GROUP 6

GROUP MEMBERS:

ROHAN DESAI

TANVI ZUNJARRAO

MANEE UPADHYEE

Cardiac Data Analysis

Report

Table of Contents

Overview

Cardiac diseases are one of the leading causes of death in middle aged and old aged population. About 610,000 die of heart diseases in United States every year, that's 1 in 4 deaths.

Healthcare is an inevitable task to be done in human life. Cardiovascular disease is a broad category for a range of diseases that are affecting heart and blood vessels. The early methods of forecasting the cardiovascular diseases helped in making decisions about the changes to have occurred in high-risk patients which resulted in the reduction of their risks.

Age, gender, smoking, family history, cholesterol, poor diet, high blood pressure, obesity, physical inactivity, and alcohol intake are risk factors for heart disease, and hereditary risk factors such as diabetes also lead to heart disease. But it is difficult to manually determine the probability of getting a heart disease based on the risk factors.

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels and they include:

- coronary heart disease – disease of the blood vessels supplying the heart muscle
- cerebrovascular disease – disease of the blood vessels supplying the brain
- peripheral arterial disease – disease of blood vessels supplying the arms and legs
- rheumatic heart disease – damage to the heart muscle and heart valves from rheumatic fever, caused by streptococcal bacteria
- congenital heart disease – malformations of heart structure existing at birth
- deep vein thrombosis and pulmonary embolism – blood clots in the leg veins, which can dislodge and move to the heart and lungs

Heart attacks and strokes are usually acute events and are mainly caused by a blockage that prevents blood from flowing to the heart or brain. The most common reason for this is a build-up of fatty deposits on the inner walls of the blood vessels that supply the heart or brain. Strokes can also be caused by bleeding from a blood vessel in the brain or from blood clots. The cause of heart attacks and strokes are usually the presence of a combination of risk factors, such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol, hypertension, diabetes and hyperlipidemia. These heart diseases are more prevalent in men than women. More than half of deaths related to heart diseases in 2009 were in men.

The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These “intermediate risks factors” can be measured in primary care facilities and indicate an increased risk of developing a heart attack, stroke, heart failure and other complications.

Cessation of tobacco use, reduction of salt in the diet, consuming fruits and vegetables, regular physical activity and avoiding harmful use of alcohol have been shown to reduce the risk of cardiovascular disease. In addition, drug treatment of diabetes, hypertension and high blood lipids may be necessary to reduce cardiovascular risk and prevent heart attacks and strokes. Health policies that create conducive environments for making healthy choices affordable and available are essential for motivating people to adopt and sustain healthy behavior.

Cardiovascular Diseases - A development issue in low- and middle-income countries

- At least three quarters of the world's deaths from CVDs occur in low- and middle-income countries.
- People in low- and middle-income countries who suffer from CVDs and other noncommunicable diseases have less access to effective and equitable health care services which respond to their needs. As a result, many people in low- and middle-income countries are detected late in the course of the disease and die younger from CVDs and other noncommunicable diseases, often in their most productive years.
- The poorest people in low- and middle-income countries are affected most. At the household level, enough evidence is emerging to prove that CVDs and other noncommunicable diseases contribute to poverty due to catastrophic health spending and high out-of-pocket expenditure.
- At macro-economic level, CVDs place a heavy burden on the economies of low- and middle-income countries.

Objective

The health care industry contains lots of medical data, therefore machine learning algorithms are required to make decisions effectively in the prediction of heart diseases. Recent research has delved into uniting these techniques to provide hybrid machine learning algorithms. In the proposed research, data pre-processing uses techniques like the removal of noisy data, removal of missing data, filling default values if applicable and classification of attributes for prediction and decision making at different levels. The performance of the diagnosis model is obtained by using methods like classification, accuracy, sensitivity and specificity analysis. This project proposes a prediction models to predict whether a person has a heart disease or not and to provide an awareness or diagnosis on that. This is done by comparing the accuracies of applying rules to the individual results of Support Vector Machine and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease.

Dataset

The dataset has been taken from the following website :

https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

The dataset consists of 14 attributes and 303 instances. There are 8 categorical attributes and 6 numeric attributes. The description of the dataset is shown below:

S.No	Attribute Name	Description	Range of Values
1	Age	Age of the person in years	29 to 79
2	Sex	Gender of the person [1: Male, 0: Female]	0, 1
3	Cp	Chest pain type [1-Typical Type 1 Angina 2- Atypical Type Angina 3-Non-angina pain 4-Asymptomatic)	1, 2, 3, 4
4	Trestbps	Resting Blood Pressure in mm Hg	94 to 200
5	Chol	Serum cholesterol in mg/dl	126 to 564
6	Fbs	Fasting Blood Sugar in mg/dl	0, 1
7	Restecg	Resting Electrocardiographic Results	0, 1, 2
8	Thalach	Maximum Heart Rate Achieved	71 to 202
9	Exang	Exercise Induced Angina	0, 1
10	OldPeak	ST depression induced by exercise relative to rest	1 to 3
11	Slope	Slope of the Peak Exercise ST segment	1, 2, 3

