Fall 2022

Project: BUAN.6356.006 Business Analytics with R

Heart Attack Prediction

Team members: (Group 16)
Draksharam, Vishnu Paschyanti (vxd200023)
Jawahar Vasagam, Premi (pxj220007)
Farzana M B, Ashika (axf210029)
Tallapally, Jahnavi (jxt200051)
Parthasarathi, Prriyamvradha (pxp220005)

Table of Contents

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

<u>**Dependent Variable**</u>	<u>**Type**</u>	<u>**Description**</u>
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.

```
> sum(!complete.cases(heart.df))
[1] 0
```

Exploratory Data Analysis:

Summary:

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

```
> head(heart.df)
  age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output
1  63   1  3    145  233   1       0      150    0     2.3   0   0     1      1
2  37   1  2    130  250   0       1      187    0     3.5   0   0     2      1
3  41   0  1    130  204   0       0      172    0     1.4   2   0     2      1
4  56   1  1    120  236   0       1      178    0     0.8   2   0     2      1
5  57   0  0    120  354   0       1      163    1     0.6   2   0     2      1
6  57   1  0    140  192   0       1      148    0     0.4   1   0     1      1
```

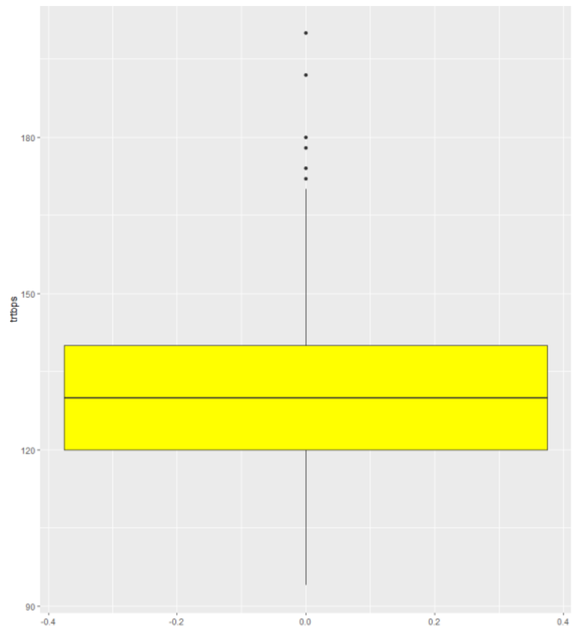
The following image shows the summary statistics for the different variables that we have in our dataset.

```
> summary(heart.df)
      age             sex               cp              trtbps           chol            fbs        
 Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0   Min.   :126.0   Min.   :0.0000  
 1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.0000  
 Median :55.00   Median :1.0000   Median :1.000   Median :130.0   Median :240.0   Median :0.0000  
 Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6   Mean   :246.3   Mean   :0.1485  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:0.0000  
 Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0   Max.   :564.0   Max.   :1.0000  
    restecg          thalachh          exng           oldpeak          slp             caa        
 Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
 Median :1.0000   Median :153.0   Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
 Mean   :0.5281   Mean   :149.6   Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
 3rd Qu.:1.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
 Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
     thall           output      
 Min.   :0.000   Min.   :0.0000  
 1st Qu.:2.000   1st Qu.:0.0000  
 Median :2.000   Median :1.0000  
 Mean   :2.314   Mean   :0.5446  
 3rd Qu.:3.000   3rd Qu.:1.0000  
 Max.   :3.000   Max.   :1.0000  
```

Finding Outliers using Box Plot:

