INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

A STATISTICAL STUDY ON HEART DISEASE PREDICTION USING RELEVANT FEATURES

A GROUP PROJECT BY

ACTION SPACE

Submitted By:

Rayyan Ahmed Khan (231080068)

Arun Kumar Singh (231080022)

Omkar Yadav (231080062)

Sunita Kulariya (221106)

Madhav Madhukar (231080054)

Submitted To:

Prof. Subhajit Dutta

Content

A Statistical Data Analysis on Heart Disease Prediction

Abstract

World Health Organization has estimated 12 million deaths occur worldwide, every year due to heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and turn reduce the complications. This project intends to pinpoint the most relevant risk factors of heart disease as well as predict the overall risk using logistic regression.

1 INTRODUCTION

1.1 WHAT IS HEART DISEASE:

A type of disease that affects the heart or blood vessels.The term "Heart disease" refers several types of heart conditions. The most common heart disease is coronary artery disease (narrow or blocked coronary arteries), which can lead to chest pain, heart attacks, or stroke. Other heart diseases include congestive heart failure, heart rhythm problems, congenital heart disease (heart disease at birth), and endocarditis (inflamed inner layer of the heart). Also called cardiovascular disease. Heart disease is the leading cause of death in the United States, according to the Centers for Disease Control and Prevention (CDC). In the United States, 1 in every 4 deaths in is the result of a heart disease.

1.2 HOW THE HEART WORKS: To understand the causes of heart disease, it helps to understand how the heart works. heart is a pump. It's a muscular organ about the size of our fist, located slightly left of center in our chest. Heart is divided into the right and the left sides.

- The right side of the heart includes the right atrium and ventricle. It collects and pumps blood to the lungs through the pulmonary arteries.
- The lungs give the blood a new supply of oxygen. The lungs also breathe out carbon dioxide, a waste product.

- Oxygen-rich blood then enters the left side of the heart, including the left atrium and ventricle.
- The left side of the heart pumps blood through the largest artery in the body (aorta) to supply tissues throughout the body with oxygen and nutrients.

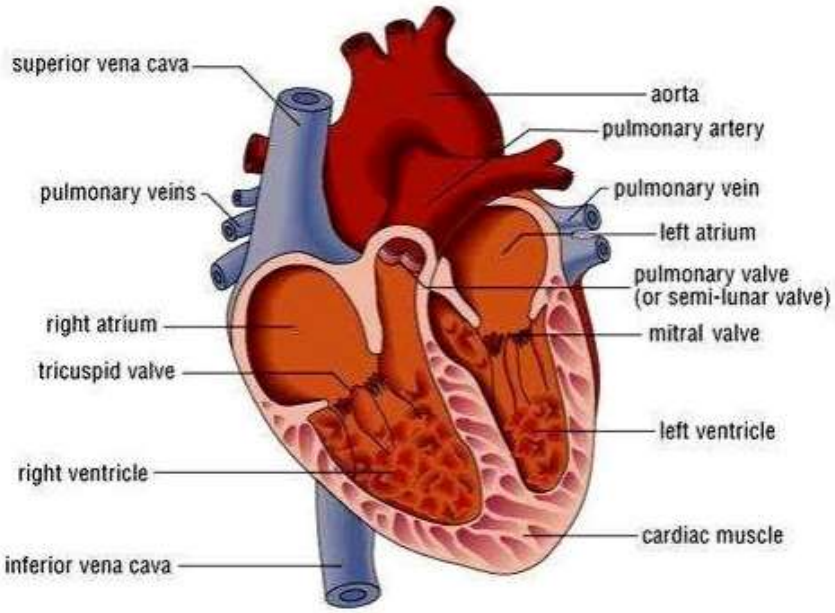
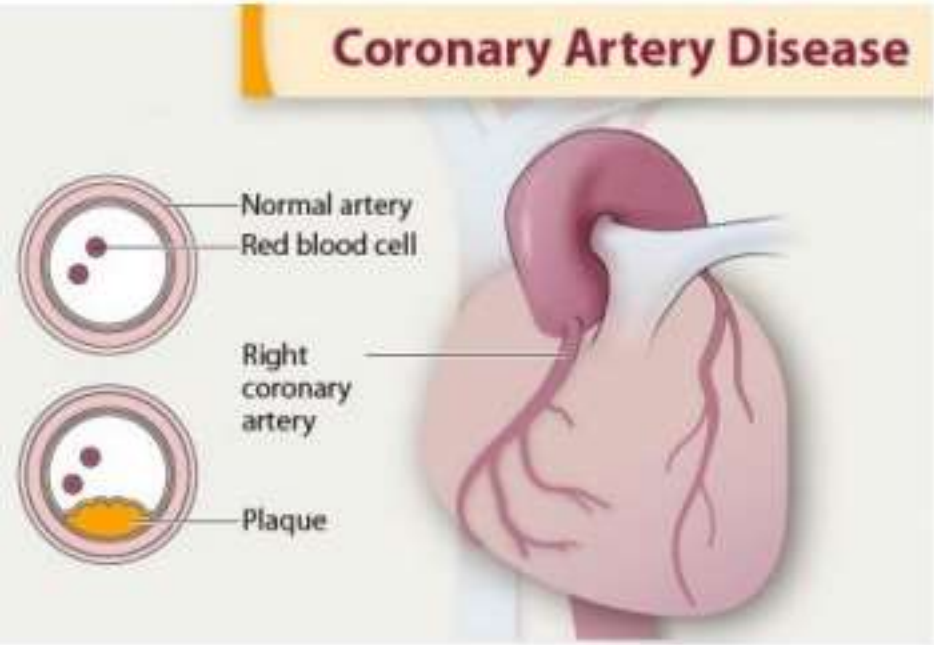


FIGURE 1: Cross section of heart

1.3 DIFFERENT TYPES OF HEART DISEASE AND THEIR CAUSES:

There are many types of heart disease. Some of them include:

(1) Coronary Artery Disease OR CAD



Causes: A buildup of fatty plaques in arteries (atherosclerosis) is the most common cause of coronary artery disease.Plaque buildup causes the inside of the arteries to narrow over time, which can partially or totally block the blood flow. This process is called atherosclerosis.

(2) Congenital Heart Disease: Congenital heart defects usually develop while a baby is in the womb. Heart defects can develop as the heart develops, about a month after conception, changing the flow of blood in the heart.

Causes: Some medical conditions, medications and genes may play a role in causing heart defects, Drinking alcohol, having diabetes or having habit of smoking during pregnancy can defect baby's heart. Heart defects can also develop in adults. As human ages, heart's structure can change, causing a heart defect.