S.No	Attribute Name	Description	Range of Values
12	Ca	Number of major vessels colored by fluoroscopy	0 to 3
13	Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect	3, 6, 7
14	Target	Class Attribute	0 or 1

Dataset Description

Patients from age 29 to 79 have been selected in this dataset. Male patients are denoted by a gender value 1 and female patients are denoted by gender value 0. Four types of chest pain can be considered as indicative of heart disease. Type 1 angina is caused by reduced blood flow to the heart muscles because of narrowed coronary arteries. Type 1 Angina is a chest pain that occurs during mental or emotional stress. Non-angina chest pain may be caused due to various reasons and may not often be due to actual heart disease. The fourth type, Asymptomatic, may not be a symptom of heart disease. The next attribute *trestbps* is the reading of the resting blood pressure. *Chol* is the cholesterol level. *Fbs* is the fasting blood sugar level; the value is assigned as 1 if the fasting blood sugar is below 120 mg/dl and 0 if it is above. *Restecg* is the resting electrocardiographic result, *thalach* is the maximum heart rate, *exang* is the exercise induced angina which is recorded as 1 if there is pain and 0 if there is no pain, *oldpeak* is the ST depression induced by exercise, *slope* is the slope of the peak exercise ST segment, *ca* is the number of major vessels colored by fluoroscopy, *thal* is the duration of the exercise test in minutes, and *num* is the class attribute. The target attribute has a value of 0 for normal and 1 for patients diagnosed with heart disease.

Indicators Description

1. **Age**: Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. Coronary fatty streaks can begin to form in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55.

2. **Sex**: Men are at greater risk of heart disease than pre-menopausal women. Once past menopause, it has been argued that a woman's risk is similar to a man's although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease than a male with diabetes.

3. **Angina (Chest Pain)**: Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion.

4. **Resting Blood Pressure**: Over time, high blood pressure can damage arteries that feed your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol or diabetes, increases your risk even more.

5. **Serum Cholesterol**: A high level of low-density lipoprotein (LDL) cholesterol (the "bad" cholesterol) is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to your diet, also ups your risk of a

heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the "good" cholesterol) lowers your risk of a heart attack.

6. **Fasting Blood Sugar**: Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body's blood sugar levels to rise, increasing your risk of a heart attack.

7. **Resting ECG**: For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits. For people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening.

8. **Max heart rate achieved**: The increase in cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.

9. **Exercise induced angina**: The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe. Angina is usually felt in the center of your chest but may spread to either or both of your shoulders, or your back, neck, jaw or arm. It can even be felt in your hands. o Types of Angina a. Stable Angina / Angina Pectoris b. Unstable Angina c. Variant (Prinzmetal) Angina d. Microvascular Angina.

10. **Peak exercise ST segment**: A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80 ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an 'equivocal' test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel disease. The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test. Another finding that is highly indicative of significant CAD is the occurrence of ST-segment elevation > 1 mm (often suggesting transmural ischemia); these patients are frequently referred urgently for coronary angiography.

Data Cleaning

Datasets in a perfect world is a perfectly curated group of observations with no missing values or anomalies. However, this is not true. Real world data comes in all shapes and sizes. It can be messy, which means it needs to be clean and wrangles. Data cleaning is a necessary part in data science problems. Machine learning models learn from data. It is crucial, however, that the data you feed them is specifically preprocessed and refined for the problem you want to solve. This includes data cleaning, preprocessing, feature engineering, and so on.

The data cleaning process in this project includes removal of null values from dataset.

Raw data also consisted of some junk values and unwanted rows which we removed by using python script. We also changed names of columns to easily understand data since raw column names were difficult to understand.

Raw dataset was cleaned by using python and then it was exported in csv format which we used in tableau and R for analysis.

Data Visualization

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it's processed.

Why is data visualization important?