Trtbps

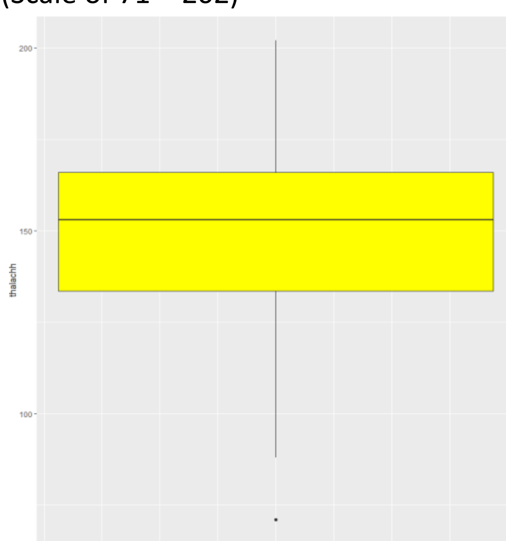
Resting Blood Pressure (in mm Hg)



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

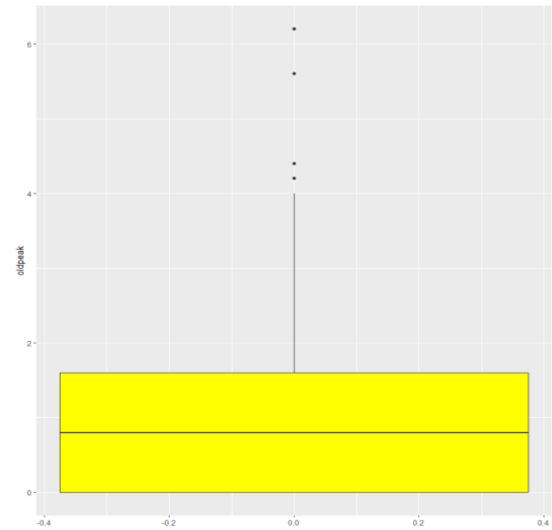
Maximum Heart Rate Achieved (Scale of 71 – 202)



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

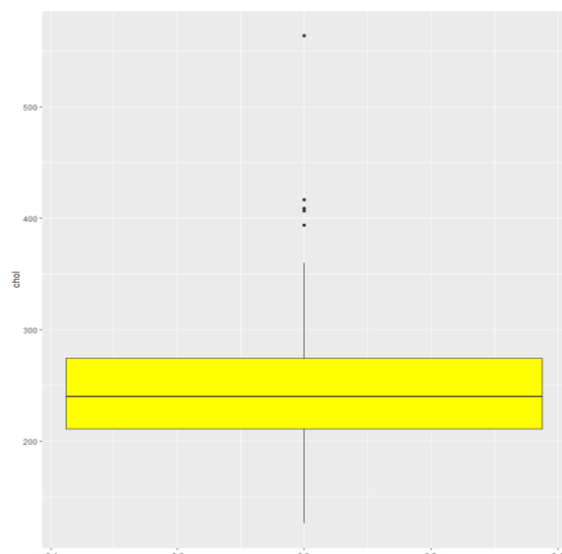
ST Depression Induced by Exercise Relative to Rest



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)

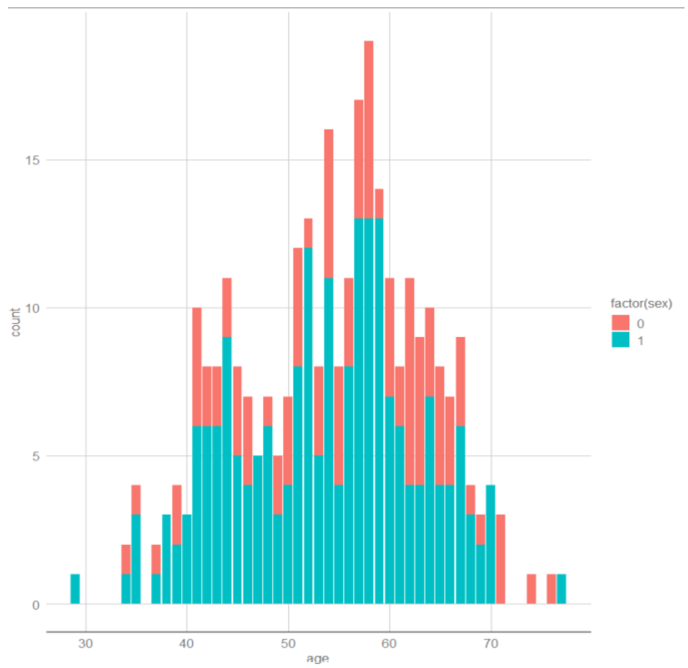


We observed 5 outliers that are around 400 or more. The normal cholesterol levels are around 200-240 so we do not remove these outliers since they represent critical levels which would require immediate attention.