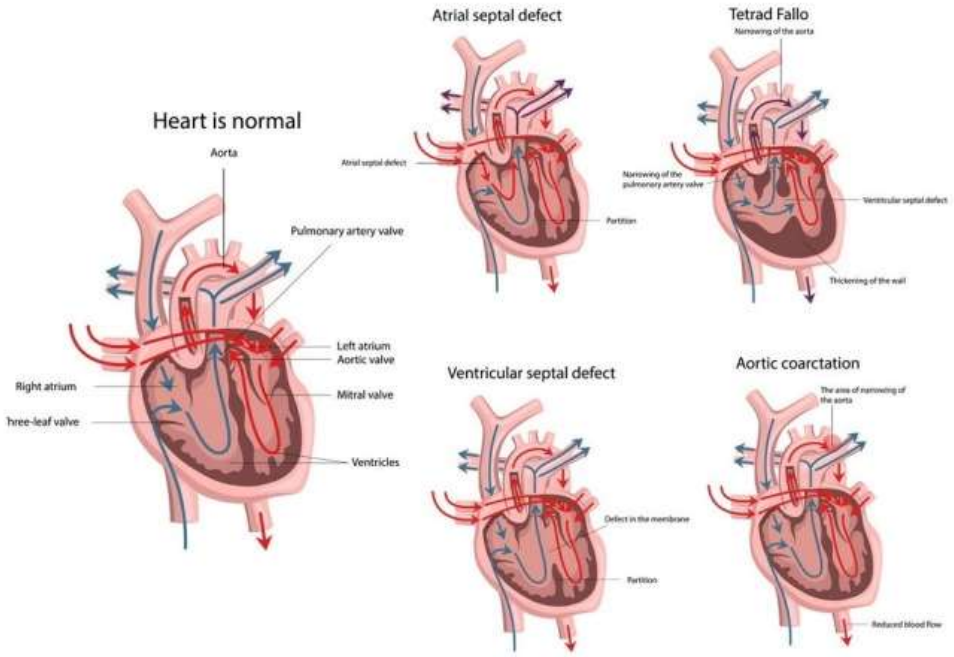
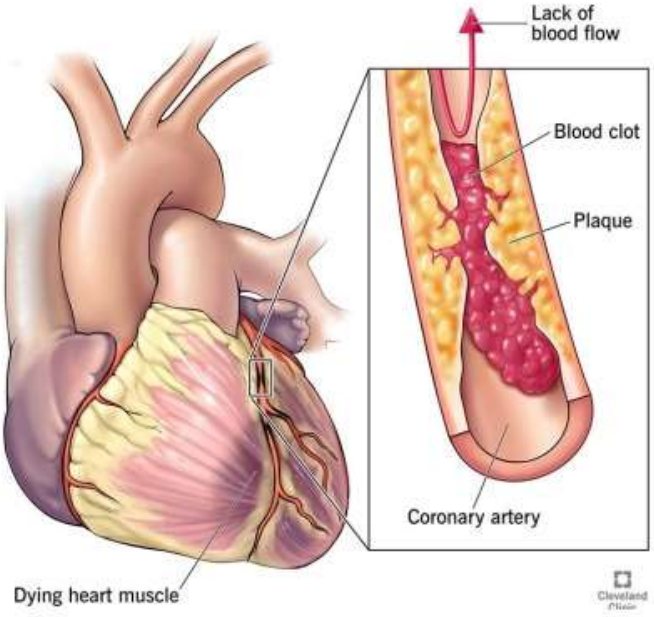


FIGURE 2: Congenital heart disease

(3) Heart Attack or Myocardial Infraction

Heart attack happens when the arteries leading to the heart become blocked, disrupting blood flow.



Causes: The leading cause of heart attacks is coronary heart disease. This is usually due to cholesterol containing deposits called plaques. Plaques can narrow the arteries, reducing

blood flow to the heart. If a plaque breaks open, it can cause a blood clot in the heart. A heart attack may because by a complete or partial blockage of a heart (coronary) artery.

(4) Heart Muscle Disease or Cardiomyopathy:

This condition can lead to heart failure. It occurs when the heart muscle becomes larger and stiffens, preventing it from pumping blood away from the heart. Some times blood can pool in the lungs.

Causes:

(i)Long-term high blood pressure

(ii)Heart tissue damage from a heart attack

(iii)Long-term rapid heart rate, can lead to Cardiomyopathy

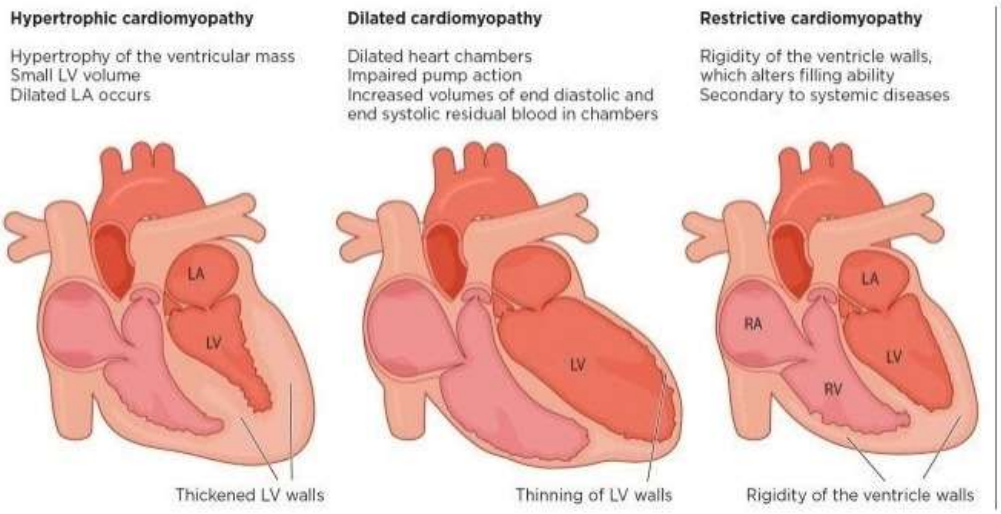
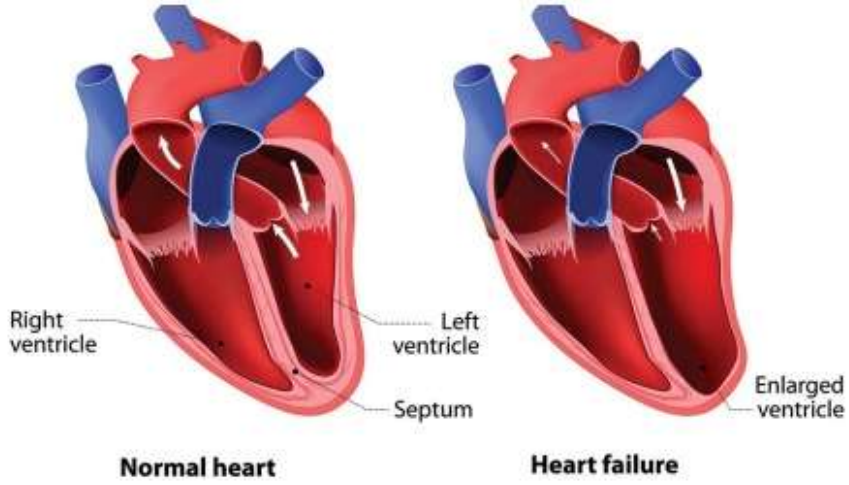


FIGURE 3: Cardiomyopathy

(5) Heart Failure or Congestive Heart Failure

This condition occurs when stiffness in the heart prevents the organ from pumping blood adequately through the body.



Causes: Conditions including high blood pressure,valve disease, thyroid disease, kidney disease, diabetes, or heart defects present at birth can all cause heart failure. In addition, heart failure can happen when several diseases or conditions are present at once.

(6) Heart Valve Disease

Valve disease happens when any of the four valves in the heart don't open or close properly and interrupt blood flow. If the defect in the valve happens at birth, it's called congenital heart disease.

Causes: of valve disease: History of certain infections that can affect the heart History of certain forms of heart disease or heart attack High blood pressure, high cholesterol, diabetes and other heart disease risk factors Heart valve disease can cause many complications, including:,Heart failure, Stroke, Blood clots.

