Fall 2022

**Team members: (Group 16)**

Draksharam, Vishnu Paschyanti (vxd200023)

Jawahar Vasagam, Premi (pxj220007)

Farzana M B, Ashika (axf210029)

Tallapally, Jahnavi (jxt200051)

Parthasarathi, Prriyamvradha (pxp220005)

**Project: BUAN.6356.006 Business Analytics with R**

**Heart Attack Prediction**

Table of Contents

[Objective: 2](#_Toc121185956)

[Summary: 2](#_Toc121185957)

[Introduction: 2](#_Toc121185958)

[Research: 4](#_Toc121185959)

[Data Description: 3](#_Toc121185960)

[Dataset: 4](#_Toc121185961)

[Data Pre-processing: 4](#_Toc121185962)

[Exploratory Data Analysis: 4](#_Toc121185963)

[Summary: 4](#_Toc121185964)

[Finding Outliers using Box Plot: 5](#_Toc121185965)

[ggplot – Histogram 7](#_Toc121185966)

[Scatterplot 8](#_Toc121185967)

[Data Modelling 8](#_Toc121185968)

[Splitting the dataset: 8](#_Toc121185969)

[Logistic Regression: 8](#_Toc121185970)

[Decision Tree 11](#_Toc121185971)

[Performance Evaluation: 13](#_Toc121185972)

[Confusion Matrix: 13](#_Toc121185973)

[ROC Curve: 15](#_Toc121185974)

[Process Flow Diagram 16](#_Toc121185975)

[Conclusion 17](#_Toc121185976)

# Objective:

Heart disease has increased significantly over the past few years for various reasons, including the environment, food, and people's different lifestyle choices. Our primary goal in this project is to identify the risk factors that significantly increase a person's likelihood of experiencing a heart attack. With this data, we can better understand how to use the items to improve our health, which also enables us to identify all probable causes of heart problems.

# Summary:

To create our models, we utilized R, a static computation and graphical application. To comprehend our dataset and build a broad concept for subsequent research, we first studied the data. The variables we looked at, showed some correlations. Then, we built classification models using this data set, such as the decision tree model and logistic regression model, to determine whether someone with specific diagnostic parameters has a high risk of developing heart disease. We used 20% of the records as validation datasets and 80% of the records as training datasets for the decision tree model. We then plotted the decision tree and assessed it using the confusion matrix and curve. We also sample the training and validation datasets for the logistic regression model, compute the odds ratio, and assess the logistic regression model using the confusion matrix in ROC. The best model was chosen after thorough evaluation of each one.

# Introduction:

The elements that are creating heart problems, how people are altering their habitats through time, and how to spot them before it's too late are what motivate us to work on this. By examining the diseases that are the main causes of death worldwide, we might calculate the death rates using the information supplied. By 2030, cardiovascular disease is expected to impact 44% of US adults, meaning that 92.1 million Americans should have had at least one kind of cardiovascular illness.

The greatest cause of death and a significant contributor to disability in the US is heart disease. By preventing and reducing risk factors, the likelihood of getting cardiovascular disease can be decreased. Identify the medical problems or way of living that may contribute to the development of the illness and evaluate your present cardiovascular status. Take steps to lower your degree of risk, such as letting you know about the variety of resources Catholic Health System has to offer. Smoking, obesity, a lack of physical exercise, and other factors that stress the heart and lungs have all slowly raised the incidence of heart disease in emerging adults. One of the body's main organs, the heart, is necessary for life; without it, the body is meaningless. These are the reasons that are medically Shown.

In our study we will be considering few business factors that could help in the health sector:

1. Which causes heart attacks the most?
2. Compared to younger individuals, are older people more susceptible to heart attacks?
3. Does having high cholesterol increase your risk of heart attacks?

# Data Description:

The heart disease dataset, which was collected from Kaggle, serves as the main original dataset for this investigation. The primary dataset, which includes one target variable and 13 predictor variables, was deemed sufficient, therefore no new areas were employed. Each independent variables and the type of variables, that is either numerical or categorical and the description of the variables has been described in the table below.

|  |  |  |
| --- | --- | --- |
| **Independent Variable** | **Type** | **Description** |
| age | Numerical | Indicates the age of the patient in years. |
| sex | Categorical | Indicates the sex of the patient in a binary format ~  1= male 0= female |
| cp | Categorical | Chest pain type ~ 0 = Typical Angina 1 = Atypical Angina 2 = Non-anginal Pain 3 = Asymptomatic |
| trtbps | Numerical | Resting blood pressure (in mm Hg on admission to the hospital) |
| chol | Numerical | Cholesterol in mg/dl fetched via BMI sensor |
| fbs | Categorical | (Fasting blood sugar > 120 mg/dl) ~ 1 = True 0 = False |
| restecg | Categorical | Resting electrocardiographic results ~ 0 = Normal 1 = ST-T wave normality 2 = Left ventricular hypertrophy |
| thalachh | Numerical | Maximum heart rate achieved in the scale of (71 to 202) |
| oldpeak | Numerical | ST depression induced by exercise relative to rest. |
| slp | Categorical | The slope of the peak exercise ST segment  0 = downsloping. 1=flat.  2=upsloping |
| caa | Categorical | Number of major blood vessels (0-4) |
| thall | Categorical | Thallium Stress Test result ~  1= fixed defect 2 = reversible defect 3=normal |
| exng | Categorical | Exercise induced angina ~ 1 = Yes 0 = No |