ggplot – Histogram

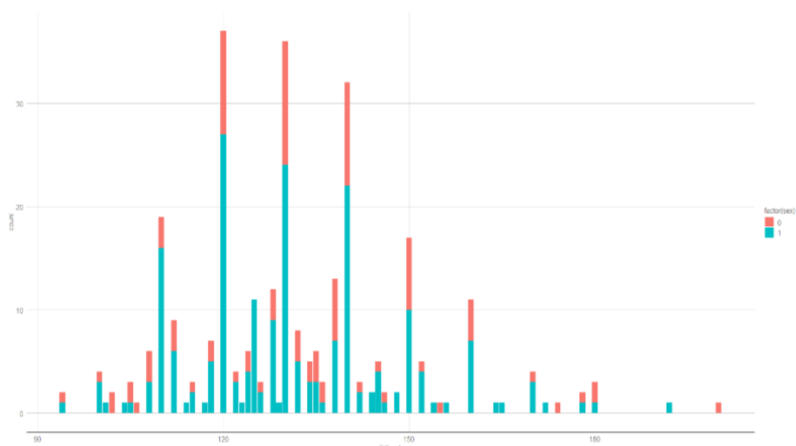
Age

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.

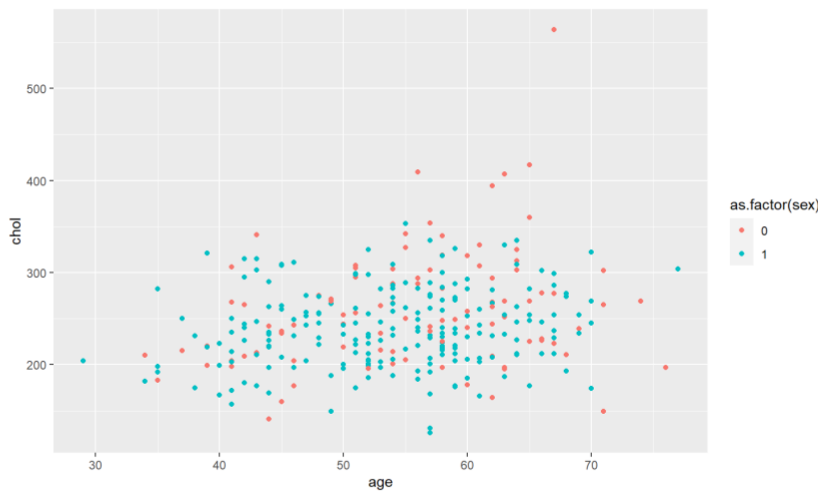


Trtbps

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

```
Call:
glm(formula = output ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5738  -0.4046   0.1633   0.6121   2.5560

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  2.172756   2.927386   0.742  0.457956
age          0.012012   0.026940   0.446  0.655682
sex         -1.478232   0.505670  -2.923  0.003463 **
cp           0.887875   0.207427   4.280 0.0000187 ***
trtbps      -0.019691   0.011811  -1.667  0.095480 .
chol        -0.004099   0.004929  -0.832  0.405626
fbs         -0.078040   0.595942  -0.131  0.895813
restecg      0.577194   0.391668   1.474  0.140567
thalachh     0.024441   0.011776   2.075  0.037948 *
exng        -0.899722   0.450007  -1.999  0.045570 *
oldpeak     -0.525260   0.233807  -2.247  0.024668 *
slp          0.627926   0.380158   1.652  0.098586 .
caa         -0.581716   0.228429  -2.547  0.010878 *
thall       -1.092531   0.329016  -3.321  0.000898 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 333.10  on 241  degrees of freedom
Residual deviance: 170.33  on 228  degrees of freedom
AIC: 198.33

Number of Fisher Scoring iterations: 6
```