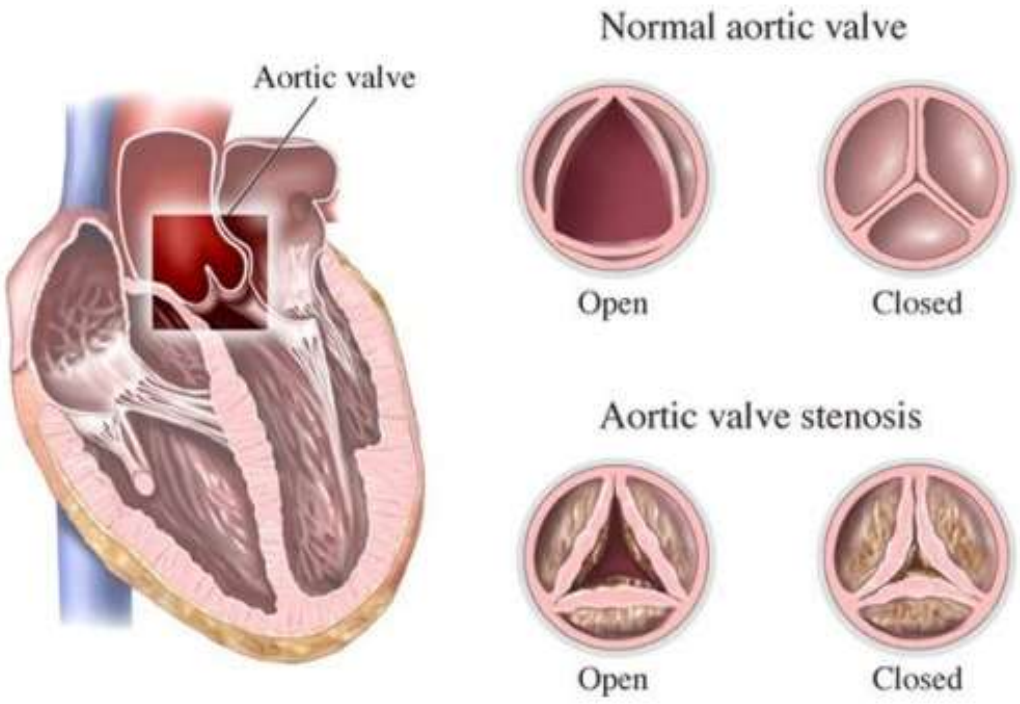


FIGURE 4: Heart Valve disease

7. Abnormal Heart Rhythms OR Arrhythmia:

This condition causes a fluctuation in the heartbeat that happens while at rest.

Causes: :Excessive use of alcohol or caffeine, diabetes, stress, valvular heart disease can causes arrhythmia.

1.4 RISK FACTORS:

Risk factors for developing heart disease include:

- **Age:** Growing older increases the risk of damaged and narrowed arteries and aweakened or thickened heart muscle
- **Sex:** Men are generally at greater risk of heart disease. The risk for women increases after menopause.
- **Family history:** A family history of heart disease increases the risk of coronary artery disease, especially if a parent developed it at an early age (before age 55 for a male relative, such as ones brother or father, and 65 for a female relative, such as ones mother or sister).
- **Smoking:** Nicotine tightens blood vessels, and carbon monoxide can damage their inner lining, making them more susceptible to atherosclerosis. Heart attacks are more common in smokers than in nonsmokers.

- **Poor diet:** A diet that's high in fat, salt, sugar and cholesterol can contribute to the development of heart disease.
- **High blood pressure:** Uncontrolled high blood pressure can result in hardening and thickening of arteries, narrowing the vessels through which blood flows.
- **High blood cholesterol levels:** High levels of cholesterol in blood can increase the risk of plaque formation and atherosclerosis.
- **Diabetes:** Diabetes increases the risk of heart disease. Both conditions share similar risk factors, such as obesity and high blood pressure.
- **Obesity:** Excess weight typically worsens other heart disease risk factors.
- **Physical inactivity:** Lack of exercise also is associated with many forms of heart disease and some of its other risk factors as well.
- **Stress:** Unrelieved stress may damage the arteries and worsen other risk factors for heart disease.

1.5 PROBLEM DEFINITION:

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience,time and expertise. Since we have a good amount of data in today's world, we can use various statistical analysis to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

2 OBJECTIVE OF THE PROJECT:

In this project work our objective is to model the chance of occurance of heart disease based on several associated covariates or features (age, sex, cholesterol, blood sugar blood pressure etc.).The goal of our heart disease prediction project is to determine if a patient should be diagnosed with heart disease or not, which is a binary outcome, so: Positive result = 1, the patient will be diagnosed with heart disease. Negative result = 0, the patient will not be diagnosed with heart disease. We have to find which model has the greatest accuracy and identify correlations in our data. Finally, we also have to determine which features are the most influential in our heart disease diagnosis. This will act as a tool to the physicians for predicting the probability of heart disease of a patient given the values of the relevant covariates of him/her.

3 DATA DESCRIPTION:

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5 CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

3.1 SOURCE OF THE DATA:

- **Dataset:** (https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction)

This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features . The five datasets used for its curation are:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Total: 1190 observations

Duplicated: 272 observations

Final dataset: 918 observation

3.2 THE DATA:

This dataset contains 11 features age, sex chest pain type, resting blood pressure, fasting blood sugar, resting ecg, ST-slope, Exercise angina, Maximum heart rate achieved, cholesterol, oldpeak and the respons variable is Heart disease. From the original dataset we have converted the categorical covariates (Sex, Chestpain type, RestingECG, ExcerciseAgina, ST-slope) and the dependent variable or response heart disease into numerical form. Here some of the observations are shown in the following table:

```
##    Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR
## 1   40   M           ATA       140         289         0     Normal   172
## 2   49   F           NAP       160         180         0     Normal   156
## 3   37   M           ATA       130         283         0         ST    98
## 4   48   F           ASY       138         214         0     Normal   108
```

```
## 5   54   M           NAP       150         195         0     Normal   122
## 6   39   M           NAP       120         339         0     Normal   170
## 7   45   F           ATA       130         237         0     Normal   170
## 8   54   M           ATA       110         208         0     Normal   142
## 9   37   M           ASY       140         207         0     Normal   130
## 10  48   F           ATA       120         284         0     Normal   120
##    ExerciseAngina Oldpeak ST_Slope HeartDisease
## 1               N     0.0       Up            0
## 2               N     1.0     Flat            1
## 3               N     0.0       Up            0
## 4               Y     1.5     Flat            1
## 5               N     0.0       Up            0
## 6               N     0.0       Up            0
## 7               N     0.0       Up            0
## 8               N     0.0       Up            0
## 9               Y     1.5     Flat            1
## 10              N     0.0       Up            0
```

4 GRAPHICAL REPRESENTATION OF THE DATA:

(i)Presence and absence of heart disease:

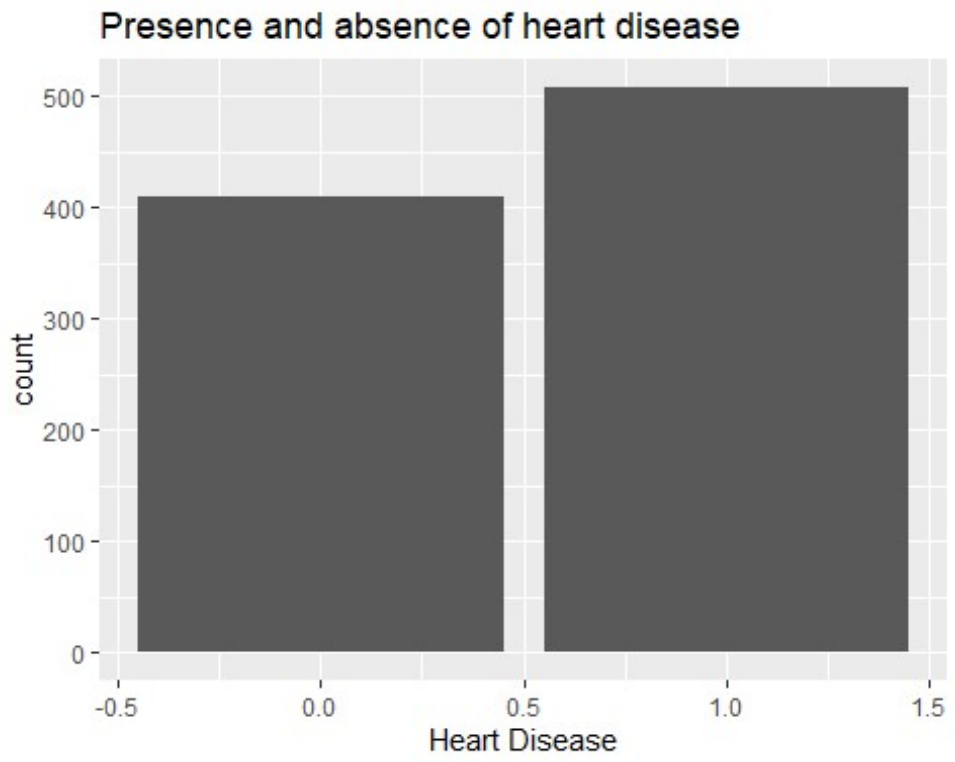


FIGURE 5: Presence and absence of Heart Disease.

