Northeastern University

IE6600 Computation and Visualization Fall -2022

Group 5

Project Report

HEART DISEASES ANALYSIS

Group Members

Riddhi Gupta

Simrat Kaur

Nidhi Jagadeesh

Contribution: Contributed equally

INTRODUCTION

The biggest challenge in the medical industry is predicting and diagnosing heart disease based on factors like physical examinations, symptoms, and signs of the patient. In addition to identifying anomalies in a dataset, it also reveals hidden structure, and it builds parsimonious models to test the assumptions. Heart disease is influenced by cholesterol levels, smoking habits, obesity, family history of diseases, blood pressure, and working environment. With the advancement of technology, machine language can be used to deal with unstructured and exponentially growing data. K means the clustering method is proposed in a big data environment and visualization is done with a tableau dashboard.

ABSTRACT

Heart disease is regarded as the deadliest disease in human life worldwide. Specifically, in this type of disease, the heart is incapable of pumping enough blood to the remaining organs of the body to function properly analyzing the data in healthcare assists in predicting diseases, improving diagnosis, analyzing symptoms, providing appropriate medicines, improving quality of care, minimizing costs, extending life expectancy, and reducing mortality.

PROBLEM STATEMENT

Real wealth is in good health. During the epidemic, we were all aware of the brutal impact that COVID-19 had on everyone, regardless of status. For better future planning, you must analyze this health and medical data. Although there are 76 attributes in this database, all published experiments only use a subset of 14 of them. The Cleveland database in particular is the only one that ML researchers have used up until this point. The "goal" field alludes to the patient's having heart disease. It is an integer valued from 0 (no presence) to 4. Attribute Information: age> 2. sex> 3. chest pain type (4 values)> 4. resting blood pressure>serum cholesterol in mg/dl> 6. fasting blood sugar > 120 mg/dl> 7. resting electrocardiographic results (values 0,1,2)> 8. maximum heart rate achieved> 9. Exercise-induced angina> 10. old peak = ST depression induced by exercise relative to rest> 11. the slope of the peak exercise ST segment> 12. A number of major vessels (0-

3) colored by fluoroscopy> 13. Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect Find key metrics and factors and show the meaningful relationships between attributes. Do your own research and produce your findings.

AIM

To predict whether the patient has heart disease or not. Age, the type of chest pain, blood pressure, blood sugar level, resting ECG, heart rate, the four diverse types of chest pain, and exercise-induced angina are the variables here considered to predict heart disease. Pre-processing the heart disease dataset efficiently involves removing irrelevant records and assigning values to tuples that are missing.

DATA DESCRIPTIVE

The recommended technique was developed using the Cleveland heart disease raw dataset, which contains 76 features of 303 patients. During the pre-processing stage, some samples are removed in order to remove inaccuracies caused by inconsistent data. Age, the type of chest discomfort, blood pressure, blood glucose level, ECG at rest, heart rate, frequency of physical activity, and ECG are seven independent features among 209 samples that are used to predict heart disease. Recently, the database's dummy values were used in place of the patients' names and social security numbers.

The attributes used are

•age (Age in years)

•sex (1 = male; 0 = female)

• chest pain 0: asymptomatic, 1: atypical angina, 2: non-anginal pain, 3: typical angina

• trestbps (Resting Blood Pressure in mm/hg)

• cholesterol (Serum Cholesterol in mg/dl)

• fbs (Fasting Blood Sugar > 120 mg/dl): [0 = no, 1 = yes]

• restecg (Resting ECG): [0: normal, 1: having ST-T wave abnormality, 2: showing probable or definite left ventricular hypertrophy]

• thalach (maximum heart rate achieved)

• exang (Exercise Induced Angina)

•oldpeak (ST depression induced by exercise relative to rest)

•slope (the slope of the peak exercise ST segment): [0: downsloping; 1: flat; 2: upsloping]

• ca [number of major vessels (0–3)]

• thal (thalassemia) [1 = normal, 2 = fixed defect, 3 = reversible defect]

•disease

TABLEAU

Tableau is a business intelligence software that analyzes data and displays insights in the form of graphs and charts. used can develop interactive dashboards which show hidden patterns, trends, densities, and variations in the data, Tableau separates the data into K-number of clusters using the centroid-based K-means clustering method. The K-means technique is used to produce dashboards using the data set. To estimate the occurrence of heart disease from the provided dataset, it offers visually appealing clusters.