Exponents of Coefficients

```
Call:  glm(formula = output ~ ., family = "binomial", data = train.df)

Coefficients:
(Intercept)          age          sex           cp       trtbps         chol          fbs      restecg
   3.125534    -0.001503    -1.824557     0.986633    -0.020590    -0.004876     0.043432     0.248093
   thalachh         exng      oldpeak          slp          caa        thall
   0.022209    -0.562949    -0.392256     0.706198    -0.645739    -0.883016

Degrees of Freedom: 241 Total (i.e. Null);  228 Residual
Null Deviance:      334.4
Residual Deviance: 174.7        AIC: 202.7
```

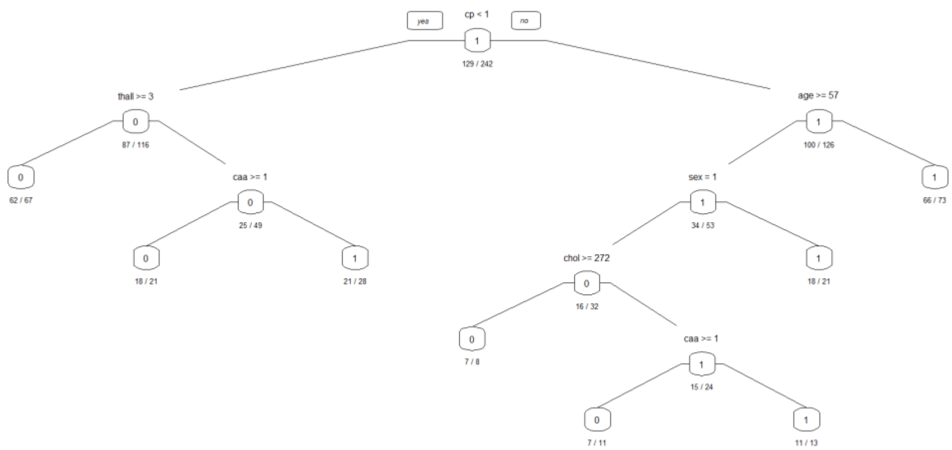
AIC:
We choose the best model we can with all the important variables by using AIC. The residual deviation is less than the null variance though in the IC model, indicating some predictive potential for the r model.