(ii) Age analysis:

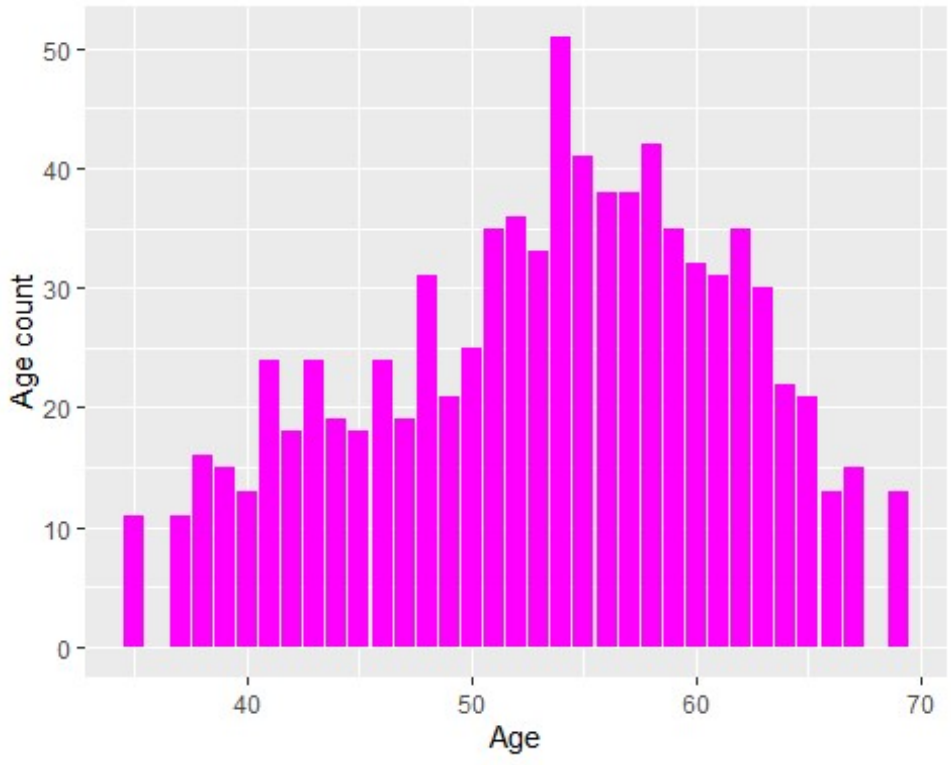


FIGURE 6: Age count

- From the above graph we can say that most of patients are between age 50 to 65 in the data

(iii) BP analysis:

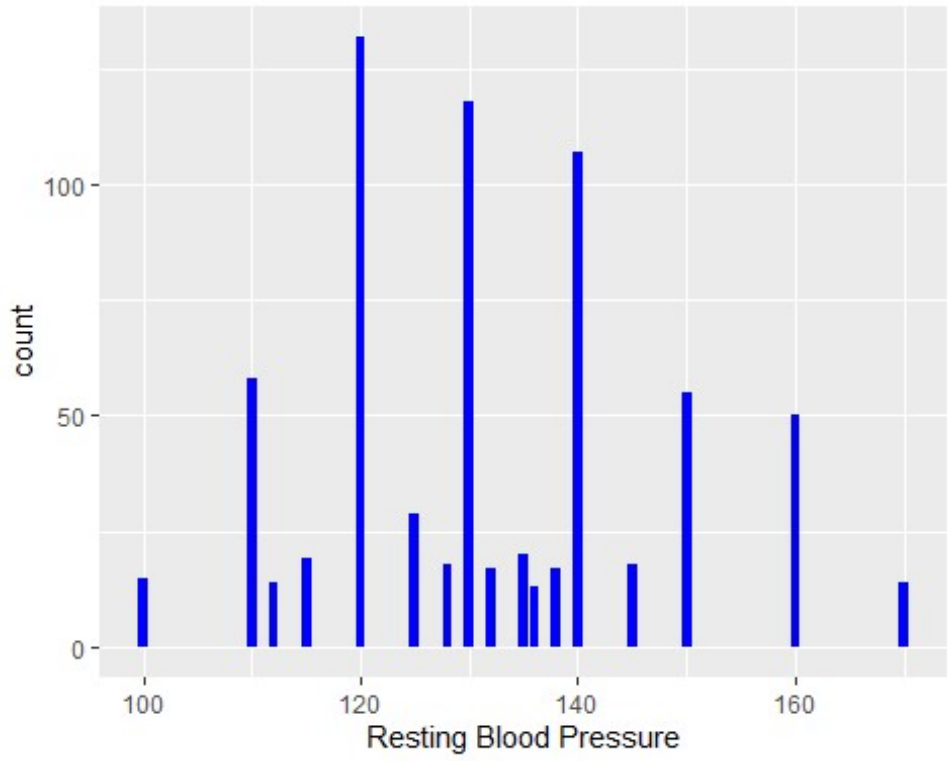


FIGURE 7: BP count

(iv) Compare Blood pressure across chest pain:

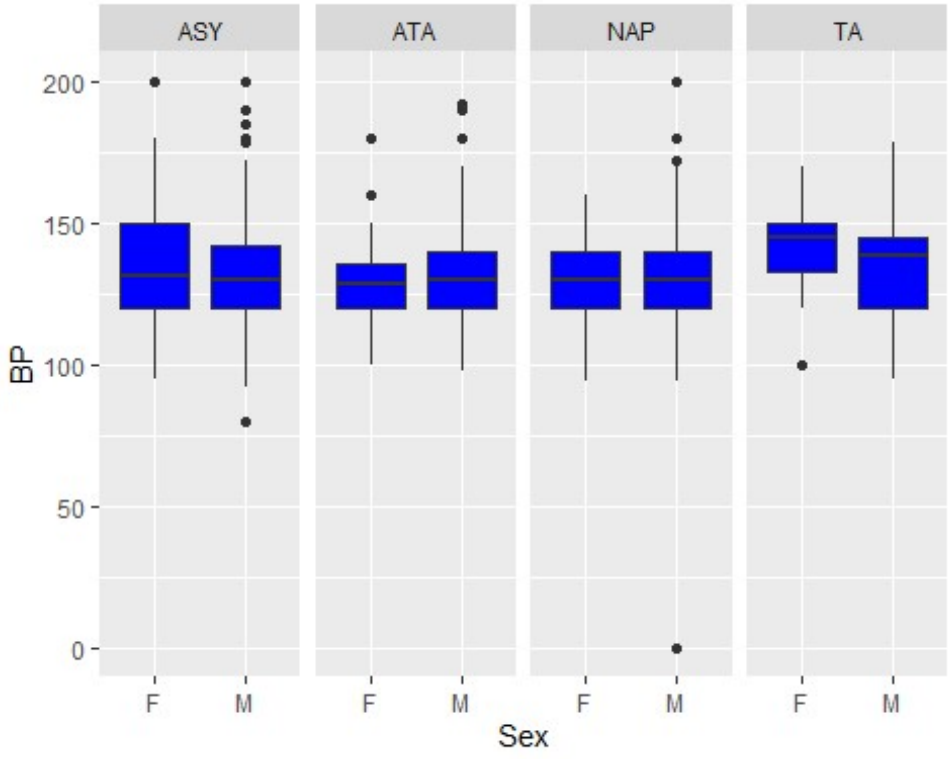


FIGURE 8: Compare Blood pressure across chest pain

- Here we are comparing blood pressure across chest pain for males and females. From the above graph we can see that there are some outliers.