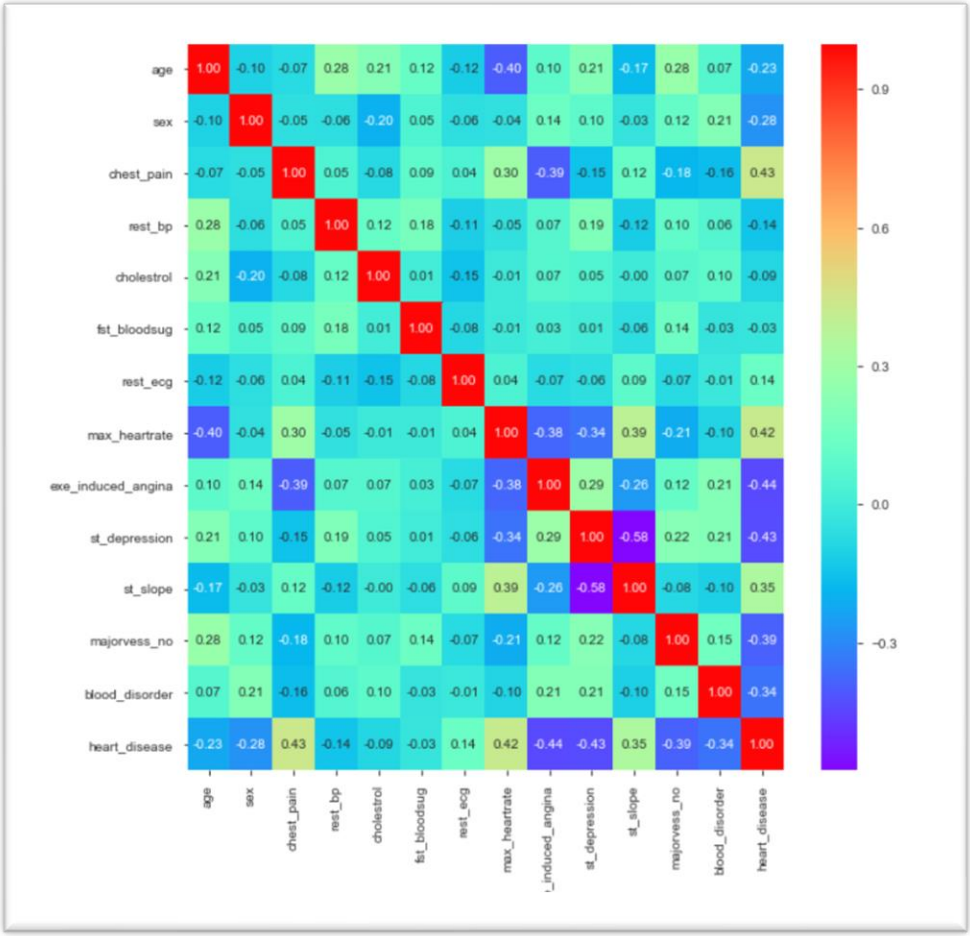
Because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports. Data visualization is a quick, easy way to convey concepts in a universal manner – and you can experiment with different scenarios by making slight adjustments.

In healthcare, the use of interactive visualizations is to represent data/information related to determinants of health and healthcare indicators, and to investigate the benefits of such visualizations for health policymaking.

Correlation Matrix

To begin with, let's see the correlation matrix of features and try to analyse it. The figure size is defined to 12 x 8 by using rcParams. The pyplot is used to show the correlation matrix. Using xticks and yticks in the code, names are added to the correlation matrix. colorbar() shows the colorbar for the matrix.

It's easy to see that there is no single feature that has a very high correlation with our target value. Also, some of the features have a negative correlation with the target value and some have positive.