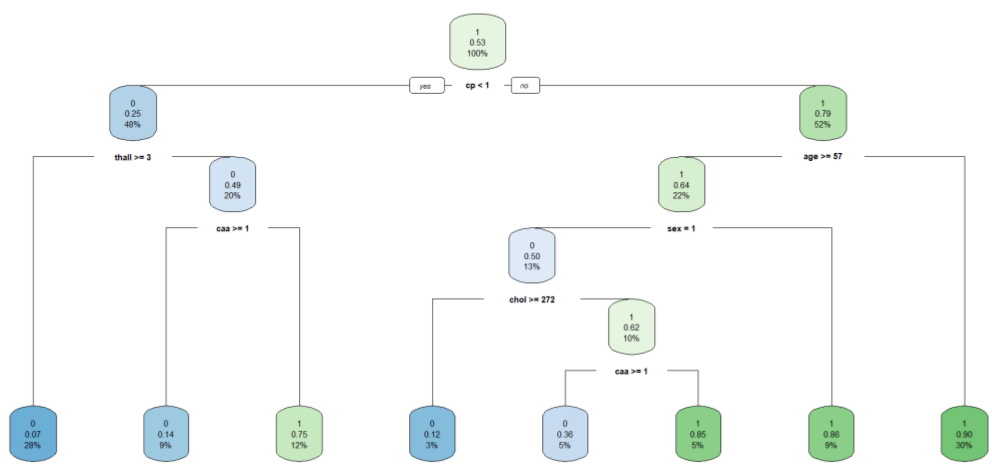
Summary of Logistic Regression Model with stepAIC

```
> summary(backwards)

Call:
glm(formula = output ~ sex + cp + trtbps + thalachh + oldpeak +
    slp + caa + thall, family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4866  -0.4732   0.1478   0.5906   2.5266

Coefficients:
            Estimate Std. Error z value     Pr(>|z|)
(Intercept)  1.93478    1.99215   0.971     0.331447
sex         -1.72314    0.48313  -3.567     0.000362 ***
cp           1.07367    0.20125   5.335 0.0000000955 ***
trtbps      -0.02389    0.01047  -2.282     0.022503 *
thalachh     0.02417    0.01020   2.370     0.017792 *
oldpeak     -0.41966    0.22605  -1.857     0.063378 .
slp          0.75322    0.37947   1.985     0.047151 *
caa         -0.64001    0.20420  -3.134     0.001723 **
thall       -0.95692    0.30594  -3.128     0.001761 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 334.42  on 241  degrees of freedom
Residual deviance: 178.40  on 233  degrees of freedom
AIC: 196.4

Number of Fisher Scoring iterations: 6
```

Decision Tree

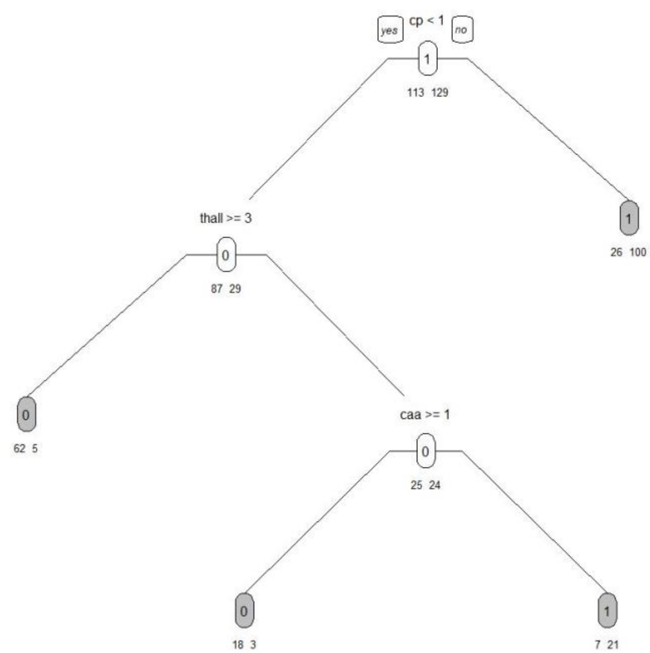
The decision tree is the most efficient and renowned classification and prediction method. A decision tree is a tree structure resembling a flow chart, in which each internal node indicates a test on an attribute, each branch shows the test's result, and each leaf node has a class label. From the decision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





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



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 and Statistics

          Reference
Prediction  0  1
         0 18  3
         1  7 33

               Accuracy : 0.8361
                 95% CI : (0.7191, 0.9185)
    No Information Rate : 0.5902
    P-Value [Acc > NIR] : 0.00003428

                  Kappa : 0.6526

 Mcnemar's Test P-Value : 0.3428

            Sensitivity : 0.7200
            Specificity : 0.9167
         Pos Pred Value : 0.8571
         Neg Pred Value : 0.8250
             Prevalence : 0.4098
         Detection Rate : 0.2951
   Detection Prevalence : 0.3443
      Balanced Accuracy : 0.8183

