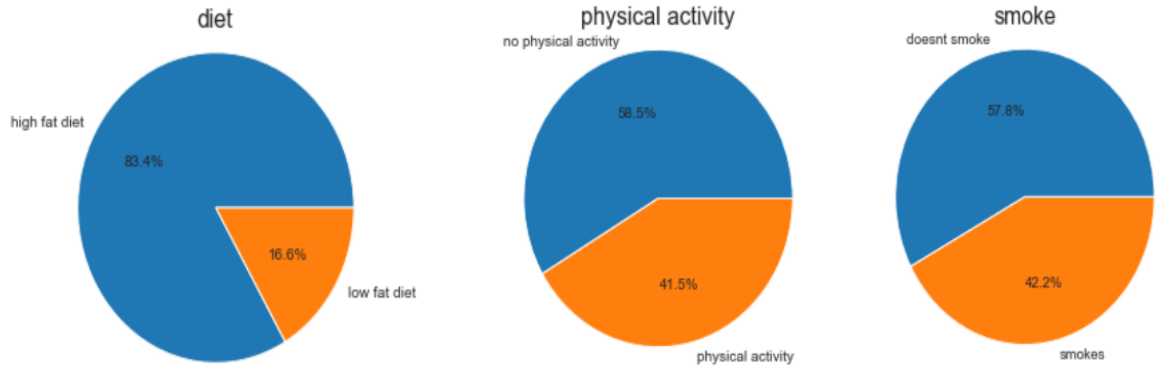
xi. Fasting Blood Sugar vs Output, Gender vs Thalassemia:



➢ **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.

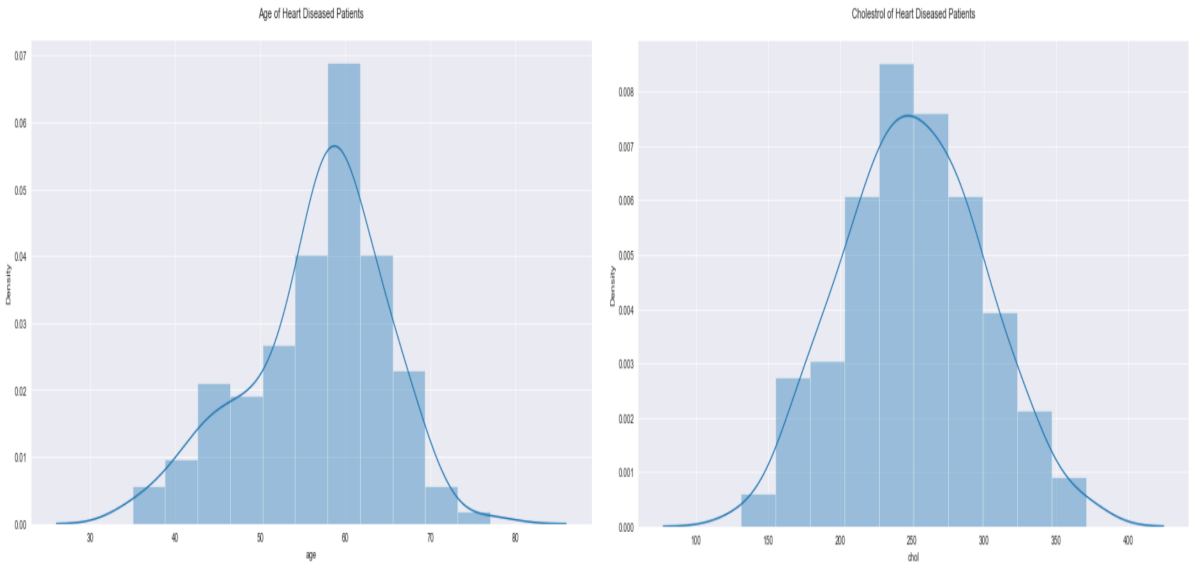
Statistical Interpretation of Diet, Smoke habits and Physical Activity



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%).

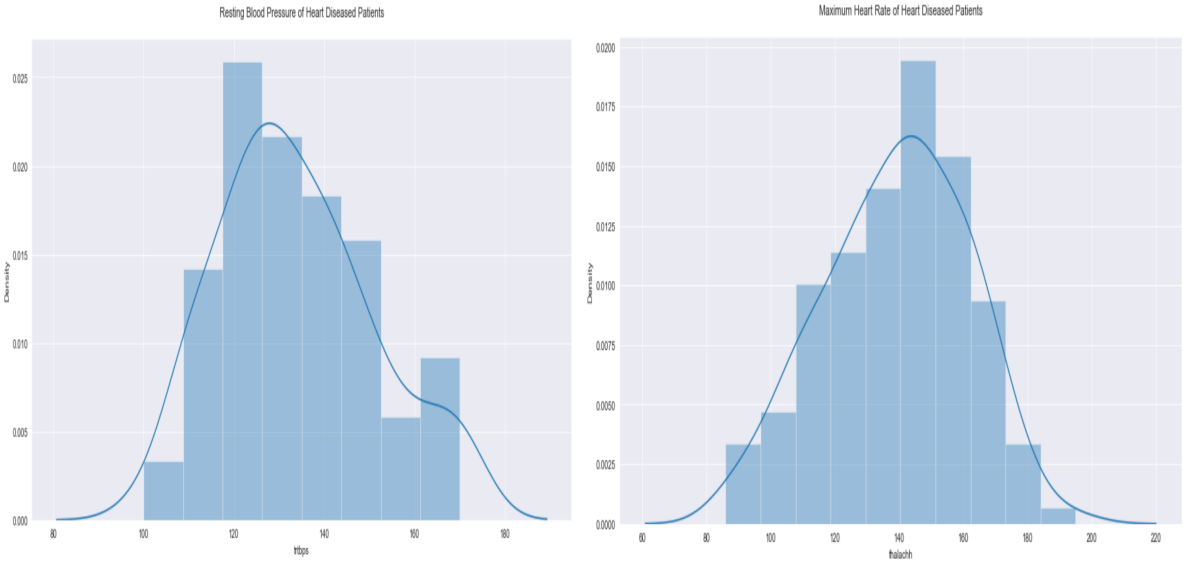
Statistical Distribution of Age, Cholesterol, Heart Rate, Blood Pressure on Heart Disease Patients