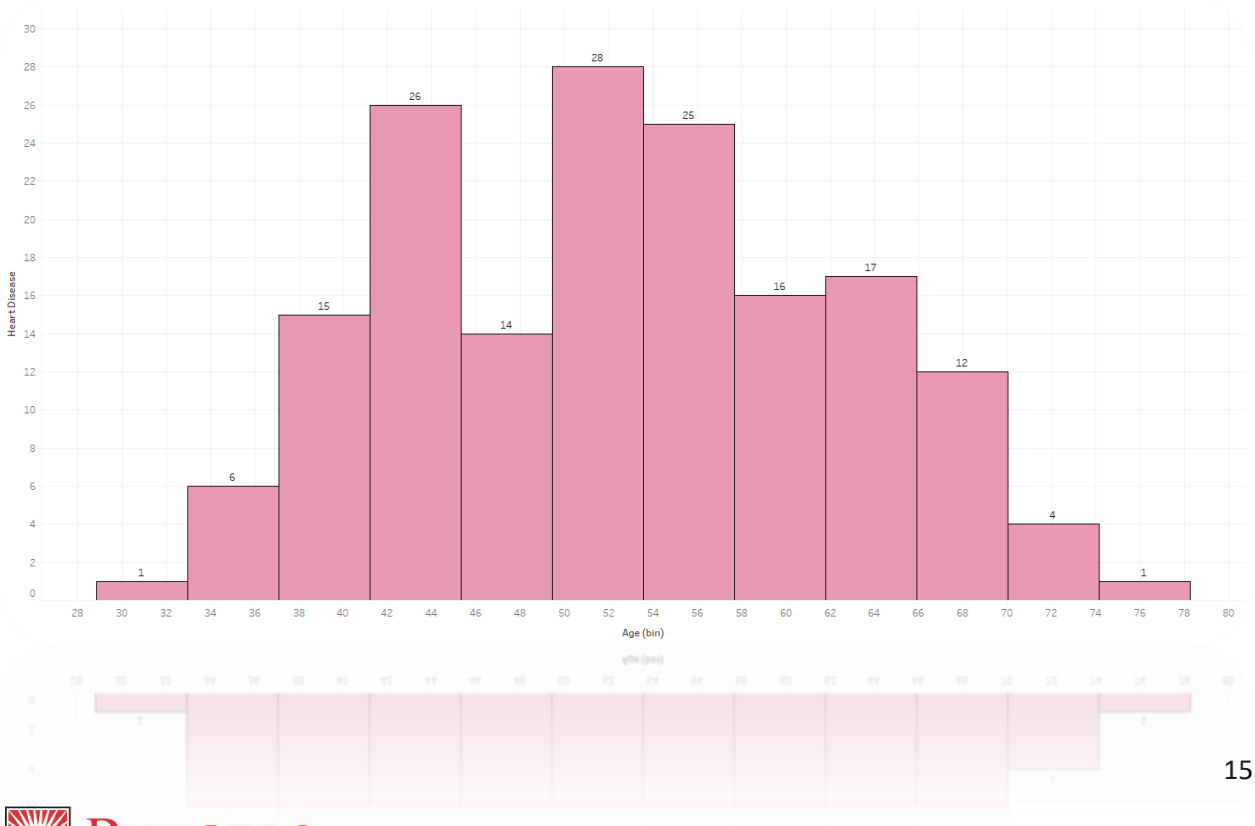


Analysis by Gender

According to the data and on the contrary to the general belief, heart diseases are more prevalent in women as compared to men. The graph below shows that the total count of males is 207 and the total count of females are 96. Based on this, 75% of females and 45% of makes suffer from heart diseases.

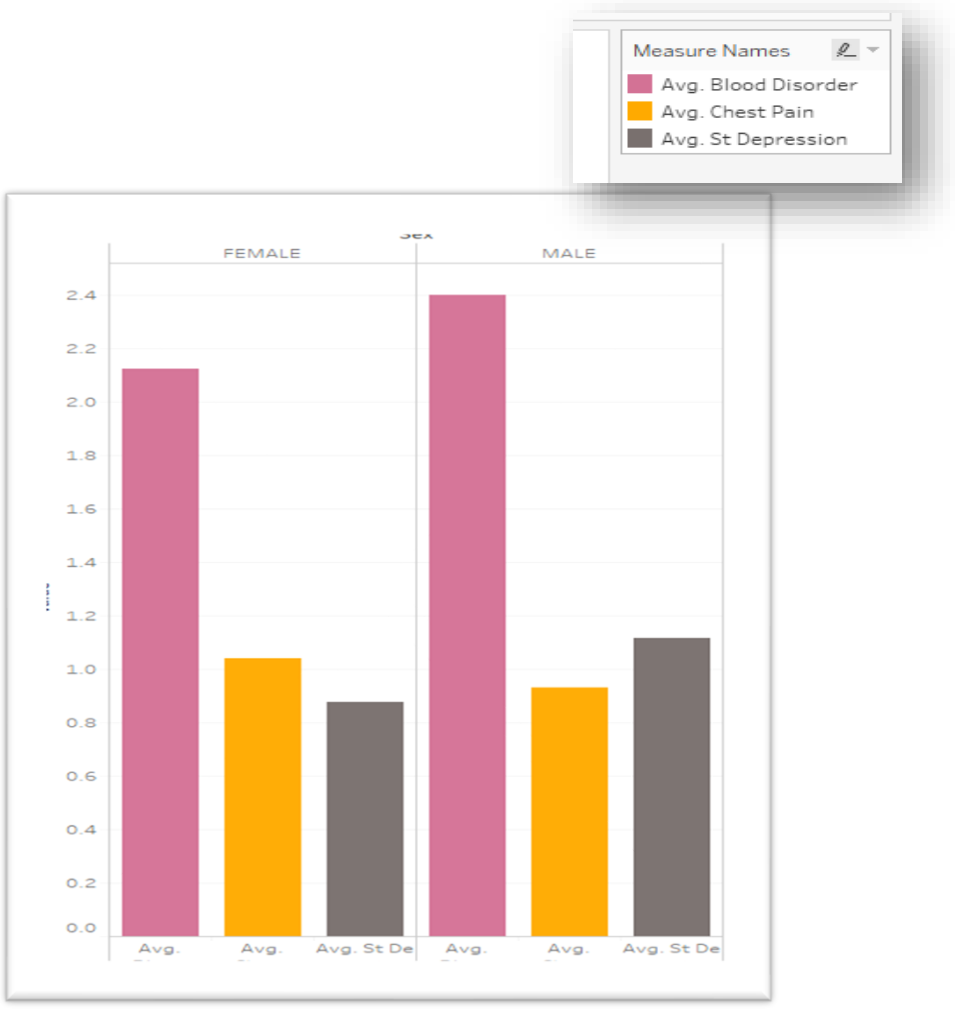
Analysis by Age group

The chart below has the data of people of different age groups ranging from 28 to 78. Here we can see that the age group falling between 38 years to 66 years have had the highest number of heart related disease.