(v) Compare Cholesterol across chest pain:

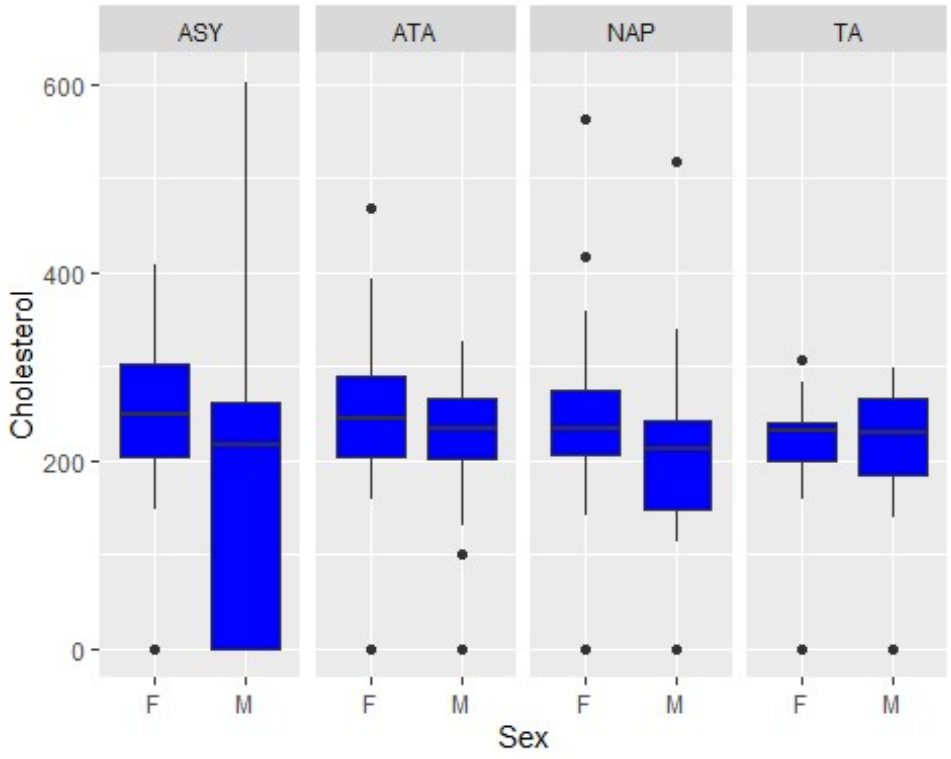


FIGURE 9: Compare Cholesterol across chest pain

- Here we are comparing Cholesterol across chest pain for males and females. From the above graph we can see that there are some outliers.

(vi) Correlation between the attributes:

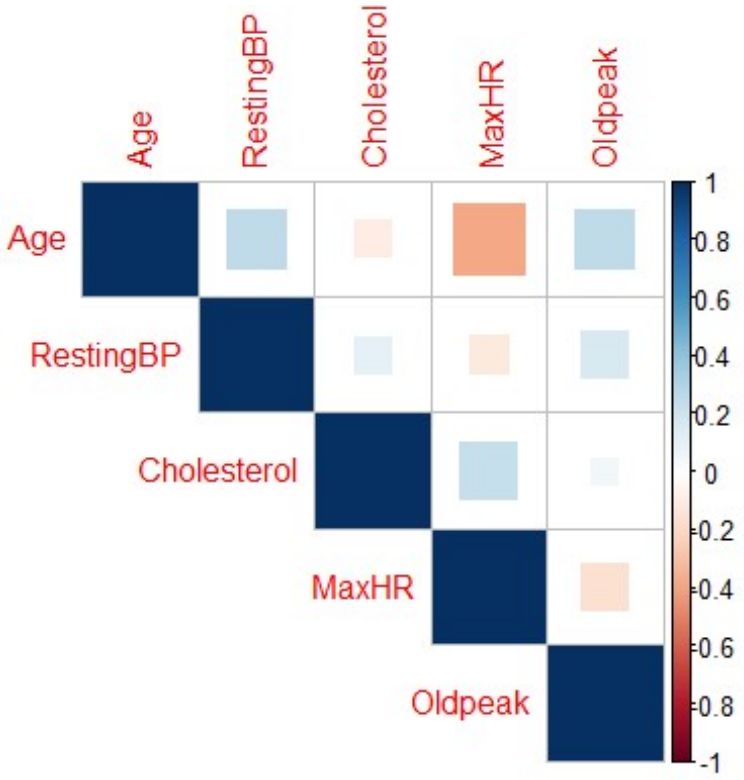


FIGURE 10: Correlation plot

- From the above plot we can see that correlation between Age and cholesterol;BP and cholesterol; BP and maximum Heart rate; cholesterol and oldpeak is very low.

5 MATERIALS AND METHODS FOR ANALYZING THE DATA:

In the data target variable is a binary dependent variable (yes or no). For categorical response data, logistic regression is the most important model. We will use

Logistic Regression to model the probability that the chance of occurance of heart disease based on several associated covariates i.e, 'age', 'sex', 'chest pain type', 'resting bp', 'cholesterol', 'FastingBS', 'RestingECG', 'MaxHR', 'ExcerciseAngina', 'Oldpeak', 'ST-slope'. To measure the association between the target variable and covariates we will use

Relative risk and odds ratio. Here we will discuss about (i) logistic regression and (ii) Relative risk and odds ratio in brief.

Before starting the discussion on logistic regression method we need to know about

5.1 Generalized Linear Model:

- **Why Generalized Linear Model (GLM) ? :**

  Consider a situation when the response variable takes only two values, i.e 0 and 1 (binary) with probabilities π & $1-\pi$.

  Here, $E(Y) = 0 \times (1-\pi) + 1 \times \pi$

  Again from the model $E(Y) = X\underline{\beta}$, then $X\underline{\beta} = \pi$

  Note that π is a probability and hence $0 \leq \pi \leq 1$ but there is no guarantee that $X\underline{\beta}$ lies between 0 and 1. For sufficiently large or small values of the explanatory variable, $X\underline{\beta}$ maybe outside the limits. Thus there is a contradiction so that the classical linear model fails.

- **Extension to the GLM in case of binary response:**

  Replace $X\underline{\beta}$ in the model by a quantity that necessarily lies between 0 and 1. A distribution function is a very common choice. Thus in case of a GLM,

  $$E(\underline{Y}) = F\left(X\underline{\beta}\right)$$

  where F is a distribution function.

  Here, $\underline{\eta} = X\underline{\beta}$ is the linear predictor of covariates. For alinear model, $\underline{\eta}) = \underline{\mu}$ but for GLM, $\underline{\mu} = F\left(\underline{\eta}\right)$ or $\underline{\eta} = F^{-1}\left(\underline{\mu}\right)$ We can arrange the usual LM into the following three components

  (i) **Random component:** The components of $\underline{Y}$ have independent normal distribution with common mean μ and variance σ^2

(ii) **Systematic Component:** The covariates $X_1, X_2, \ldots, X_p$ produce a linear predictor $\eta = \sum_{j=1}^{p} X_j \beta_j$

(iii) **The link function :** $\underline{\eta} = \underline{\mu}$

- **Some link function:** When Y is a binary response variable,

(i) Logit link $\eta = ln\left(\frac{\mu}{1-\mu}\right)$

(ii) Probit link $\eta = \phi^{-1}(\mu)$

(iii) Complementary log-log link $\eta = ln[-ln(1-\mu)]$

5.2 LOGISTIC REGRESSION:

- **Binary response and logistic regression:**

Let Y be a binary response variable taking only two values 0 and 1. $\therefore E(Y) = P(Y=1) = \pi(\underline{x})$ (say) We use the notation $\pi(\underline{x})$ to denote that π is dependent on the explanatory variables $X1, X2, \ldots, Xp$, For simplicity we start with the only one covariate X. For a binary response variable Y, $E(Y) = \alpha + \beta x$, is called the linear probability model. When the observation on Y are independent, it is a GLM with identity link function.