|  |  |  |
| --- | --- | --- |
| **Dependent Variable** | **Type** | **Description** |
| output | Categorical | 0=less chance of heart attack 1=more chance of heart attack |
|  |

# Dataset:

<https://www.kaggle.com/code/namanmanchanda/heart-attack-eda-prediction-90-accuracy/data>

# Data Pre-processing:

We loaded the selected dataset from Kaggle into R

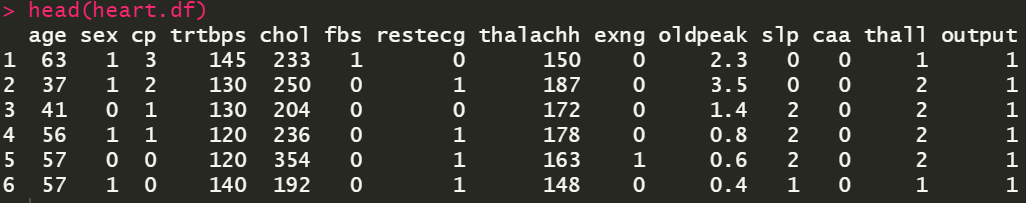
We found no null values in the dataset, so we did not have to remove any null values.



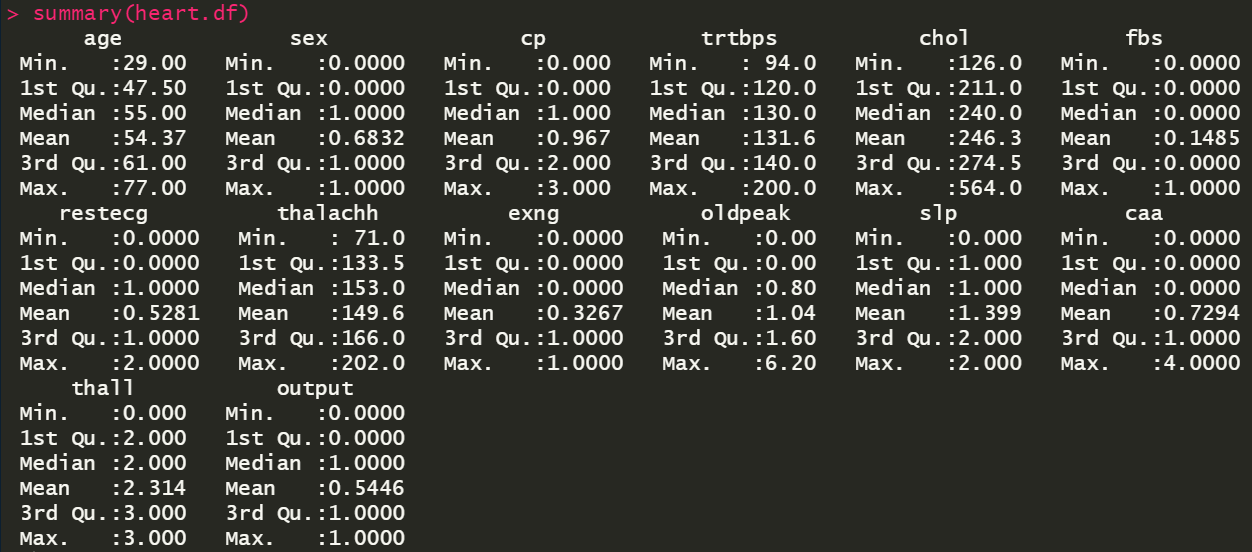
## Exploratory Data Analysis:

### Summary:

The following image shows the sample data of the heart attack prediction dataset



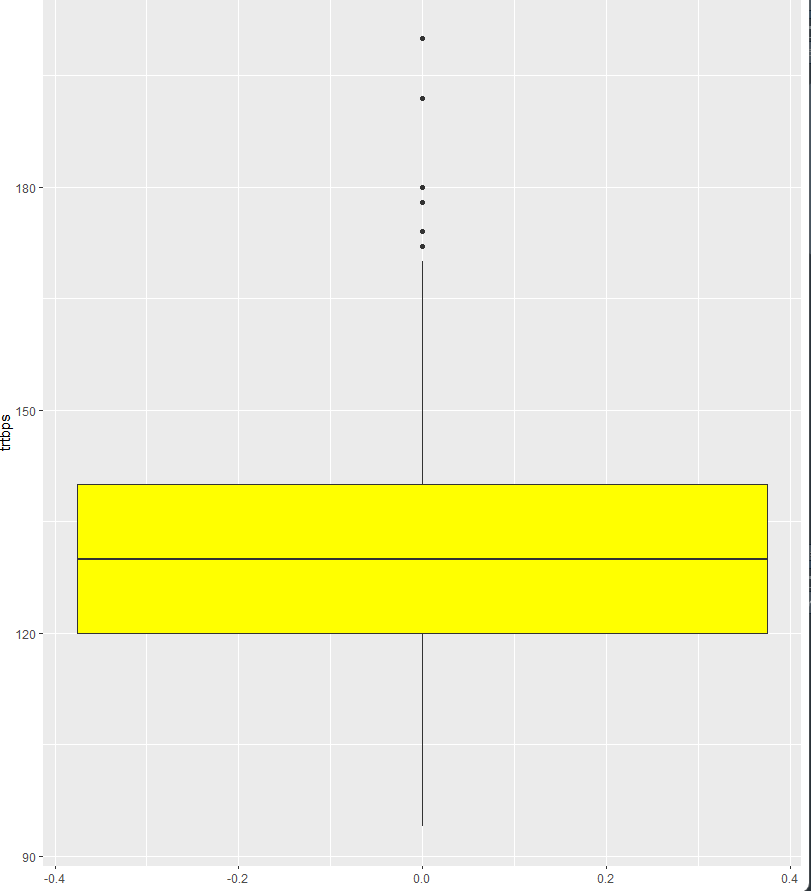
The following image shows the summary statistics for the different variables that we have in our dataset.



### Finding Outliers using Box Plot:

**Trtbps**

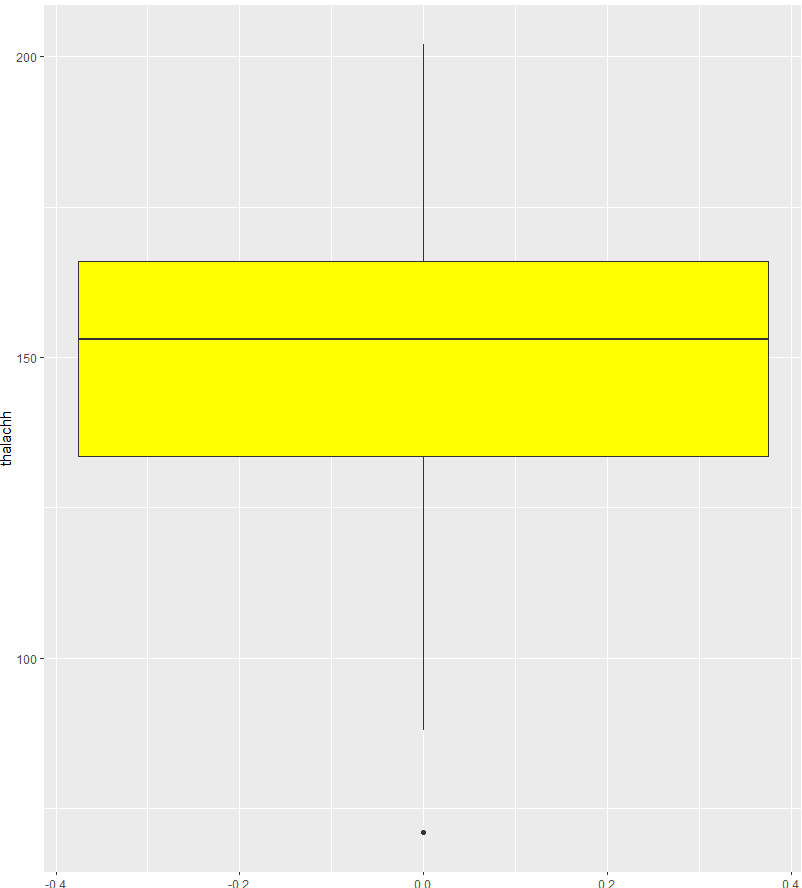
Resting Blood Pressure (in mm Hg)



We can see that we were able to observe 6 outliers for resting blood pressure out of which 3 are 180 or above. Since blood pressure readings more than 180/110 are considered hypertensive emergencies, we do not remove these outliers.

