# VIDYASAGAR UNIVERSITY

## HEALTH CARE : HEART ATTACK POSSIBILITY

**Registration No: - 1160480 of 2020-2021**

**Roll No: - 1126116-200137**

**Session: - 2022-2023**

# ACKNOWLEDGEMENT

It gives me tremendous pleasure in acknowledging that valuable assistance extended to me by various people in the successful completion of this program.

Primarily I would thank Haldia Govt. College, Vidyasagar University for providing me with an opportunity to do the project work. Then I wish to feel great pleasure to express my deep sense of gratitude to Dr Shyamsundar Sahoo, Mr Sibsankar Karan, Mr Tanmay Kumar Maity, and Mr Bijitsh Halder for their excellent guidance, instructions, constant encouragement and a keen interest in the progress of this project work.

Last but not least I would like to thank all of my friends who have helped and given me timely support due to which the project has been completed successfully.

CONTENTS ~

# 1. INTRODUCTION

In India, heart attacks have emerged as a critical contributor to mortality, accounting for a staggering 28% of all reported deaths. This alarming statistic highlights the urgent need to delve deeper into the underlying causes and contributing factors behind this widespread occurrence. Analyzing the data surrounding heart attack-related deaths is imperative to not only comprehend the extent of the issue but also to develop targeted strategies for prevention and management.

One key aspect of analyzing heart attack data involves understanding the demographic distribution of these cases. Factors such as age, gender, socioeconomic status, and geographic location play pivotal roles in influencing the prevalence of heart attacks. For instance, certain age groups might be more susceptible due to physiological changes, while socioeconomic factors can impact lifestyle choices that contribute to heart health. Moreover, the disparity between urban and rural areas in terms of healthcare access and lifestyle patterns could further elucidate patterns in heart attack occurrences.

An in-depth analysis of heart attack data in India considers factors like age, gender, chest pain, blood pressure, blood sugar, cholesterol levels, maximum heart rate achieved, and thalassemia presence. This comprehensive approach seeks to uncover potential correlations between these variables and the occurrence of heart attacks, aiding in targeted prevention and treatment strategies.

# 2. Objective:

i. Associating Explanatory Variables: The primary goal of our analysis is to establish associations between various explanatory variables and the likelihood of heart attacks. These variables include age, sex, chest pain characteristics, resting blood pressure, slope of a cardiac vessel, number of major vessels, thalassemia type, and maximum heart rate achieved. By examining these relationships, we aim to identify key factors that significantly influence the occurrence of heart attacks.

ii. Logistic Regression Fitting: We intend to employ a logistic regression model to effectively capture and quantify the relationships between the mentioned explanatory variables and the binary outcome variable, which likely represents the presence or absence of a heart attack. By fitting this model, we aim to mathematically express how changes in these variables are linked to changes in the probability of experiencing a heart attack.

iii. Predictive Modeling: Utilizing the logistic regression model, we seek to develop a predictive tool that can estimate the likelihood of a heart attack based on a given set of explanatory variables. This predictive aspect enables us to assess the risk associated with individual patients or populations, aiding in early detection and informed decision-making regarding medical interventions and preventive measures.

## 3. DATA DESCRIPTION

This data collected from the website - https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

Creators :

1. Hungarian Institute of Cardiology . Budapest : Andras Janosi , M.D.

2. University Hospital , Zurich , Switzerland : William Steinbrunn , M.D.

3. University Hospital , Basel , Switzerland : Matthias Pfisterer , M.D.

4. V.A. Medical Center , Long Beach and Cleveland Clinic Foundation : Robert Detrano , M.D. , Ph.D.

- **Age**:

  Here age is specified as the age of the patient.

- **Sex**:

  Here sex is specified as the sex of the patient.

- **Chest Pain**:

  Chest pain refers to discomfort or pain that occurs in the chest area, between the neck and the abdomen. It is a common symptom that can have various causes, ranging from mild and temporary issues to serious medical conditions.

- **Resting Blood Pressure**:

  Resting blood pressure refers to the measurement of blood pressure when a person is in a calm and relaxed state, typically after sitting or lying down for a few minutes. It is an essential parameter used to assess an individual's cardiovascular health and is often measured as part of a routine health examination or medical evaluation.

- **Serum Cholesterol:**

  Serum cholesterol refers to the measurement of cholesterol levels in the blood. Cholesterol is a waxy, fat-like substance that is produced by the liver and also obtained from certain foods. It is an essential component of cell membranes and is used in the body to produce hormones, vitamin D, and substances that aid in digestion.