Structural defect of linear probability model:

For significantly large x values, we may have $\pi(x) > 1$ or $\pi(x) < 0$. We usually expect a non linear relationship between X and $\pi(x)$. For fixed change in X, we have less impact on $\pi(x)$ when π is near 0 or 1 then that when π is moderate (i.e between 0 and 1). This model can be valid for a finite value of x. Also conditions that make the least square estimate optimal, are not satisfied. Here $Var(Y) = \pi(x)(1-\pi(x)) \to 0$ as $\pi(x) \to 0$ *or* 1.

The conditional distribution of y is most nearly concentrated at a point.

Logistic Regression model:

Because of structural problem with the linear probability model, it is more fruitful to study models implying a curvilinear relationship when we expect a monotonic relationship. An S-shaped curve is the natural shape for the regression curve. A function having the shape is

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

which is termed as logistic regression function.

Properties

$$\pi'(x) = \frac{d}{dx}\left(1 - \frac{1}{1+e^{\alpha+\beta x}}\right)$$

$$= \frac{\beta e^{\alpha+\beta}}{(1+e^{\alpha+\beta x})^2}$$

$$= \beta \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \frac{1}{1+e^{\alpha+\beta x}}$$

$$= \beta\pi(x)(1-\pi(x))$$

Since $0 < \pi(x) < 1, \pi(x)\big(1-\pi(x)\big) > 0$

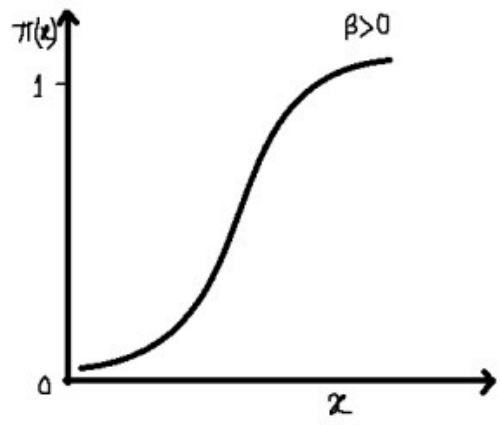
So, if $\beta > 0, \pi'(x) > 0$ if $\beta <, \pi'(x) > 0$

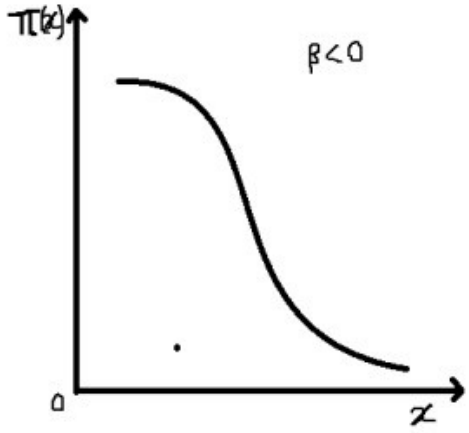
Also if $\beta > 0, \pi(x) \to 1\, as\, x \to \infty$

and $\beta < 0, \pi(x) \to 0\, as\, x \to \infty$

As $\beta \to 0, \pi(x) \to \frac{e^{\alpha}}{1+e^{\alpha}}$

The curve approaches a horizontal straight line, $\beta = 0$, implies the absence of the affect of covariates in independents of X and Y





Maximum slope of the curve:

We have $\pi'(x) = \beta\pi(x)\big(1-\pi(x)\big)$ Max $\pi'(x) = \frac{\beta}{4}$ attained when $\pi(x) = \frac{1}{2}$

$\pi(x) = \frac{1}{2}$

$\Rightarrow \frac{2e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = 1$

$\Rightarrow 2e^{\alpha+\beta x} = 1 + e^{\alpha+\beta x}$

$\Rightarrow e^{\alpha+\beta x} = 1$

$\Rightarrow \alpha + \beta x = 0$

$\Rightarrow x = -\frac{\alpha}{\beta}$

As $|\beta|$ increases, the curve has a steeper rate of increase at x values.

An approximation to β

Let, x_1 and x_2 are such that $\pi(x_1) = \frac{1}{4}, \pi(x_2) = \frac{3}{4}$

$\pi(x_1) = \frac{1}{4}$

$\Rightarrow 3e^{\alpha+\beta x_1} = 1$

$\Rightarrow \alpha + \beta x_1 = -\ln 3$

Again,

$\pi(x_2) = \frac{3}{4}$

$\Rightarrow 4e^{\alpha+\beta x_2} = 3 + 3e^{\alpha+\beta x_2}$

$\Rightarrow \alpha + \beta x_2 = \ln 3$

Therefore, $\beta(x_1 - x_2) = 2\ln 3 \Rightarrow \beta = \frac{2\ln 3}{x_2 - x_1}$

Interpretation of the parameters:

Here $\frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta x}$ LHS is the odds of making response 1, the odd changes multiplicatively at the rate β per unit change in x.

$ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x$ Thus β can be interpreted as the change in log odds per unit change in x. α can be interpreted as the baseline odds when there is no covariate.

- The appropriate link function here is the logit link $ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$ The advantages of this model over models using other links is that the effect can be estimated when the sampling design is prospective or retrospective.
- To determine the appropriate form of the systematic component of the logistic regression, plot the empirical logits against the $x_i's$.

- Suppose for the i-th setting of x i.e x_i, let we observe n_i responses and y_i be the number of yes responses. The i-th sample logit is

$$ln\left(\frac{\frac{y_i}{n_i}}{1-\frac{y_i}{n_i}}\right) = ln\left(\frac{y_i}{1-y_i}\right)$$

This is undefined for $y_i = 0$ or $y_i = n_i$. So we consider $ Code$ called empirical logit.

Inverse CDF link

$E(\alpha + \beta x) \rightarrow$ class of link functions for binary response. $\pi(x) = F(\alpha + \beta x) = F(\eta)$ Therefore $\eta = \alpha + \beta x = F^{-1}(\pi(x))$ $F^{-1}(.)$ maps (0,1) onto $(-\infty, \infty)$. It is a GLM with link $F^{-1}(.)$. For a logistic distribution with location parameter μ and scale parameter τ. $Codes$

When $F(x) = \phi(x)$=c.d.f. of N(0,1) then the model is called probit model. - Normal tails are than logistic, so far the probit model, $\pi(x) \rightarrow 0$ or $\rightarrow 1$ more quickly than the logit model.

5.3 ODDS RATIO AND RELATIVE RISK:

Relativerisk:

The difference of proportions of successes, $\pi_1 - \pi_2$, is a basic comparison of the two rows. A value $\pi_1 - \pi_2$ of fixed size may have greater importance when both π_i are close to 0 or 1 than when they are not. For a study comparing two treatments on the proportion of subjects who die, the difference between 0.010 and 0.001 may be more noteworthy than the difference between 0.410 and 0.401, even though both are 0.009. In such cases, the ratio of proportions is also informative. The relative risk is defined to be the ratio $\frac{\pi_1}{\pi_2}$ It can be any non-negative real number. A relative risk of 1.0 corresponds to independence.