**Thalachh**

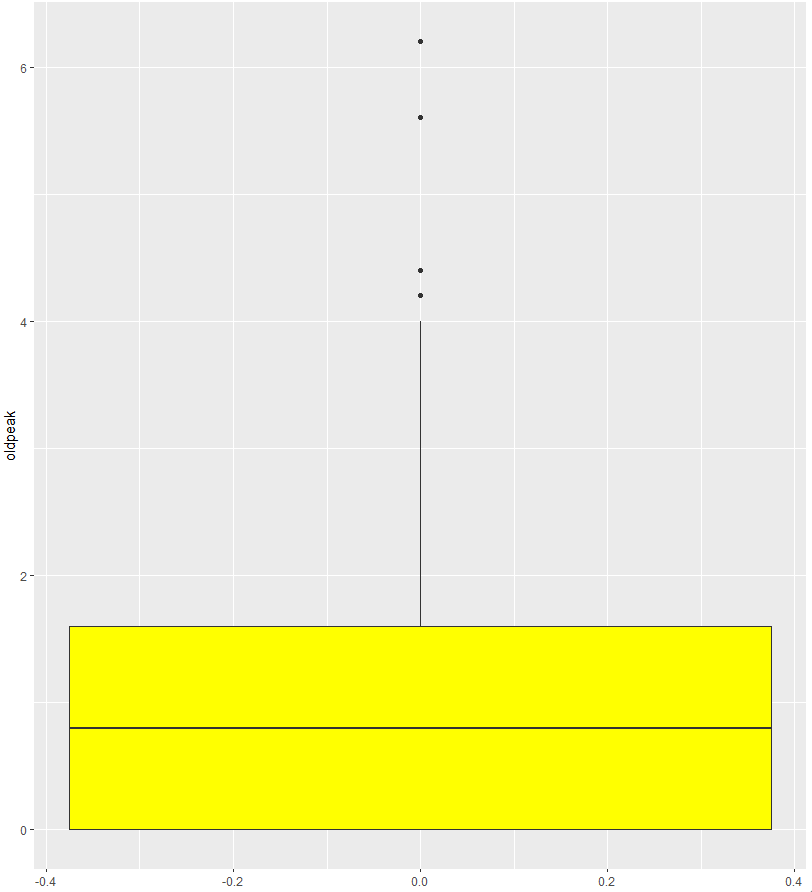
Maximum Heart Rate Achieved (Scale of 71 – 202)



We observe one outlier that is below 75 but since it lies in the normal resting heart rate range, we do not remove it since it indicates a healthy subject

**Oldpeak**

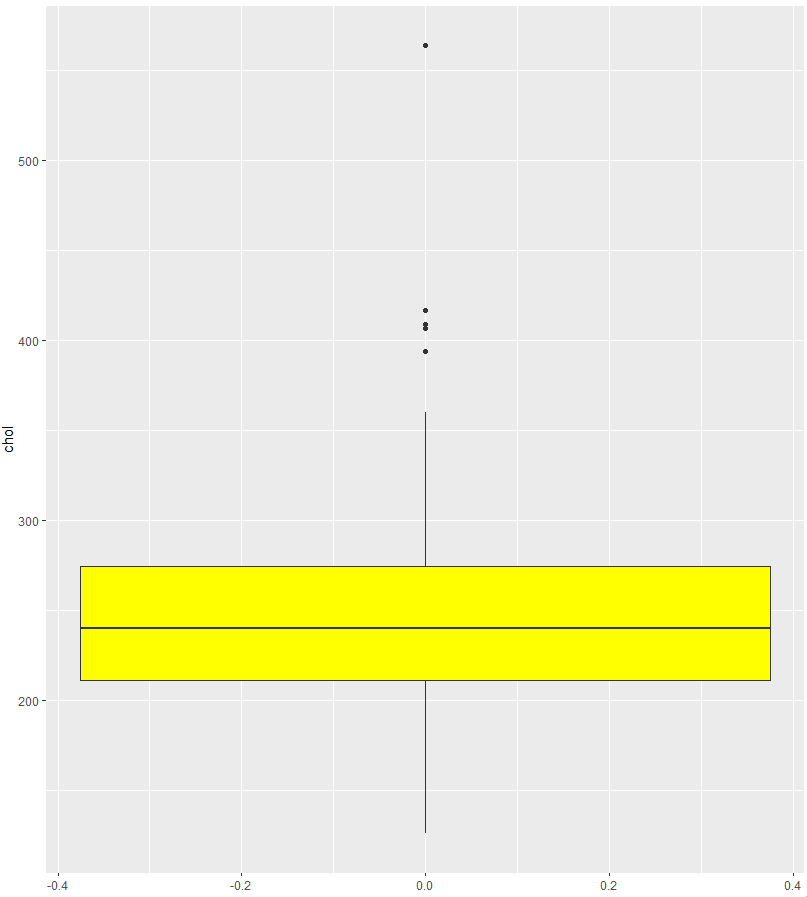
ST Depression Induced by Exercise Relative to Rest



We observe 4 outliers that are greater than 4 in this variable but since higher oldpeak values can indicate serious heart conditions, we do not remove them.