DATA ANALYTICS

This section presents the findings of the data analysis conducted to find the necessary hidden patterns for forecasting cardiac illnesses. The K-means technique is then used to put together the pre-processed heart disease data set. Here, four diverse types of heart conditions—asymptomatic pain, atypical angina pain, non-anginal pain, and non-anginal pain—are explored. All four categories of chest discomfort are used in the computation of the results together with other determining factors.

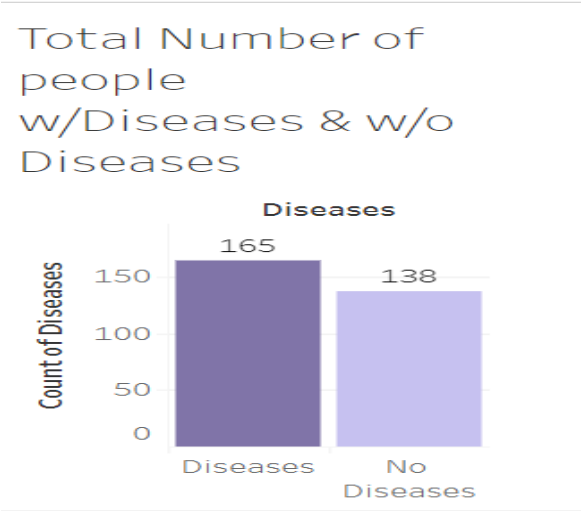

Fig1. Total Number of people w/Diseases & w/o Diseases

This Chart helps us to understand that the no. of people effected by Heart Diseases is more than the no. of people who are not effected by Heart Diseases. Further we will discuss the reasons for this

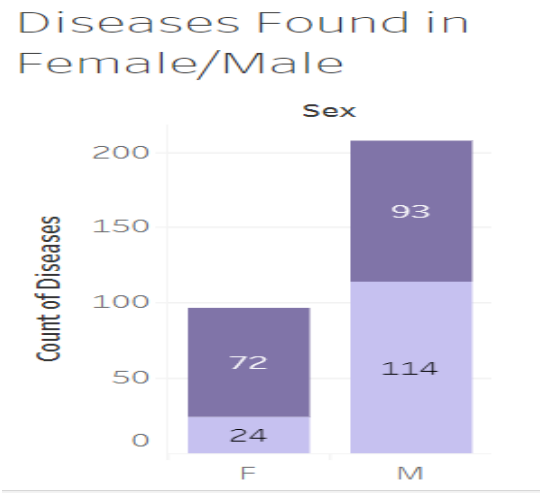

Fig 2. Diseases Found in Female/Male

When we dig further deep we can see that more males are effected by the Heart Diseases.

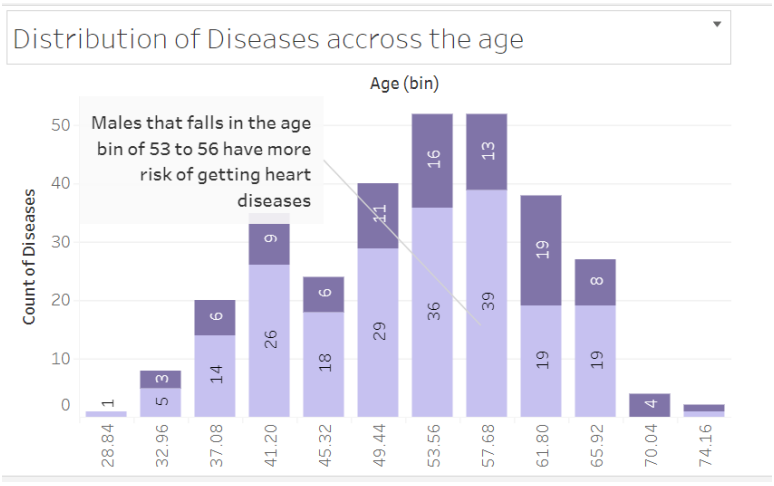

Fig 3. Distribution of Diseases across the age

Males that falls in the age bin of 53 to 56 have more risk of getting heart diseases