Odds Ratio

For a probability π of success, the odds are defined to be $\Omega = \frac{\pi}{1-\pi}$

The odds are nonnegative, with Ω>1.0 when a success is more likely than a failure. When $\pi = 0.75$,for instance, then $\Omega = \frac{0.75}{0.25} = 3.0$; a success is three times as likely as a failure, and we expect about three successes for every one failure. When $\omega = \frac{1}{3}$, a failure is three times as likely as a success. Inversely,

$$\pi = \frac{\Omega}{1-\Omega}$$

Refer again to a 2×2 table. Within row i, the odds of success instead of failure are $\Omega_i = \frac{\pi_i}{1-\pi_i}$. The ratio of the odds Ω1 and Ω2 in the two rows,

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}$$

is called odds ratio.

6 ANALYSIS OF DATA:

In the data there are 11 covariates or predictor variables to predict the probability of heart disease. Among them 5 are continuous variables (age, resting blood pressure, cholesterol, maximum heart rate, oldpeak). Let us denote age, resting blood pressure, cholesterol, maximum heart rate, oldpeak, sex, chest pain type, fasting blood sugar, resting ECG, Excercise angina, ST-slope with X_1, X_2, X_3, X_4, $X_5, \ldots, X_{11}$ respectively.

$\underline{X} = X_1, X_2, X_3, X_4, X_5, \ldots, X_{11}$. Our target variable or response variable is HeartDisease. Let us denote it with Y, takes binary responses 1 and 0 with probability, $\pi(x_i) = \pi_i$ and $1 - \pi i$.

Define, $X_i = (x_{i1}, x_{i2}, .. x_{ip})$ where i = 1(1)918 and p = 1(1)11 Y_i = Number of **yes** response for the i-th setting of X; i = 1(1)918 n_i = Number of response for the i-th setting of X As we have 918 observations. Therefore the logistic regression equation is

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{p} \beta_j \, x_{ij}$$

independent with x_{io}, i = 1(1)918 and p = 1(1)11 Estimate of β:

$$\underline{\beta}^{(t+1)} = \left[X'\left(Diag\left(n_i\pi_i(1-\pi_i)\right)\right)^{-1} X\right]$$

Thus the mathematics works behind logistic regression.

6.2 DATA PREPARATION:

Now to analyse this large data first need to import the data in R studio. Then we need to do some data preparation to apply logistic regression method. Several predictor variable that we will use, some of it does not have a correct type of data. Therefore, we adjust variable type of data. Then we check the availability of any missing value in each variables.

6.3 TWO BY TWO TABLE ANALYSIS:

After preparing data we can do some case control by making sure all of the factor levels are represented by patients with and without heart disease and the results are given below:

TABLE 1

SEX

HeartDisease	FEMALE	MALE
NO	143	267
YES	50	458

TABLE 2

ChestPainType

HeartDisease	ASY	ATA	NAP	TA
0	104	149	131	26
1	392	24	72	20

TABLE 3

FastingBS

HeartDisease	<=120	>120
NO	366	44
YES	338	170

TABLE 4

RestingECG

HeartDisease	LVH	Normal	ST
0	82	267	61
1	106	285	117

TABLE 5

ChestPainType

HeartDisease	ASY	ATA	NAP	TA
0	104	149	131	26
1	392	24	72	20

TABLE 6

ST_Slope

HeartDisease	Down	Flat	Up
0	14	79	317
1	49	381	78

TABLE 7

ExerciseAngina

HeartDisease	N	Y
0	355	55
1	192	316

- **Analysis of 2×2 table (TABLE:2):**

Looking at the raw data from Table 2 we can say that most of the females don't have heart disease and most of the males have heart disease. Being female is likely to decrease the odds in having heart disease. Odds in case of female patients = 0.3496 In other words, if a sample is female, the odds are against it that result will be Yes. Being male is likely to increase the odds in having heart disease. Odds in case of male patients = 1.7153 In other words, if a sample is male, the odds are for it that the result will be yes. Here is the result of 2×2 table analysis for comparison test in R:

```
## 2 by 2 table analysis:
## ------------------------------------------------------
## Outcome   : FEMALE
## Comparing : NO vs. YES
##
##     FEMALE MALE    P(FEMALE) 95% conf. interval
## NO     143  267       0.3488    0.3042   0.3962
## YES     50  458       0.0984    0.0754   0.1275
##
##                                    95% conf. interval
##              Relative Risk: 3.5436    2.6395   4.7574
##          Sample Odds Ratio: 4.9059    3.4378   7.0011
## Conditional MLE Odds Ratio: 4.8971    3.3972   7.1467
##     Probability difference: 0.2504    0.1972   0.3030
##
##              Exact P-value: 0.0000
##         Asymptotic P-value: 0.0000
## ------------------------------------------------------
```

6.4 ANALYSIS USING LOGISTIC REGRESSION:

Now we will jump to the logistic regression. Let us create a very simple model that uses sex to predict Heart disease. Let us denote sex with X. So here the model will be $ln\left(\frac{\pi}{1-\pi}\right) = \beta_o + \beta_1 X$

```
##              Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -1.050822  0.1642953 -6.395931 1.595726e-10
## SexMALE      1.590442  0.1814433  8.765503 1.859444e-18
```

Interpretation:

Going through the first coefficient, the intercept is the log(odds) a female has heart disease. This is because female is the first factor in "sex" (the factors are ordered, alphabetically by default,"female", "male"). Now the second coefficient, sexMALE is the log(odds ratio) that tells that if a sample has sex=MALE, the odds of having heart disease are, on a log scale, 1.59 times greater than if a sample has sex=FEMALE. Now let us see what this logistic regression predicts, given that a patient is either female or male through a plot;

Since there are only two probabilities (one for females and one for males), we can use a table to summarize the predicted probabilities. The result is given below:

```
##                     Sex
## probability.of.hd   FEMALE MALE
##   0.259067357513526    193    0
##   0.63172413793103       0  725
```

TABLE 8: Predicted probabilities of getting heart disease.

Now we will use other available covariates for prediction. Before creating model, we check our target variable proportions, i.e

```
##
##        NO       YES
## 0.4466231 0.5533769
```

Selection of covariates: The common approach to statistical model building is minimization of variables until the most parsimonious model model that describes the data which also results in numerical stability and generalizability of the results. Inclusion of all other relevant variables in the model regardless of their significance can lead to numerical unstable estimates. Therefore We have to made purposeful selection of variables in logistic regression. **A decision to keep a variable in the model might be based on the statistical significance.** Finally, statistical significance should not be the sole criterion for inclusion of a term in a model. It is sensible to include a variable that is central to the purposes of the study and report its estimated effect even if it is not statistically significant.