1. ***Age, Cholesterol:***



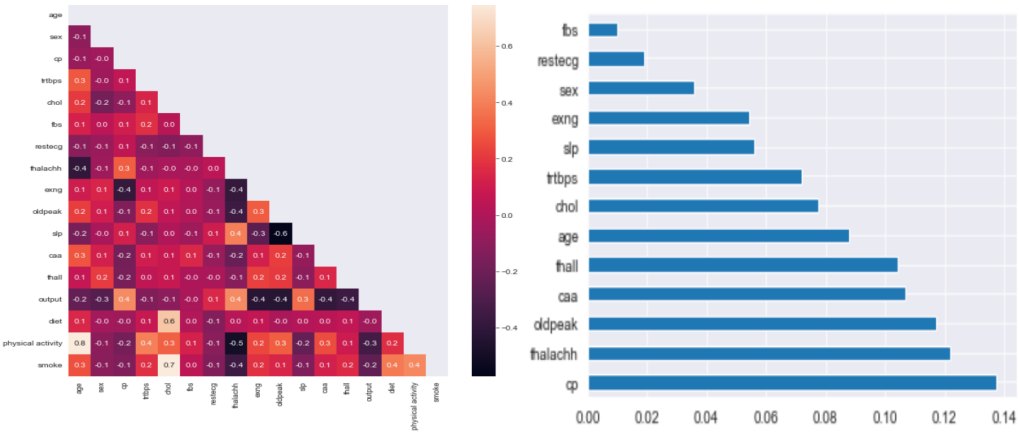
- Heart Disease is very common in the seniors, which is composed of age group centred at 60 and above and common among adults which belong to the age group of 40 to 60 (range) and is clustered between 55 to 65 age group. But it's rare among the age group of 19 to 40 and very rare among the age group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

2. *Resting Blood Pressure, Maximum Heart Rate:*



- The graph of the Resting Blood pressure distribution of heart disease 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 disease patients shows the Maximum Heart Rate levels ranging from 100 to 180, the mean is around 145, and is clustered between 120 to 170.

6.Feature Selection:



- ➢ 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.

7. Model Building

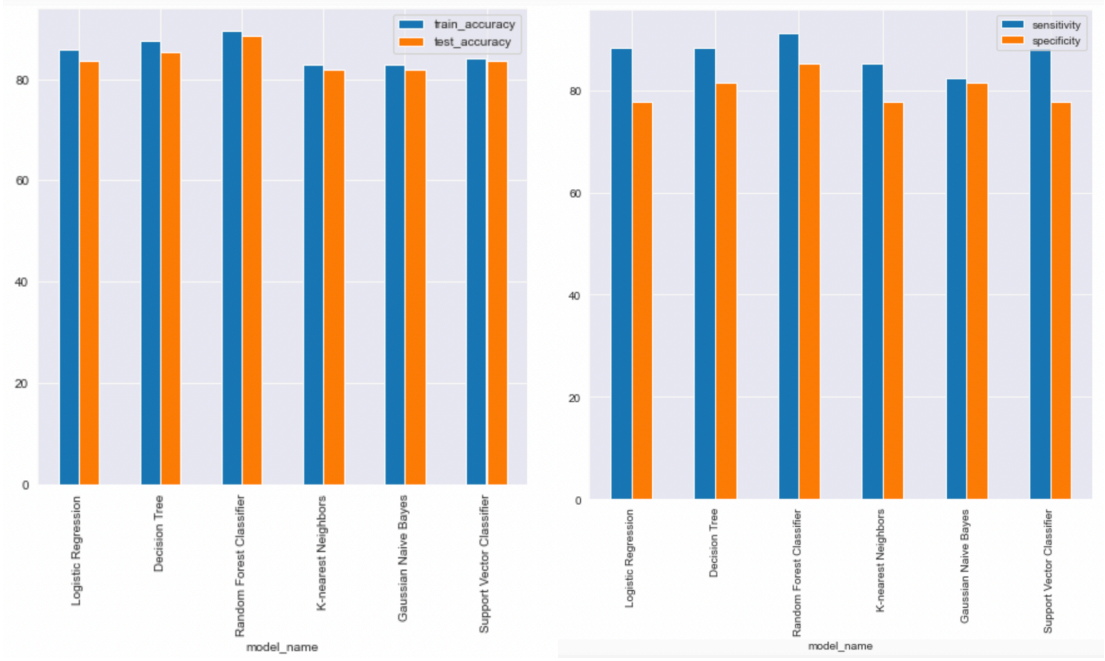
As there are no challenges in the data like multicollinearity or imbalanced target variable, all of the classification models are implemented to check which model works best for the taken dataset.

Models implemented:

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

8. 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**.



9.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.

10.Future Scope

- Current dataset has only 300 records. So, more data can be collected to re-train the model and improve accuracy of the predictions.
- Data can be collected for different cardiovascular diseases and AI can be used to identify a connection between different illnesses and recommend apt treatment and medication.