- **Fasting Blood Sugar**:

  Fasting blood sugar refers to the measurement of glucose levels in the blood after a period of fasting, typically overnight or for at least 8 hours. It is an important diagnostic tool used to assess and monitor blood glucose control, particularly in the diagnosis and management of diabetes.

- **Resting Electrocardiographic Results:**

  A resting electrocardiographic (ECG) result refers to the recording of the electrical activity of the heart while the individual is at rest. It is a commonly performed non-invasive diagnostic test used to assess the electrical rhythm and conduction of the heart. A resting ECG provides valuable information about the heart's health, identifies irregularities, and helps diagnose various cardiac conditions.

- **Maximum Heart Rate Achieved:**

The maximum heart rate achieved refers to the highest heart rate a person reaches during physical activity or exercise. It is often used as a reference point to determine exercise intensity, monitor cardiovascular fitness, and guide exercise prescriptions.

- **Exercise Induced Angina:**

Exercise-induced angina, also known as exertional angina, is chest pain or discomfort that occurs during physical activity or exercise. It is typically a symptom of underlying coronary artery disease (CAD), which is the narrowing or blockage of the blood vessels that supply oxygen-rich blood to the heart.

- **Oldpeak:**

Oldpeak, also known as ST depression, refers to a specific finding on an electrocardiogram (ECG) that is associated with heart attacks, also known as myocardial infarctions. It represents a downward displacement of the ST segment of the ECG waveform.

- **The Slope of The Peak Exercise ST Segment:**

The slope of the peak exercise ST segment refers to the direction and steepness of the ST segment on an electrocardiogram (ECG) during peak exercise or physical stress testing. The ST segment is a portion of the ECG waveform that represents the period between ventricular depolarization (contraction) and repolarization (relaxation).

- **Number of Major Vessels:**

During certain diagnostic procedures or interventions, such as coronary angiography or percutaneous coronary intervention (PCI), fluoroscopy is used to visualize the coronary arteries. Fluoroscopy is a real-time X-ray imaging technique that allows healthcare professionals to observe the flow of contrast dye as it is injected into the blood vessels.

- **Thalassemia:**

Thalassemia is an inherited blood disorder characterized by abnormal haemoglobin production, which affects the production of red blood cells. It is not directly associated with an increased risk of heart attacks. However, thalassemia can indirectly contribute to an increased risk of heart complications in some individuals.

- **Target:**

Indicative of the presence or absence of heart attack.

# 4. METHODOLOGY

## a) Pearsonian Chi-Square Test:

Pearson's chi-square test is a statistical test that can be used to determine if there is a significant association between two categorical variables. It is commonly used in research to test the hypothesis that there is no significant difference between the expected frequencies and the observed frequencies of the variables being studied. To conduct a Pearson's chi-square test to check whether a factor is significant or not, follow these steps:

**1.** Define the null hypothesis: The null hypothesis states that there is no significant association between the two variables being studied.

**2.** Collect data: Collect data on the two variables being studied in the form of a contingency table. The contingency table shows the frequency of each category of the two variables.

**3**. Calculate the expected frequencies: Calculate the expected frequencies for each cell in the contingency table using the formula E = (row total x column total) / total.

**4.** Calculate the chi-square statistic: Calculate the chi-square statistic using the formula X^2 = Σ [(O - E)^2 / E], where O is the observed frequency and E is the expected frequency.

**5.** Determine the degrees of freedom: Calculate the degrees of freedom (df) as (number of rows - 1) x (number of columns - 1).

**6.** Determine the critical value: Determine the critical value of chi-square from a chi-square distribution table using the degrees of freedom and the desired level of significance (usually 0.05).

**7.** Compare the calculated chi-square statistic with the critical value: If the calculated chi-square statistic is greater than the critical value, reject the null hypothesis and conclude that there is a significant association between the two variables. If the calculated chi-square statistic is less than the critical value, fail to reject the null hypothesis and conclude that there is no significant association between the two variables.

In summary, Pearson's chi-square test can be used to determine if a factor is significant or not by analyzing the association between two categorical variables.