**Chol**

Cholesterol (in mg/dl)

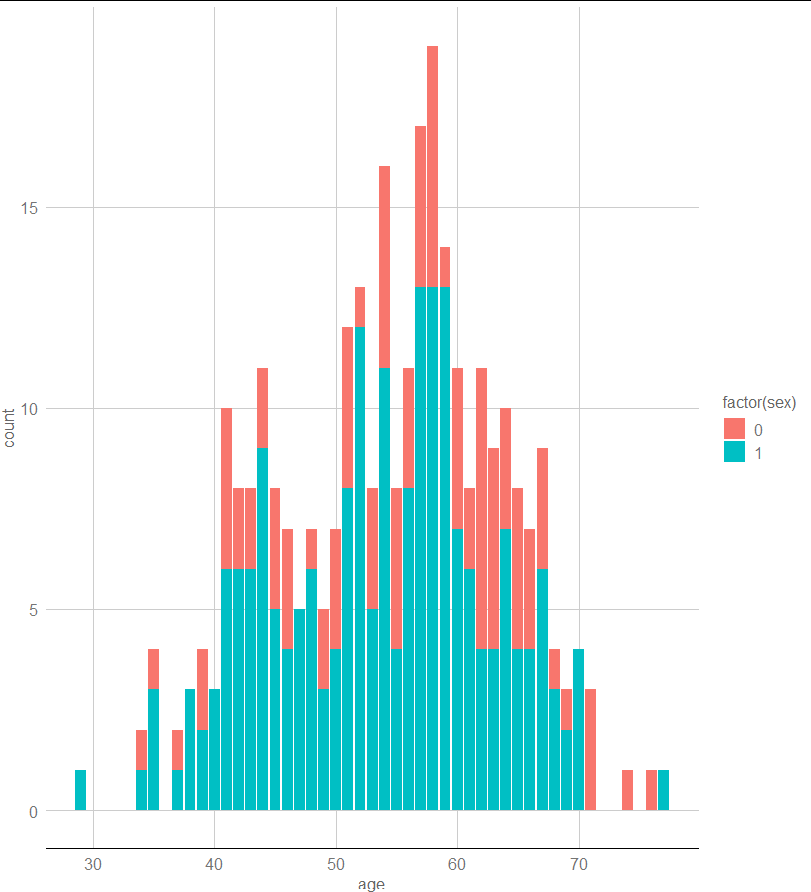


We observed 5 outliers that are around 400 or more. The normal cholesterol levels are around 200-240 so we do not remove these outliers since they represent critical levels which would require immediate attention.

### ggplot – Histogram

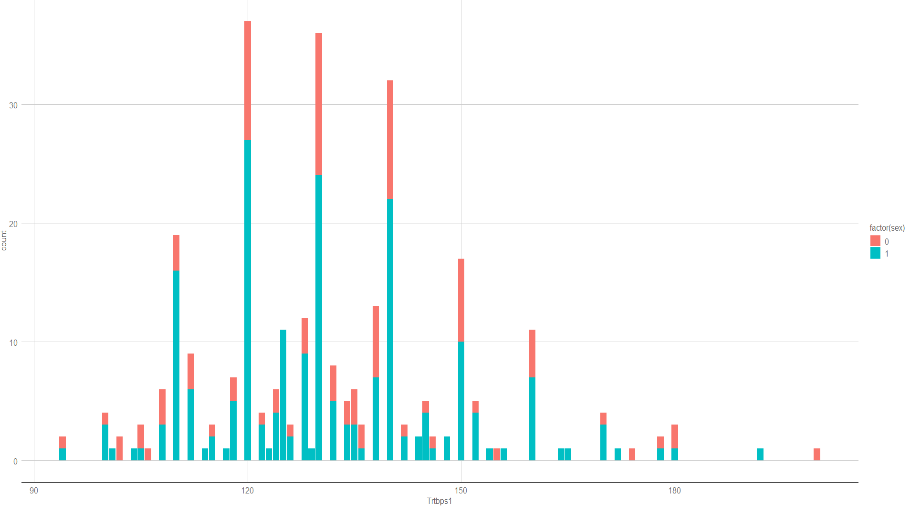
**Age**

The histogram below shows us the graphical representation of the data gathered on age. The minimum value is 29 years, mean is 54.37 years, and the maximum value is 77 years.



**Trtbps**

This graph shows us the resting blood pressure graphical data. We can see that the minimum value is 94 units, mean is 131.6 units, and the maximum value is 200 units.



### Scatterplot



As shown above, elderly women between the ages of 60-70 have higher chances of having cholesterol. Whereas coming to men, we cannot really infer which ages of men have higher chances of cholesterol.