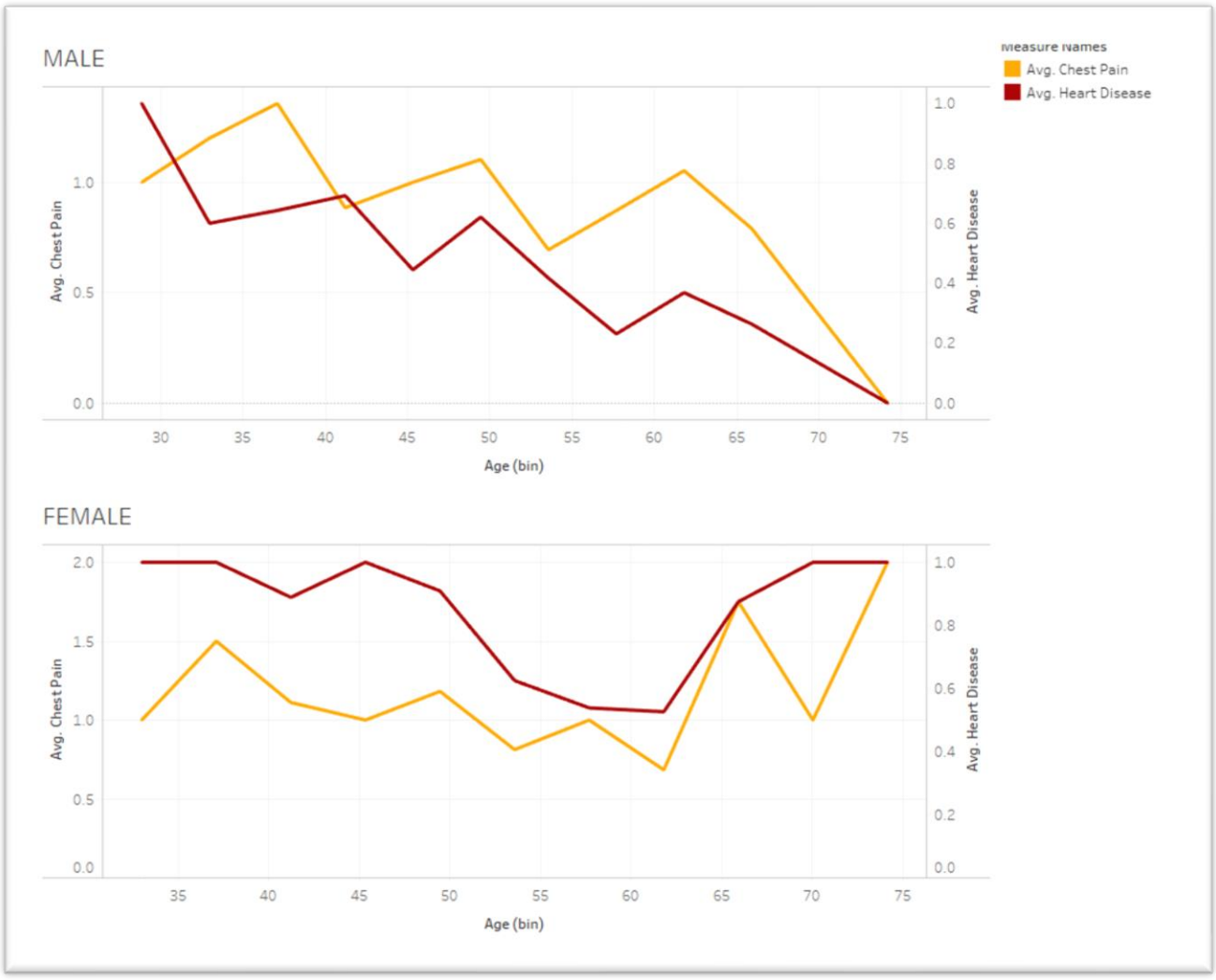
Analysis by Blood Disorder, Chest pain and Depression

Here we are correlating chest pain & depression with heart diseases. It's not conclusive whether chest pain related issues or depression may cause heart disease. But we can observe that chest pain is more in female than in male.



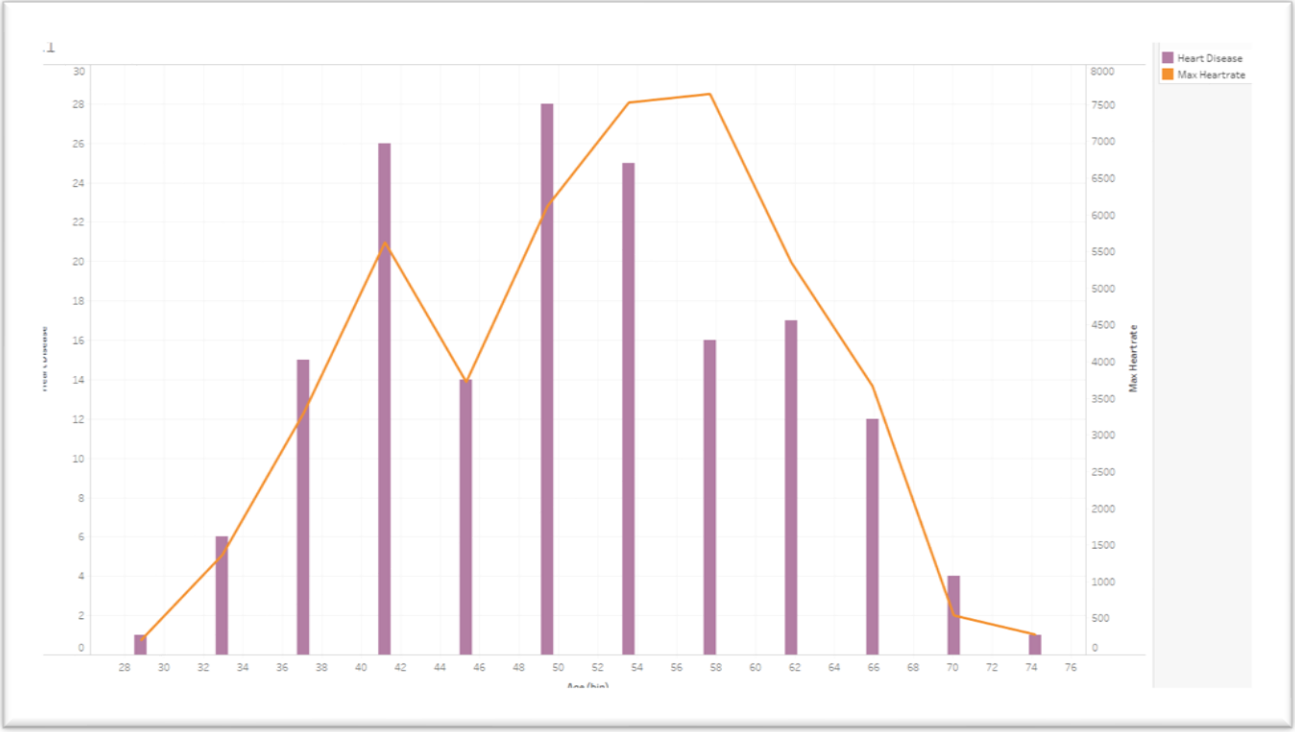
Analysis by Chest pain and Heart disease