## Pearsonian chi-Square Test in R:

To perform a chi-square test in R, you can use the "chisq.test()"function. Here's an example: The output includes the chi-square statistic, degrees o freedom (df), and p-value. The p-value is greater than 0.05, indicating that there is not enough evidence to reject the null hypothesis that there is no significant association between the two variables. The null hypothesis in this case is that there is no association between the two categorical variables. If the p-value is less than the significance level (usually 0.05), we reject the null hypothesis and conclude that there is a significant association between the variables. If the p-value is greater than or equal to the significance level, we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest a significant association between the variables

## Logistics Regression :

Multiple logistic regression using the logit function theory is a statistical method used to model the relationship between a binary response variable and multiple predictor variables. The *logit* function is a mathematical function used to model the relationship between the probability of the response variable being 1 and the predictor variables.

The *logit* function is defined as the natural logarithm of the odds of the response variable being 1, given the values of the predictor variables. Mathematically, the *logit* function can be expressed as:

$logit(\pi) = ln(\pi / (1 - \pi)),$

where $\pi$ is the probability of the response variable being 1. In multiple logistic regression using the logit function theory, the log odds of the response variable being 1 is model as a linear combination of the predictor variables. Mathematically, the model can be expressed as:

$logit(\pi) = \alpha + \beta 1 * x1 + \beta 2 * x2 + ... + \beta k * xk,$

where $\alpha$ is the intercept term, $\beta 1, \beta 2, ..., \beta k$ are the coefficients of the predictor variables $x1, x2, ..., xk$, respectively.

The coefficients of the model are estimated using maximum likelihood estimation, which involves finding the set of coefficients that maximize the likelihood of the observed data given the model. The significance of each predictor variable is determined using hypothesis testing, which involves testing whether the coefficient of each predictor variable is significantly different from zero.

Multiple logistic regression using the *logit* function theory is a powerful tool for analyzing data with a binary response variable and multiple predictor variables. It allows researchers to model the relationship between the response variable and the predictor variables while controlling for other relevant factors.

ACI Interpretation: AIC values are not interpretable on their own in terms of "good" or "bad." They are meaningful only when comparing models. A lower AIC indicates that a model is likely to fit the data well while penalizing for complexity. However, the absolute value of AIC doesn't provide insight into how well the model fits the data; it's only useful for comparing models within the same dataset.

## 5. Result & Analysis

1. **Chi- Square Test :** Hear I want to analysis to any independency our response and exploratory variable.

**Chi-square test for Age and Target:** Age is in numerical variable so first we spite the age into 3 groups. Groups are below 45, 45-60, and 60 above.

```
chisq.test(heart_data$split_age,heart_data$target)

	Pearson's Chi-squared test

data:  heart_data$split_age and heart_data$target
X-squared = 13.104, df = 2, p-value = 0.001427
```

We observed that the chi-square test p-value is less than 0.05 so we rejected the null hypothesis of 5% level of significance. We conclude that they are dependent

**Chi-square test for Sex and Target:** Here all are categorical variables so we test the chi-square test.

```
> chisq.test(heart_data$sex,heart_data$target)

	Pearson's Chi-squared test with Yates'
	continuity correction

data:  heart_data$sex and heart_data$target
X-squared = 22.717, df = 1, p-value =
1.877e-06
```

We observed that the chi-square test p-value is less than 0.05 so we rejected the null hypothesis of 5% level of significance. We conclude that they are dependent.

**Chi-square test for Chest pain and Target:** Here all are categorical variables so we test the chi-square test.

```
> chisq.test(heart_data$cp,heart_data$target)

	Pearson's Chi-squared test

data:  heart_data$cp and heart_data$target
X-squared = 81.686, df = 3, p-value < 2.2e-16
```

We observed that the chi-square test p-value is less than 0.05 so we rejected the null hypothesis of 5% level of significance. We conclude that they are dependent.

**Chi-square test for Fasting blood sugar and Target:** Here all are categorical variables so we test the chi-square test.

```
> chisq.test(heart_data$fbs,heart_data$target)

	Pearson's Chi-squared test with Yates'
	continuity correction

data:  heart_data$fbs and heart_data$target
X-squared = 0.10627, df = 1, p-value = 0.7444
```

We observed that the chi-square test p-value is less than 0.05 so we accepted the null hypothesis of a 5% level of significance. We conclude that they are independent.

**Chi-square test for resting electrocardiographic results and Target:** Here all are categorical variables so we test the chi-square test.

```
> chisq.test(heart_data$restecg,heart_data$target)

	Pearson's Chi-squared test

data:  heart_data$restecg and heart_data$target
X-squared = 10.023, df = 2, p-value = 0.006661
```

We observed that the chi-square test p-value is less than 0.05 so we rejected the null hypothesis of 5% level of significance. We conclude that they are dependent.