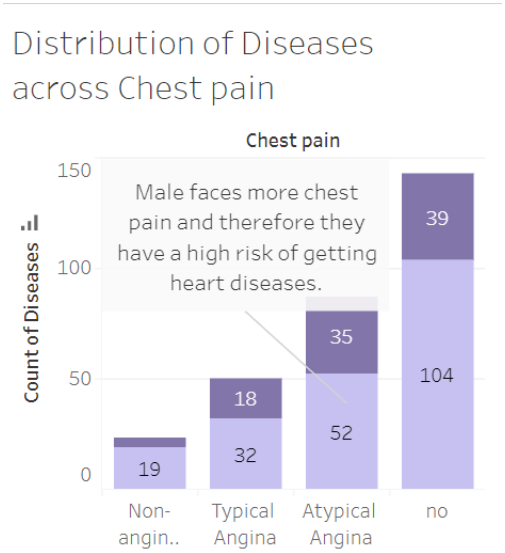

Fig.4 Distribution of Diseases across Chest pain

This sheet shows that most of the male got heart diseases because of the chest pain - Atypical Angina

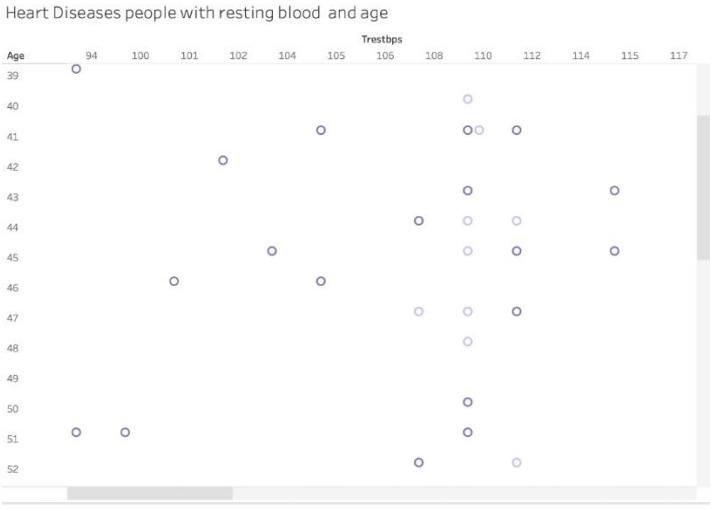

Fig.5 circle viewers of heart disease people with resting blood and age

Circle viewers given in Fig.5 depicts the distribution of ages and the resting blood pressure risk of heart disease for the targeted class. It is observed that target class with the age ranging from 29 to 70

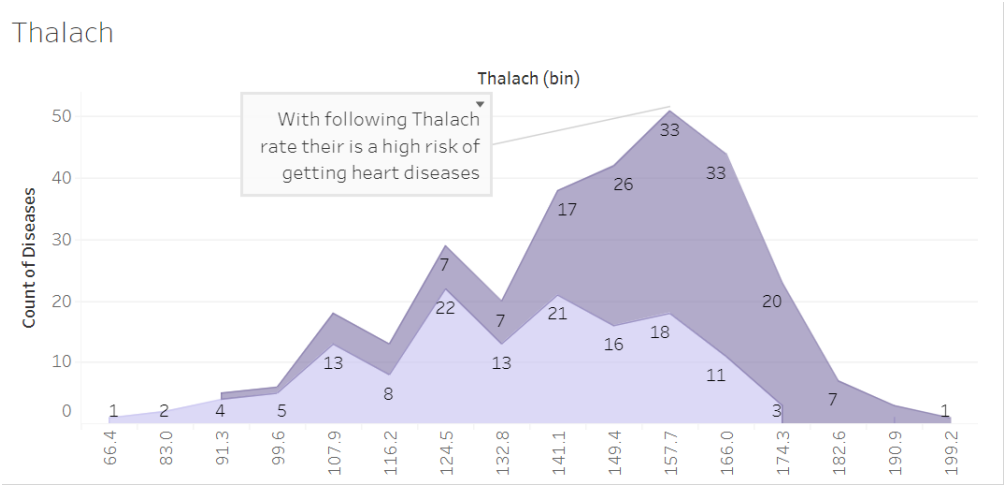

Fig.6 area chart of the maximum heart disease rate(thalach) vs count of diseases

This Chart shows that the Thalch is one of the factor for the disease. It is also shown from the area chart that if thalach rises over 140 to 160, the risk of getting heart diseases also increases.