## Data Modelling

### Splitting the dataset:

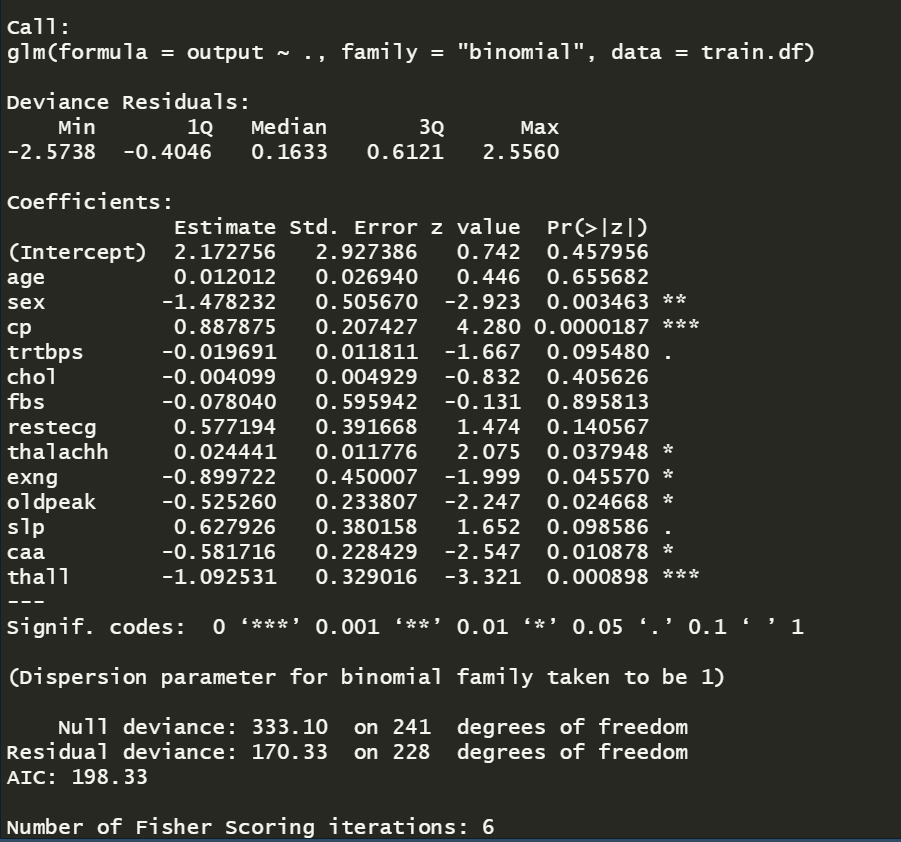
The data set was first split into two parts: training data set and validation data set. 80% of the data set was used for training, while the remaining 20% was used for validation data set.

### Logistic Regression:

It broadens the application of regression analysis to cases in which the response variable is binary, and the primary outcome is categorical, as well as the proper regression analysis to carry out in those cases. Sex, cp, trestbps, ca, thalach, exang, slope are significant variables. The deviance using null illustrated how good the response will be predicted by using the model with intercept. The deviation of residual displays how well the model predicts the answer by using the predictors which are considered while working with them. While doing this process we can find out the error of measure. This tells us that residual deviance is the error of measure. Whenever there is in a smaller amount of residual deviance the good is the model’s predictive power. In the output we get the residual deviance smaller than the null deviance, or a logistic regression model has some predictive power, and the variables will have some explanatory power.

The odds ratio in logistic regression indicates the ongoing influence of a predictor X just on probability of a particular outcome. It is expected that the logistic transformation of the variable outcome has a linear correlation with the predictor factors whenever a binary variable outcome is by using this regression analysis. This makes it challenging to interpret the regression coefficients. Therefore, we use the odds ratio.

**Summary of Logistic Regression Model**



**Exponents of Coefficients**

Text

Description automatically generated

AIC:

We choose the best model we can with all the important variables by using AIC. The residual deviation is less than the null variance though in the IC model, indicating some predictive potential for the r model.