**Chi-square test for exercise-induced angina and Target:** Here all are categorical variables so we test the chi-square test.

```
chisq.test(heart_data$exang,heart_data$target)

	Pearson's Chi-squared test with Yates'
	continuity correction

data:  heart_data$exang and heart_data$target
X-squared = 55.945, df = 1, p-value =
7.454e-14
```

We observed that the chi-square test p-value is less than 0.05 so we rejected the null hypothesis of 5% level of significance. We conclude that they are dependent.

**Chi-square test for Slope and Target:** Here all are categorical variables so we test the chi-square test.

```
chisq.test(heart_data$slope,heart_data$target)

        Pearson's Chi-squared test

data:  heart_data$slope and heart_data$target
X-squared = 47.507, df = 2, p-value =
4.831e-11
```

We observed that the chi-square test p-value is less than 0.05 so we rejected the null hypothesis of 5% level of significance. We conclude that they are dependent.

**Chi-square test for CA and Target:** Here all are categorical variables so we test the chi-square test.

```
chisq.test(heart_data$ca,heart_data$target)

        Pearson's Chi-squared test

data:  heart_data$ca and heart_data$target
X-squared = 74.367, df = 4, p-value =
2.712e-15
```

We observed that the chi-square test p-value is less than 0.05 so we rejected the null hypothesis of 5% level of significance. We conclude that they are dependent.

**Chi-square test for Thalassemia and Target:** Here all are categorical variables so we test the chi-square test.

```
chisq.test(heart_data$thal,heart_data$target)

        Pearson's Chi-squared test

data:  heart_data$thal and heart_data$target
X-squared = 85.304, df = 3, p-value < 2.2e-16
```

We observed that the chi-square test p-value is less than 0.05 so we rejected the null hypothesis of 5% level of significance. We conclude that they are dependent.

**Chi-square test for Serum cholesterol and Target:** Serum cholesterol is in numerical variable so first we spite the age into 3 groups. Groups are below 200mg/dl, 200-239mg/dl, and above 240mg/dl.

```
chisq.test(heart_data$split_chol,heart_data$target)

	Pearson's Chi-squared test

data:  heart_data$split_chol and heart_data$target
X-squared = 3.7636, df = 2, p-value = 0.1523
```

We observed that the chi-square test p-value is less than 0.05 so we accepted the null hypothesis of a 5% level of significance. We conclude that they are independent.

**Chi-square test for Resting blood pressure and Target:** Resting blood pressure is in numerical variable so first we spite the age into 4 groups. Groups are below 90, 90-120, 120-140, and above 140.

```

	Pearson's Chi-squared test

data:  heart_data$split_tresbps and heart_data$target
X-squared = 6.557, df = 2, p-value = 0.03768
```

We observed that the chi-square test p-value is less than 0.05 so we rejected the null hypothesis of 5% level of significance. We conclude that they are dependent.

**Chi-square test for Oldpeak and Target:** Oldpeak is in numerical variable so first we spite the age into 3 groups. Groups are 0-1.5, 1.5-3, and above 3.

```
chisq.test(heart_data$splite_oldpeak,heart_data$target)

	Pearson's Chi-squared test

data:  heart_data$splite_oldpeak and heart_data$target
X-squared = 40.015, df = 2, p-value =
2.045e-09
```

We observed that the chi-square test p-value is less than 0.05 so we rejected the null hypothesis of 5% level of significance. We conclude that they are dependent.

**Chi-square test for Maximum heart rate achieved and Target:** Maximum heart rate achieved is in numerical variable so first we spite the age into 3 groups. Groups are below 100, 100-150, and above 150.

```
chisq.test(heart_data$splite_thalach,heart_data$target)

        Pearson's Chi-squared test

data:  heart_data$splite_thalach and heart_data$target
X-squared = 51.84, df = 2, p-value = 5.535e-12
```

We observed that the chi-square test p-value is less than 0.05 so we rejected the null hypothesis of 5% level of significance. We conclude that they are dependent.