Other criteria besides significance tests can help select a good model in terms of estimating quantities of interest. **The best known is the Akaike information criterion i.e AIC.** AIC=-2×(maximized log likelihood - number of parameters in model) .We will use several variables that may have a significant effect toward our target variable like, age, sex, blood pressure, cholesterol, maximum heart rate, oldpeak. So, We will start with age(X1), sex(X2) (Model1). Therefore as per previous denotation here the equation is,

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{p} \beta_j \, x_{ij}$$

independent with x_{io} = 1, i = 1(1)918 and p = 1, 2 Then we estimate the coefficients, the result given below

```
##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -4.63489288 0.481185190 -9.632243 5.843942e-22
## Age          0.06653121 0.008165056  8.148285 3.691216e-16
## SexM         1.64119501 0.189344877  8.667755 4.407236e-18
```

Interpretation: We can see that Age and sex are significant toward heart disease.So keeping them in the model we add blood pressure and cholesterol (Model2) and perform the previous process. The results are shown below

```
##                 Estimate   Std. Error   z value     Pr(>|z|)
## (Intercept) -4.536556683 0.6714886987 -6.755969 1.418840e-11
## RestingBP    0.008729276 0.0041910009  2.082862 3.726381e-02
## Cholesterol -0.003738244 0.0007424775 -5.034825 4.782861e-07
## Age          0.059802161 0.0084460957  7.080450 1.436868e-12
## SexM         1.477859457 0.1931725726  7.650462 2.002583e-14
```

From the above result we see that Blood pressure is significant at 5% level with heart disease. Now we will add maximum heart rate and old peak (Model3)

```
##                 Estimate   Std. Error    z value     Pr(>|z|)
## (Intercept)  1.172133089 0.9994802310  1.1727426 2.408990e-01
## RestingBP    0.004030559 0.0047297234  0.8521765 3.941161e-01
## Cholesterol -0.003847012 0.0008394794 -4.5826160 4.591949e-06
## Age          0.020919215 0.0099433645  2.1038367 3.539269e-02
## SexM         1.300474066 0.2138253510  6.0819452 1.187332e-09
## MaxHR       -0.026140546 0.0037676281 -6.9381970 3.971357e-12
## Oldpeak      0.946047331 0.0976132382  9.6917933 3.267403e-22
```

after adding two continuous covariates blood pressure becomes insignificant, which is unlikely. Therefore we need to drop them from the model as well as we will drop age (Model4) and then perform logistic regression again. And here is the results below:

```
##                 Estimate   Std. Error   z value     Pr(>|z|)
## (Intercept) -2.225978977 0.5651345868 -3.938848 8.187393e-05
## RestingBP    0.016196485 0.0040280049  4.020970 5.795908e-05
## Cholesterol -0.004181964 0.0007185532 -5.819978 5.885549e-09
## SexM         1.443330294 0.1869281226  7.721312 1.151386e-14
```

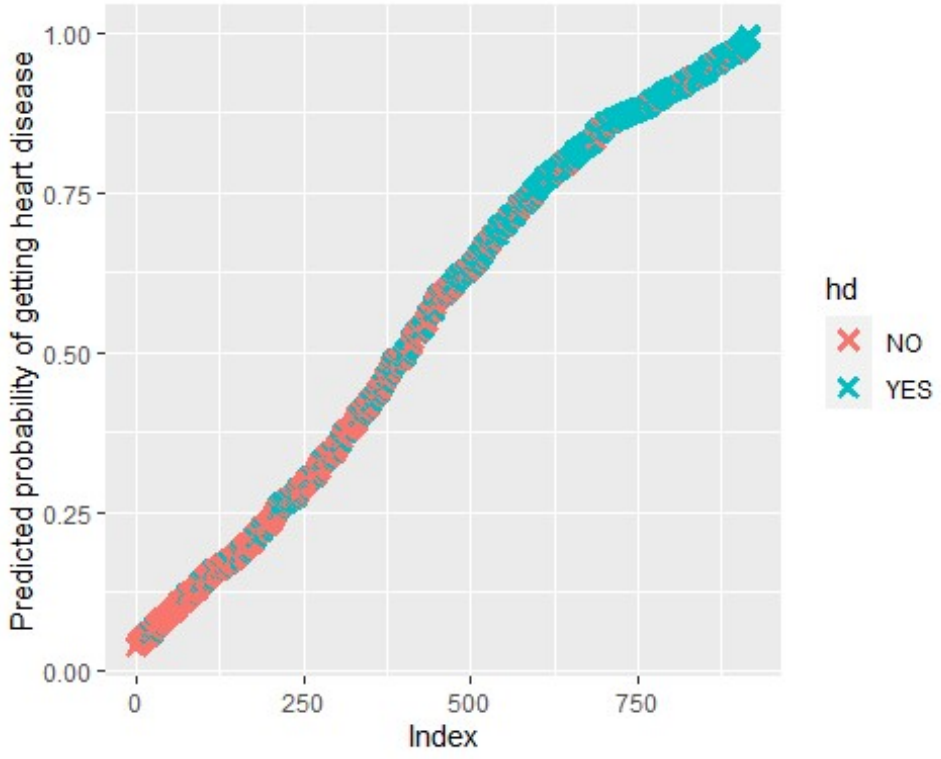
So we can see that Blood pressure is significant with heart disease when we use cholesterol and sex as other two covariate. Now we illustrate the AIC values along with residual variance for this four models.

TABLE 9

Model predictors	Residual deviance	df	AIC
Model 1	1101.6	915	1107.6
Model 2	1072.5	913	1082.5
Model 3	885.97	911	899.97
Model 4	1126.5	914	1134.5

For the 3rd model the AIC value is smallest. We will use this model for prediction as **lower AIC values indicate a better-fit model**

Prediction: Let us predict response variable using 3rd model. Then we predict the response variable heart disease using those 6 features or covariates and the plot is given below:



6.5 CONCLUSION:

The main problem in model building situations is to choose from a large set of covariates those that should be included in the 'best' model. From the simple one covariate model we got the probability of getting heart disease is more likely for male (63.17%) than female (25.91%). Here we have chosen 6 covariates. We have seen that while taking account

maximum heart rate and oldpeak, blood pressure becomes insignificant with heart disease which is not expected. In our data the correlation between heart rate and BP is negative i.e -0.112135. In real life the relationship between blood pressure and heart rate is location dependent. Therefore in spite of insignificance of Blood pressure, it can be used as a feature for prediction. So as per our analysis it is proposed to use the 6 features, Resting blood pressure, cholesterol, age, sex, maximum heart rate achieved and Oldpeak to predict heart disease

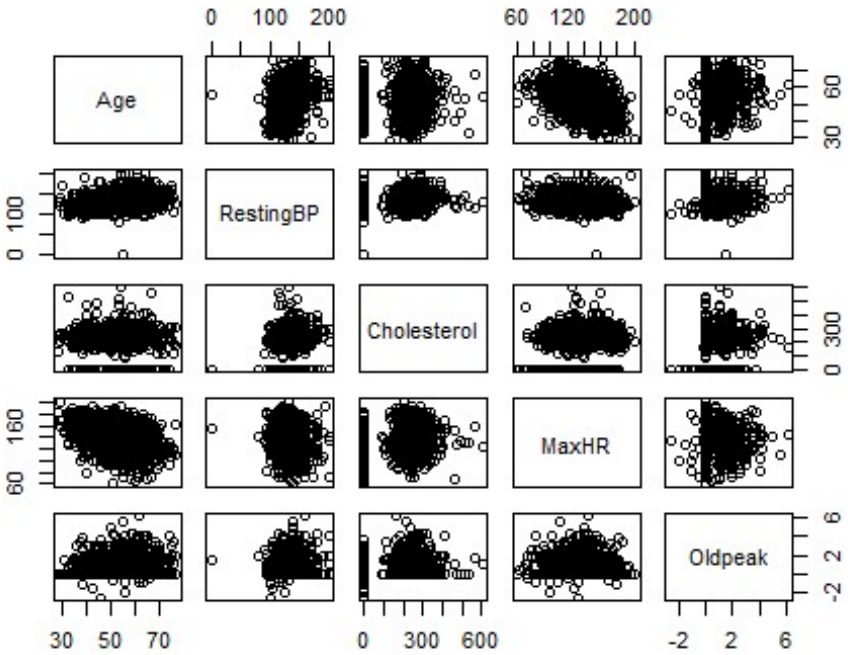


FIGURE 13: Corr elation between continuous variables

7 ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to our professor Shubhajit Dutta who gave us the golden opportunity to do this project. It helped us in doing lot of research and we came to know a lot of things to this topic.

8 REFERENCE

1. https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

2. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

3.wikipedia.

4. Categorical Data Analysis-Alan Agresti

5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2633005/

6. https://www.impcna.com/6-types-of-heart-disease-and-what-causes-them