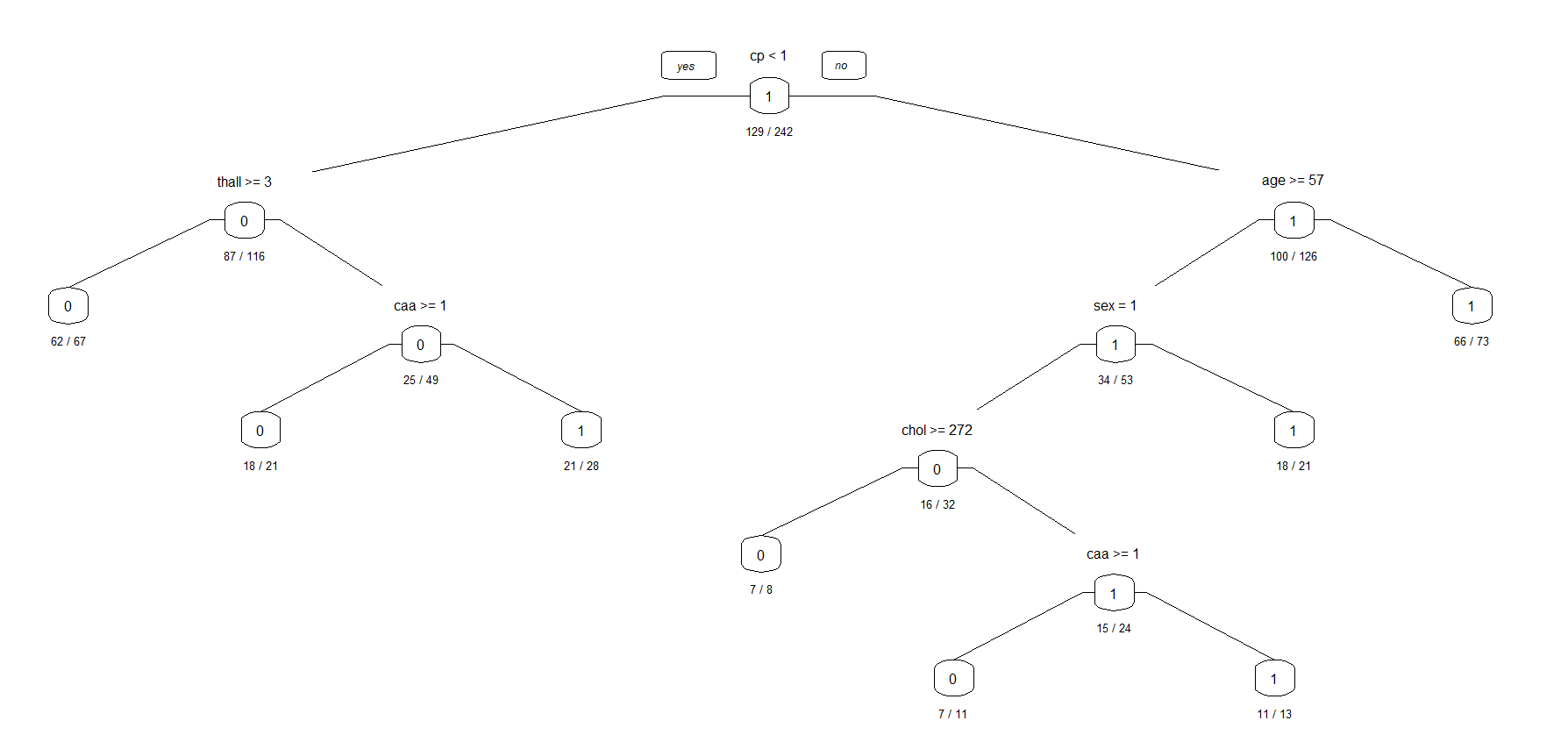
**Summary of Logistic Regression Model with stepAIC**

Table

Description automatically generated

### Decision Tree

The decision tree is the most efficient and renowned classification and prediction method. A decision tree is a tree structure resembling a flow chart, in which each internal node indicates a test on an attribute, each branch shows the test's result, and each leaf node has a class label. From the decision tree we get the following rule with the most percentage cover of cases.



Diagram, schematic

Description automatically generated

### Pruning

The pruning technique of Decision Trees is tuning the hyperparameters prior to the training pipeline. It stops the tree-building process to avoid producing leaves with small samples. During each stage of the splitting of the tree, the cross-validation error will be monitored. If the value of the error does not decrease anymore - then we stop the growth of the decision tree. We have performed pre-pruning of the decision tree to increase the accuracy of the model

Diagram, radar chart

Description automatically generated

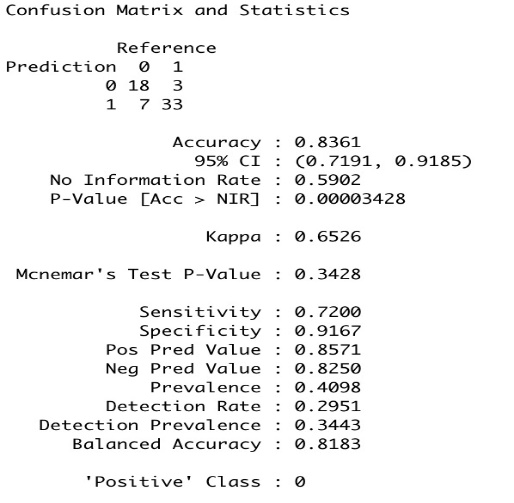
# Performance Evaluation:

Understanding and knowing the model's prediction error rate can help you understand how well a regression model performs. Understanding how well the regression line matched the dataset and understanding the correctness of such methods can also help you gauge performance.