## Target variable vs explanatory variable:

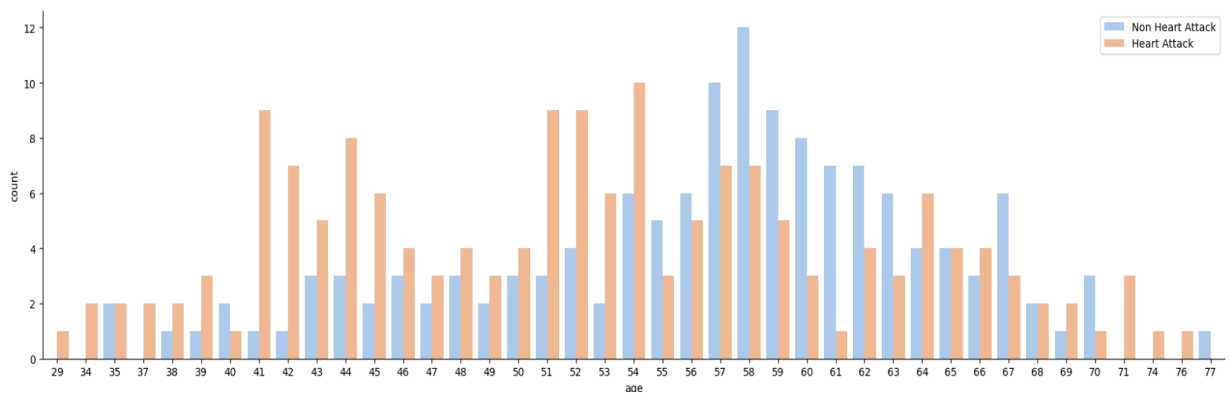
| FACTOR | CHI-SQUARE VALUE | DF | P-VALUE | COMMENT |
|---|---|---|---|---|
| Age | 13.104 | 2 | 0.001427 | H0 is rejected |
| Resting blood pressure | 6.557 | 2 | 0.03768 | H0 is accepted |
| Serum cholesterol | 3.7636 | 2 | 0.1523 | H0 is accepted |
| Maximum heart rate achieved | 51.84 | 2 | 5.54E-12 | H0 is rejected |
| Old-peak | 40.015 | 2 | 2.05E-09 | H0 is rejected |
| Sex | 22.717 | 1 | 1.88E-06 | H0 is rejected |
| Chest pain type | 81.686 | 3 | 2.20E-16 | H0 is rejected |
| Fasting blood sugar | 0.10627 | 1 | 0.7444 | H0 is accepted |
| Resting electrocardiographic results | 10.023 | 2 | 0.006661 | H0 is rejected |
| Exercise-induced angina | 55.945 | 1 | 7.45E-14 | H0 is rejected |
| The slope of the peak exercise ST segment | 47.507 | 2 | 4.83E-11 | H0 is rejected |
| Number of major vessels coloured by fluoroscopy | 74.367 | 4 | 2.71E-15 | H0 is rejected |
| Thalassemia | 85.304 | 3 | 2.20E-16 | H0 is rejected |

Table 1

In summary, while serum cholesterol, resting blood pressure, and fasting blood sugar were found to be independent of heart attacks, several other variables including age, sex, chest pain type, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, old-peak, the slope of the peak exercise ST segment, the number of major vessels coloured by fluoroscopy, and thalassemia have demonstrated a clear association with heart attacks.

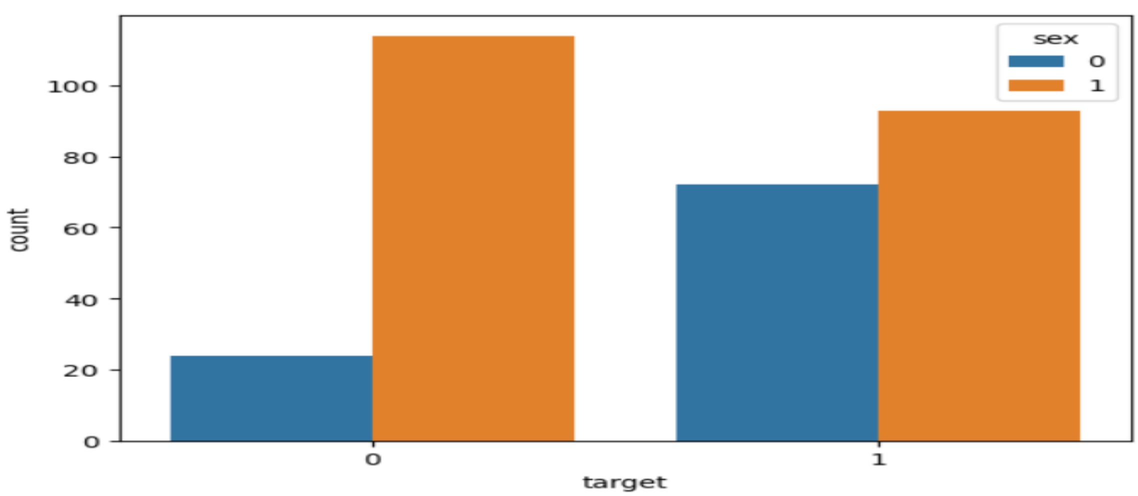
Understanding and addressing these factors can contribute to better risk assessment, prevention, and management strategies for heart disease.