       'Positive' Class : 0
```

Confusion matrix for logistic regression with StepAIC:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 17  4
         1  8 32

               Accuracy : 0.8033
                 95% CI : (0.6816, 0.894)
    No Information Rate : 0.5902
    P-Value [Acc > NIR] : 0.0003498

                  Kappa : 0.5831

 Mcnemar's Test P-Value : 0.3864762

            Sensitivity : 0.6800
            Specificity : 0.8889
         Pos Pred Value : 0.8095
         Neg Pred Value : 0.8000
             Prevalence : 0.4098
         Detection Rate : 0.2787
   Detection Prevalence : 0.3443
      Balanced Accuracy : 0.7844

       'Positive' Class : 0
```

Confusion Matrix for Decision Tree:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  94  13
         1  19 116

               Accuracy : 0.8678
                 95% CI : (0.8185, 0.9078)
    No Information Rate : 0.5331
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.7335

 Mcnemar's Test P-Value : 0.3768

            Sensitivity : 0.8319
            Specificity : 0.8992
         Pos Pred Value : 0.8785
         Neg Pred Value : 0.8593
             Prevalence : 0.4669
         Detection Rate : 0.3884
   Detection Prevalence : 0.4421
      Balanced Accuracy : 0.8655

       'Positive' Class : 0
```

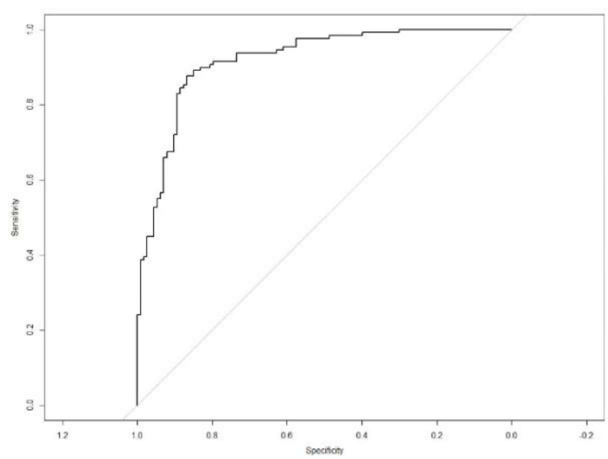
Confusion Matrix for Decision Tree after Pruning:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 100  15
         1  13 114

               Accuracy : 0.8843
                 95% CI : (0.8371, 0.9217)
    No Information Rate : 0.5331
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.7678

 Mcnemar's Test P-Value : 0.8501

            Sensitivity : 0.8850
            Specificity : 0.8837
         Pos Pred Value : 0.8696
         Neg Pred Value : 0.8976
             Prevalence : 0.4669
         Detection Rate : 0.4132
   Detection Prevalence : 0.4752
      Balanced Accuracy : 0.8843

       'Positive' Class : 0
```

ROC Curve:

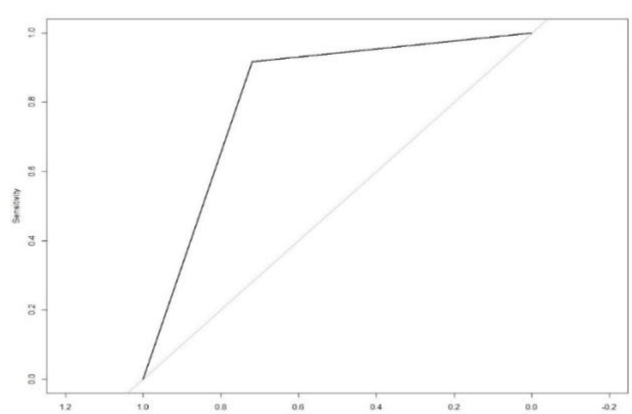
We utilize the AUC which is area under the curve and ROC which is known as Receiver Operating Characteristics curves for the classification issue to evaluate or visualize the performance of the classifier problem. It is among the most crucial assessment matrices for assessing the effectiveness of any 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