Here by plotting against age, it's clear that the chest pain issues do strongly corelate with heart diseases which means chest pain issues could cause heart diseases in men.



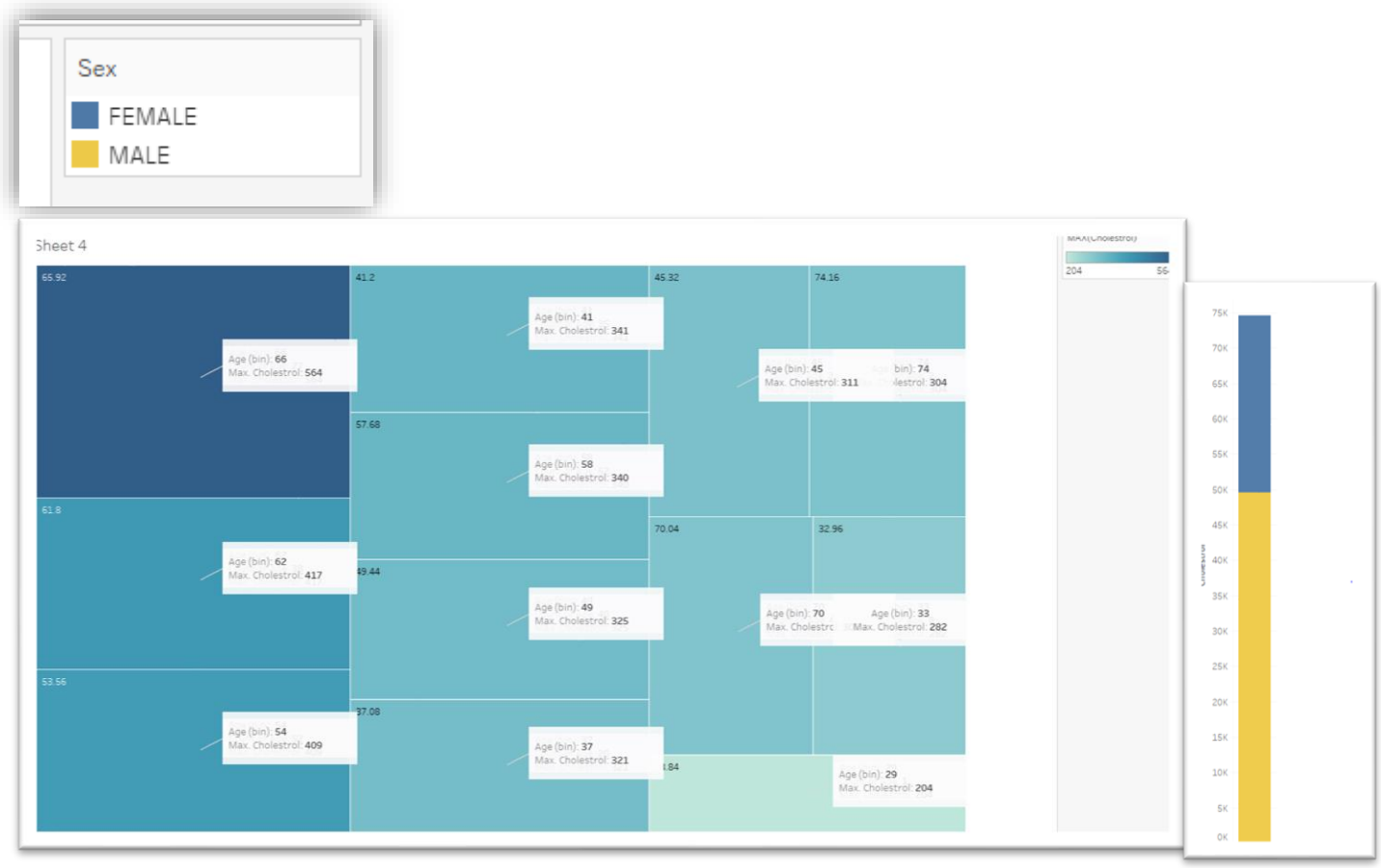
Analysis by Max Heartrate

Here by plotting Heartrate and heart diseases we observe that as heartrate increases the probability of heart disease.