## The graph of the target variable and the Age:




The graph reveals a distinct trend in heart attack risk across different age groups in India. Specifically, individuals aged 30 to 55 exhibit a higher likelihood of experiencing a heart attack, indicating a critical period of vulnerability. Surprisingly, the risk diminishes within the age range of 55 to 70. Notably, the data over time consistently underscores the elevated risk bracket of ages 40 to 55 as a particularly critical phase for heart attack occurrence, emphasizing the need for targeted preventive measures and awareness campaigns in this age window.

## Bar Diagram of target and sex:




Here “0” indicates the female and “1” indicates the male. So we said that female heart attack chance is high than male.

**Multiple logistics Regression :** We fit the logistic regression in R. In the summary that

```
glmfit=glm(target~age+sex+cp+restbps+slope+ca+thal+thalach,data=df,family=binomial)
> summary(glmfit)

Call:
glm(formula = target ~ age + sex + cp + restbps + slope + ca +
    thal + thalach, family = binomial, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4242  -0.4595   0.1616   0.5731   2.8254

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.011937   2.230491   0.454 0.650057
age         -0.008137   0.021313  -0.382 0.702602
sex         -1.695081   0.412415  -4.110 3.95e-05 ***
cp           0.948729   0.173740   5.461 4.74e-08 ***
restbps     -0.024352   0.009718  -2.506 0.012211 *
slope        1.101733   0.294341   3.743 0.000182 ***
ca          -0.777491   0.181758  -4.278 1.89e-05 ***
thal        -0.961397   0.268887  -3.575 0.000350 ***
thalach      0.028958   0.009628   3.008 0.002634 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 229.48  on 294  degrees of freedom
AIC: 247.48

Number of Fisher Scoring iterations: 5
```

- Intercept: The base log-odds of the response when all predictor variables are zero.
- Age: The change in log-odds of the response for a one-unit increase in age, which is not statistically significant ($p > 0.05$).
- Sex: Being female (sex = 0) is associated with a significant decrease in the log-odds of the response compared to being male (sex = 1).
- CP (Chest Pain): A one-unit increase in the chest pain variable is associated with a significant increase in the log-odds of the response.

- Restbps (Resting Blood Pressure): A one-unit increase in resting blood pressure is associated with a statistically significant decrease in the log-odds of the response.
- Slope: A one-unit increase in the slope variable is associated with a significant increase in the log-odds of the response.
- CA (Number of Major Vessels): A one-unit increase in the number of major vessels is associated with a significant decrease in the log-odds of the response.
- Thal: An increase in the thal variable is associated with a significant decrease in the log-odds of the response.
- Thalach (Maximum Heart Rate Achieved): A one-unit increase in maximum heart rate achieved is associated with a significant increase in the log-odds of the response.

AIC (Akaike Information Criterion): AIC is a measure of the model's goodness of fit, considering its complexity. Lower AIC values indicate a better balance between model fit and simplicity.

In summary, this logistic regression model quantifies the relationships between predictor variables (age, sex, chest pain, resting blood pressure, slope, ca, thal, thalach) and the likelihood of experiencing a heart attack. The coefficients and their associated p-values provide insights into the strength and significance of these relationships.

## 6. Conclusion:

The analysis of the Pearsonian chi-square test has provided important insights into the relationship between various factors and heart attacks. The results indicate that there is no significant association between serum cholesterol, resting blood pressure, and fasting blood sugar with heart attacks, suggesting that these variables are independent of the occurrence of heart attacks.

However, it is crucial to note that the analysis also revealed a significant relationship between several other variables and heart attacks. Factors such as age, sex, chest pain type, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, old-peak, the slope of the peak exercise ST segment, the number of major vessels coloured by fluoroscopy, and thalassemia have shown a clear association with heart attacks.