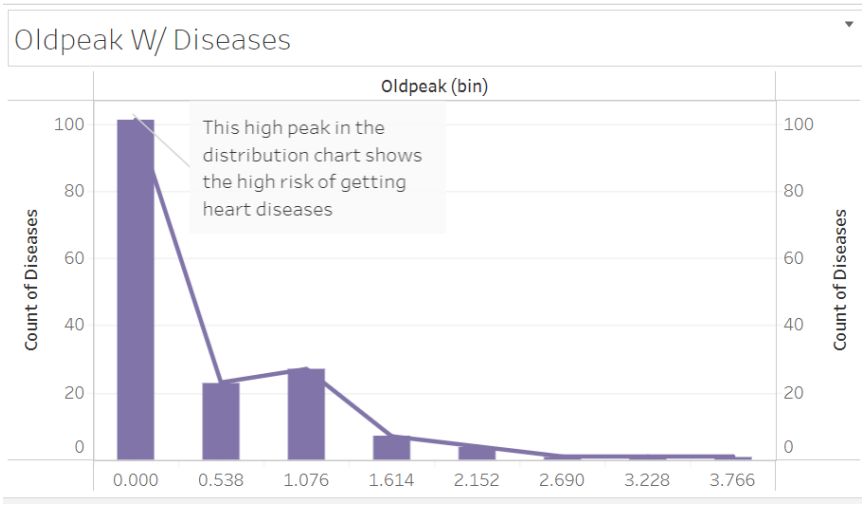

Fig.7 Distribution Plot between Oldpeak W/ Diseases

This distribution plot shows how ST depression induced by exercise relative to rest effects the heart diseases. So if i high chol./Thalch person do not do any excerise and stays in burden then risk of getting heart diseases is more.

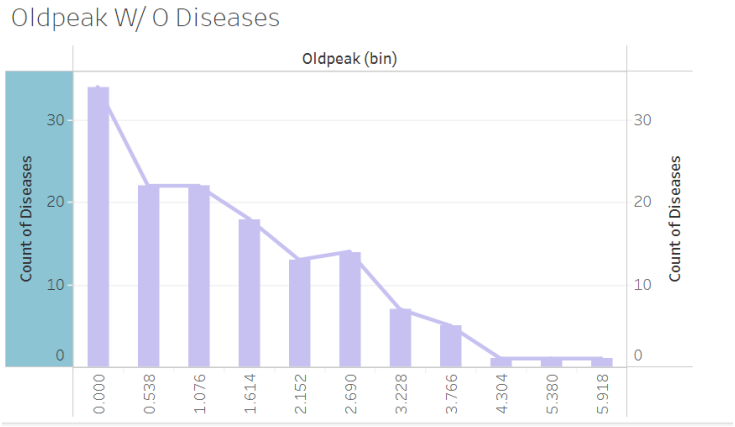

Fig.8 Distribution plot the ST depression induced by exercise relative to rest with people not having diseases

The slope of the peak exercise ST segment Value 1: upsloping vs maximum heart rate achieved

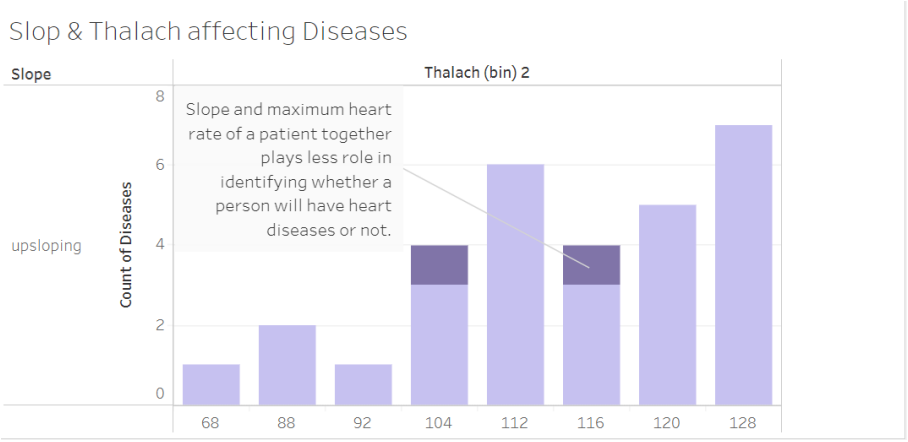

Fig.9 Slop & Thalach affecting Diseases

This Distribution plot tells us that the risk is higher for slope=2 when maximum heart rate achieved is 100 to 120