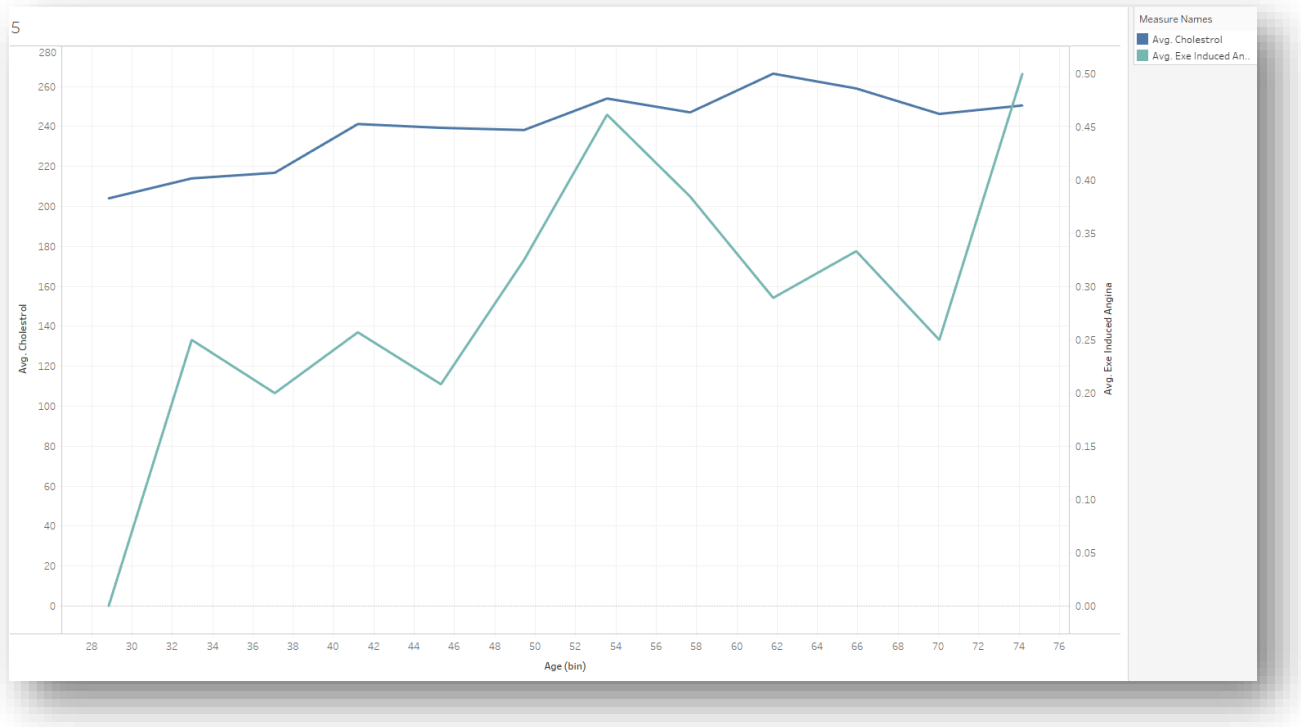
Level of Cholesterol

We observe that as age increases the level of cholesterol also increases. The highest level of cholesterol is at age 66.



Relationship between Induced Angina and Cholesterol

Angina is basically chest pain caused by reduced blood flow to the heart. We observe that as cholesterol increases chance of having angina also increases.



Machine Learning Models

In this project, I took 2 algorithms and varied their various parameters and compared the final models. I split the dataset into 70% training data and 30% testing data.

Logistic Regression

Logistic regression is a statistical model that in its basic form uses a **logistic** function to model a binary dependent variable, although many more complex extensions exist. In **regression** analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a **logistic** model (a form of binary **regression**). Here, the combined accuracy of logistic regression gives the accuracy result of 87.79%.

```
In [25]:
#Logistic regression
from sklearn.linear_model import LogisticRegression

accuracy={}
regression= LogisticRegression(solver="lbfgs",C=1000,random_state=0)
regression.fit(f_train_std,t_train)
print("\nFor Logestic Regression:")
eval_performance(regression,f_train_std,f_test_std,t_train,t_test)
accuracy['Logistic Regression']= round(regression.score(f_test,t_test)*100,2)

For Logestic Regression:
Accuracy of test data set:84.62%
Number in test data set:  91
Misclassified samples: 14
Combined Accuracy: 87.79%
Number in test  303
Misclassified samples: 37
```

Support Vector Machine

A support vector machine (SVM) is machine learning algorithm that analyzes data for classification and regression analysis.

SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible.

SVMs are used in text categorization, image classification, handwriting recognition and in the sciences.