These findings highlight the importance of considering a comprehensive range of factors when examining the risk factors for heart attacks. Age, sex, and various clinical indicators such as chest pain type, electrocardiographic results, and maximum heart rate achieved are known to be crucial in understanding heart disease. The presence of exercise-induced angina, old-peak values, and the slope of the peak exercise ST segment are additional indicators of potential heart problems. Furthermore, the number of major vessels coloured by fluoroscopy and the presence of thalassemia can also serve as significant indicators of heart disease risk.

In summary, while serum cholesterol, resting blood pressure, and fasting blood sugar were found to be independent of heart attacks, several other variables including age, sex, chest pain type, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, old-peak, the slope of the peak exercise ST segment, the number of major vessels coloured by fluoroscopy, and thalassemia have demonstrated a clear association with heart attacks. Understanding and addressing these factors can contribute to better risk assessment, prevention, and management strategies for heart disease.

Age showed no statistically significant impact on the heart attack likelihood ($p = 0.702$), while being male ($p < 0.001$), experiencing specific chest pain (cp) characteristics ($p < 0.001$), having higher rest blood pressure ($p = 0.012$), steeper slope ($p < 0.001$), fewer major vessels ($p < 0.001$), specific thalassemia type ($p = 0.00035$), and a higher maximum heart rate achieved ($p = 0.0026$) were all associated with increased or decreased probabilities of a heart attack.

The model's goodness of fit is suggested by the Null deviance and Residual deviance values, with the Residual deviance being substantially lower. The Akaike Information Criterion (AIC) value of 247.48 reflects the model's overall performance. This analysis underscores the importance of these variables in understanding and predicting heart attack occurrences, providing valuable insights for healthcare professionals and policymakers aiming to mitigate and manage heart disease risks effectively.

## 7. Reference:


Data Website :

1. *https://archive.ics.uci.edu*
2. *https://www.kaggle.com/code*
3. *https://main.mohfw.gov.in*

BOOK :

1. *Fundamental of Statistics,Vol-1, Vol-2,A.M Gun , M.K Gupta , B. Dasgupta.*
2. *Fundamental of Mathematical Statistics, Gupta & Kapoor.*
3. *Probability Distribution Theory and Statistical inference, K C Bhuyan.*

## 8. Appendix:

```
library(stats)

setwd(dir = 'C:/Users//hp//Desktop//Project//sayak')
heart_data <- read.csv("heart12.csv")
#
head(heart_data)
str(heart_data)
 x<-cor(heart_data)
 heatmap(x)
chisq.test(heart_data$age,heart_data$target)

chisq.test(heart_data$sex,heart_data$target)

chisq.test(heart_data$cp,heart_data$target)

chisq.test(heart_data$fbs,heart_data$target)

chisq.test(heart_data$restecg,heart_data$target)

chisq.test(heart_data$exang,heart_data$target)

chisq.test(heart_data$slope,heart_data$target)

chisq.test(heart_data$ca,heart_data$target)

chisq.test(heart_data$thal,heart_data$target)

chisq.test(heart_data$split_age,heart_data$target)

chisq.test(heart_data$split_chol,heart_data$target)

chisq.test(heart_data$split_tresbps,heart_data$target)
```

```r
chisq.test(heart_data$splite_oldpeak,heart_data$target)

chisq.test(heart_data$splite_thalach,heart_data$target)


str(df)

df <- heart_data
glmfit=glm(target~age+sex+cp+restbps+slope+ca+thal+thalach,data=df,family=binomial)
summary(glmfit)

plot(glmfit)
f_model <- lm(Target ~ ., data = d)
nu_model <- lm(Target~1,data=d)
library(MASS, lib.loc = "C:/Program Files/R/R-4.3.0/library")
summary(stepAIC(nu_model,direction = "forward",scope
=list(upper=f_model,lower = nu_model)))

# Load the required library
library(plotly)

# Create the histogram with box plot overlays
fig <- plot_ly(heart_data, x = ~age, color = ~factor(target),
             colors = list("0" = "lightgreen", "1" = "lightgreen"),
             type = "histogram") %>%
  add_trace(histnorm = "percent") %>%
  add_trace(y = ~factor(target), type = "box", boxpoints = "all", jitter = 0.3) %>%
  layout(showlegend = FALSE)

# Show the figure
fig
```