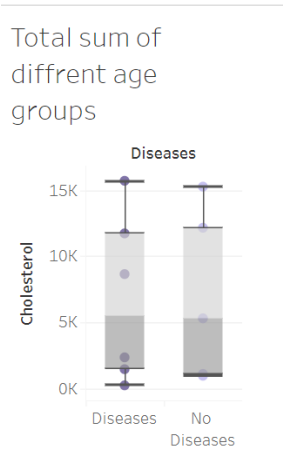

Fig.10 box-plot variation of disease /no disease vs Cholesterol

This illustrate the cholesterol range with respect to age with people having disease and not having disease

Fig.10 side -by-side circle of variation of disease for each target class values

shows the heart disease parameter for respective sex class with the maximum heart rate. It is observed from the color code that the target class with ‘CHOLESTEROL’ highest is represented by blue color is male with 49k and the blue color indicates the “CHOLESTEROL” highest in female with 25k diabetics. Target class with diabetics and acceptable heart rate are showing.

.

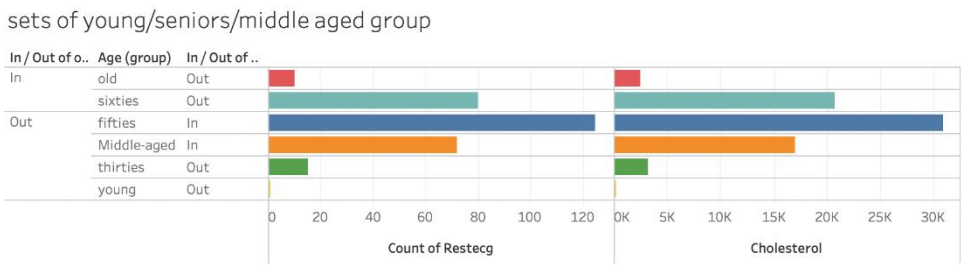

Fig.11 Horizontal Histogram of variation of age for each diseases.

It depicts the distribution of set of different ages and the risk of heart disease for the Trestbps class, and target classes ranging from 29 to 79 are having high risk of heart disease

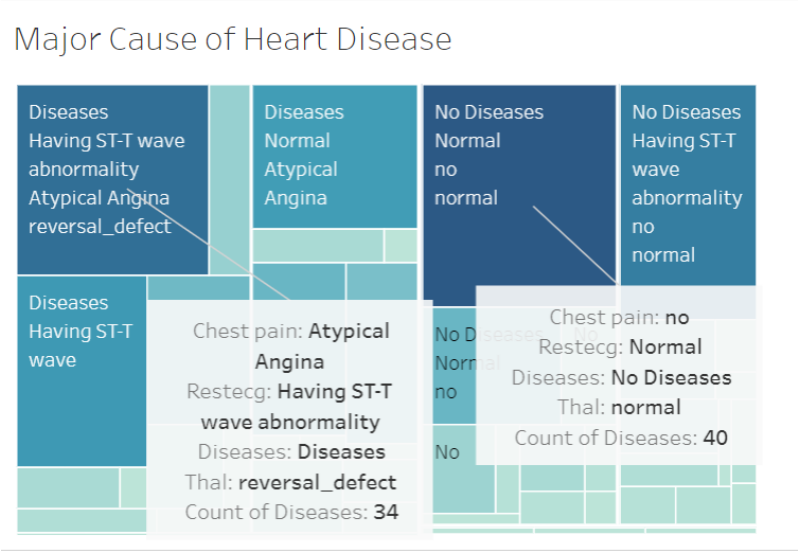

Fig 12. Major Cause of Heart Disease

This chart tells us the major causes of heart disease. So as a summary if a person has ST-T wave abnormality and chest pain as atypical angina and thalassemia is under reversal defect than there is a high risk of heart diseases and if a person do not have any chest pain, Resting ECG is normal then there is no risk of diseases.