This classifier aims at forming a hyperplane that can separate the classes as much as possible by adjusting the distance between the data points and the hyperplane. There are several kernels based on which the hyperplane is decided. I tried four kernels namely, *linear*, *poly*, *rbf*, and *sigmoid*.

In [27]:
```
#SVM
from sklearn.svm import SVC
SVC_scores=[]


SVM_rbf = SVC(kernel='rbf', C=1.0, random_state=0)
SVM_rbf.fit(f_train_std, t_train)
print("\nFor Support Vector Machine(RBF):")
eval_performance(SVM_rbf,f_train_std,f_test_std,t_train,t_test)
SVC_scores.append(SVM_rbf.score(f_test_std,t_test))

```

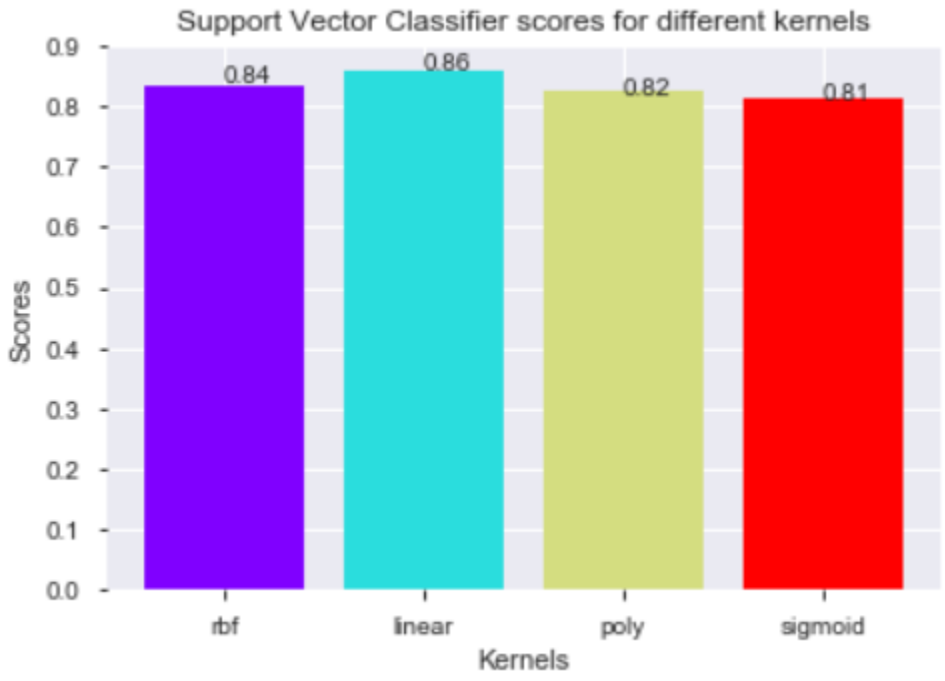
SVM accuracy result

The accuracy for test and training data has been calculated for all the four kernels. Based on the combined accuracy, the rbf kernel gives the most accurate prediction out of all.

```
For Support Vector Machine(RBF):
Accuracy of test data set:83.52%
Number in test data set:  91
Misclassified samples: 15
Combined Accuracy: 90.43%
Number in test  303
Misclassified samples: 29

For Support Vector Machine(linear):
Accuracy of test data set:85.71%
Number in test data set:  91
Misclassified samples: 13
Combined Accuracy: 88.78%
Number in test  303
Misclassified samples: 34

For Support Vector Machine(poly):
Accuracy of test data set:82.42%
Number in test data set:  91
Misclassified samples: 16
Combined Accuracy: 89.77%
Number in test  303
Misclassified samples: 31

For Support Vector Machine(sigmoid):
Accuracy of test data set:81.32%
Number in test data set:  91
Misclassified samples: 17
Combined Accuracy: 83.83%
Number in test  303
Misclassified samples: 49
{'Logistic Regression': 85.71, 'SVM': 85.71}
```

SVM Scores based on test data

As can be seen from the plot below, the linear kernel performed the best for this dataset and achieved a score of 86%.



Conclusion

Heart Disease is one of the major concerns for the society today. It is difficult to manually determine the odds of getting heart disease based on risk factors. However, machine learning techniques are useful to predict the output from existing data.

Future Scope

The scary thing for most patients after they have had a heart attack is they want to know whether they are going to have another one and unfortunately where the field of cardiology lies right now, this question cannot be answered reliably so the answer is usually no one knows.

Whether we are witnessing a short-lived phenomenon or a lasting transformational process in all this remains to be seen. However, as we have seen many times in the past, large-scale changes in consumer behavior do eventually make a difference that cannot and will not be ignored. In this case, there could be an almost perfect win-win outcome.

These heart disease prediction model can be combined with a front end user interface and a mobile or a web based application can be created wherein the users can enter their respective parameter details and the model will predict whether the person is or can suffer from heart diseases. This will help people who stay in remote areas as it can be accessed from anywhere anytime for self-diagnosis.

References

https://towardsdatascience.com/

https://www.sciencedirect.com/

https://ieeexplore.ieee.org/document/8550857

https://towardsdatascience.com/heart-disease-prediction-73468d630cfc

https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

https://en.wikipedia.org/