## Confusion Matrix:

The confusion matrix is a measurement that tracks the number of errors, including false positives and false negatives, to show how well a classification model performs. We display the accuracy of our training data using the confusion matrix.

**Confusion matrix for logistic regression:**

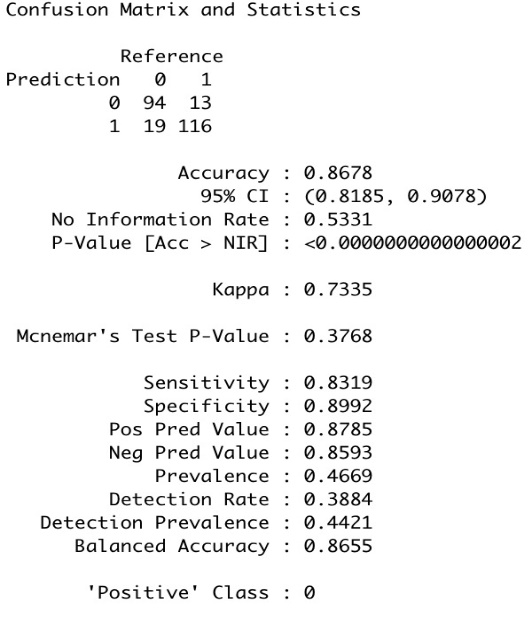


**Confusion matrix for logistic regression with StepAIC:**

Text

Description automatically generated

**Confusion Matrix for Decision Tree:**



**Confusion Matrix for Decision Tree after Pruning:**

Text

Description automatically generated

## ROC Curve:

We utilize the AUC which is area under the curve and ROC which is known as Receiver Operating Characteristics curves for the classification issue to evaluate or visualize the performance of the classifier problem. It is among the most crucial assessment matrices for assessing the effectiveness of any different classifiers. It can also be expressed as an AROC. The probability curve is called ROC, and the amount or measure of distinction is called AUC. It reveals how well the model can discriminate between classes. The model performs better at detecting 0 as 0 and 1 as 1 which will be higher the AUC. Let’s take an example on how the model is better at differentiating, as which of the model is more elevated AUC. A practical model has an AUC close to 1, which indicates that it has a strong level of separability. Our algorithm has the weakest metric of separability since its AUC is close to the 0. This indicates that the outcome is reversing. It predicts that zeros will be ones and ones will be zeros. Additionally, a model has absolutely no potential for class separation when AUC is 0.5.

**ROC Curve for Training dataset**

A picture containing chart

Description automatically generated

**Area under the curve: 0.9191**

**ROC Curve for Validation dataset**

A picture containing chart

Description automatically generated

**Area under the curve: 0.8183**

# Process Flow Diagram

**Diagram

Description automatically generated**

# Conclusion

We created and evaluated logistic regression and decision tree models and recorded the outcomes and performance measures for each. We were 83.6% accurate with the regression model, and with the decision tree, we were 86.78% accurate. After pruning, we were able to improve the accuracy to 88.43%. The decision tree approach is superior for our project's analysis because of this. They also exhibit greater accuracy than the logistic regression model in this case and are simple to use and evaluate. The constructed decision tree algorithm enabled us to pinpoint thal, ca, thalach, and cp as the most crucial heart attack predictors. This demonstrates that heart attacks are not primarily caused by old age or high cholesterol.

This model may be used to determine whether a certain patient with a particular health profile is likely to have a heart attack. This model may be used to predict outcomes with greater precision and less error for larger populations. The model's predictions can be used as a starting point to enhance methods to research different medical aspects that may aid in preventing heart attacks. Similar projects may be created to analyse other vulnerable demographics and assist them in better serving society using analytics and various categorization methods.

# Reference

1. Moonesinghe R, Yang Q, Zhang Z, Khoury MJ. Prevalence and cardiovascular health impact of family history of premature heart disease in the United States: Analysis of the National Health and Nutrition Examination Survey, 2007-2014
2. Fang I, Luncheon C, Ayala C, Odom E, Loustalot F. Awareness of heart attack symptoms and response among adults in - United States, 2008, 2014, and 2017. MMWR. 2019;58(5):101 6.
3. Singh P, Singh S, Pandi-Jain GS. “Effective heart disease prediction system using data mining techniques” in International Journal of Nanomedicine, 13(T-NANO 2014 Abstracts):121-124, 2018

# Link to the presentation

[Call with BA with R-20221206\_200847-Meeting Recording.mp4](https://cometmail-my.sharepoint.com/:v:/g/personal/pxp220005_utdallas_edu/EfruTnP6mpRDrPIIM0mY_0oBcds2-ULDV4VM_1edOjtM7g?e=ESjNKB)