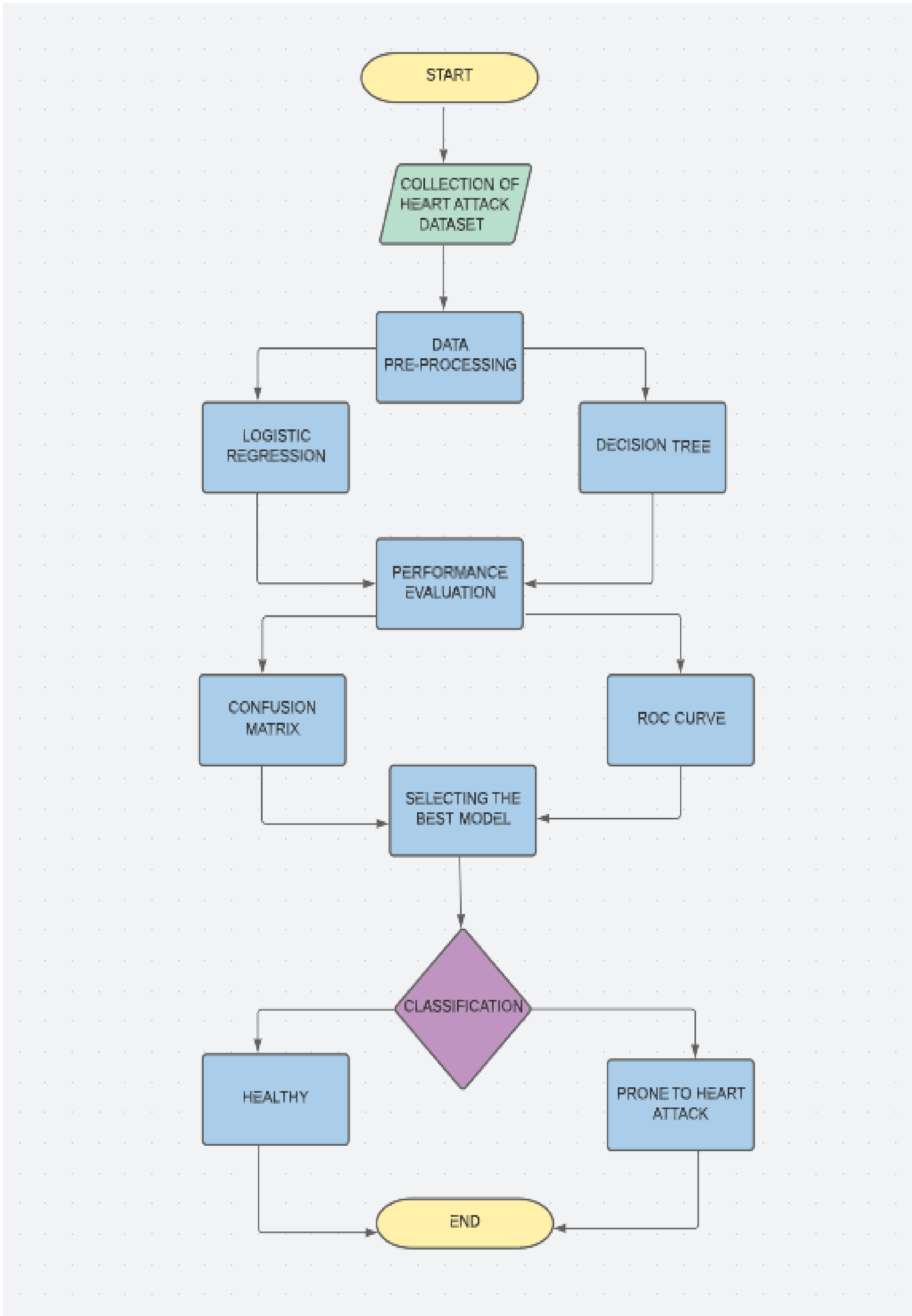
Area under the curve: 0.9191

ROC Curve for Validation dataset



Area under the curve: 0.8183

Process Flow Diagram



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]