Heart Disease – Identification of Predictors and Prevention

Darshil Shankar[1], Neel Viradiya[2], Prof. Chirag Prajapati[3]

[1,2,3]*Department of Computer Applications, SDJ International College, Surat, Gujarat, India*

Abstract – *Heart Diseases are one of the major problems prevalent today due to health and lifestyle choices. Our main objective in the project is to predict the chances of Heart Disease and the major factors contributing to it. The data set we found helps us to evaluate deeper relationships between potential factors and perhaps reshape health care products.*

***Key Words*: Heart Disease, Tree Model, Logistic Regression Model, R**

1. EXECUTIVE SUMMARY

We used R, a statistical computing and graphics tool for building our models. At first, we explored the dataset to understand our dataset and develop a general idea for further analysis. We found some correlations among variables. Then we used this dataset to build classification models, including the Decision Tree model and Logistic Regression model, to predict whether someone, with certain diagnostic measurements, has chances of getting Heart Disease or not. For the Decision Tree model, we sampled 80% of records as training datasets 20% of records as validation data sets, plotted the Decision Tree, and evaluated it using confusion matrix and ROC. For the Logistic Regression model, also we sampled training data sets and validation data sets, built the Logistic Regression model, computed the odds ratios, and evaluated the Logistic Regression model using the confusion matrix and ROC. We evaluated all models and selected the best possible model.

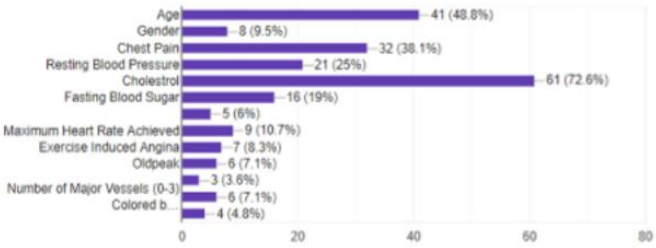
2. INTRODUCTION

The motivation of this work comes from the fact that although the death rates from Heart Diseases are declining: Heart Disease is still the major cause of death in the USA. An estimated 92.1 million US adults have at least one kind of Heart Disease and by 2030, around 44% of the US adult population is expected to have some form of Heart Disease.

Heart Disease comes under many categories such as coronary artery disease, heart rhythm problems, chest pain (angina), or stroke. In using data from the Global Burden of Disease Study, approximately 90% of the stroke risk could be attributed to modifiable risk factors. In our study, we tried identifying those risk factors that contribute the most to Heart Diseases, "The estimated direct costs of Heart Diseases and stroke increased from $ 103.5 billion in 1996 to 1997 to $ 213.8 billion from 2014 to 2015.

Research: What is the biggest contributor of Heart Disease?

84 Responses



We can see here, from the 84 respondents, 61 or 72.6% think Cholesterol and 41 or 48.8% think Age was the biggest contributor to Heart Disease.

A Machine Learning model was to predict the risk of a Heart Disease in the subjects and these predictions were compared to the actual experiences of the subjects over fifteen years. The predicted machine learning scores aligned accurately with the actual distribution of observed events. Experimental results show 100% accurate prediction for the system using Neural Networks.

3. PREPROCESSING DATA

We found a total 6 null values in the dataset. The thal column contained 2 and the ca column contained 4 of the null values. We have removed these values in order to make the dataset accurate.

The following shows the data and the summary statistics:

```
head(heart_data.dt)
```

```
##    age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1:  67   1  4      160  286   0       2     108     1     1.5     2  3    0
## 2:  67   1  4      120  229   0       2     129     1     2.6     2  2    2
## 3:  37   1  3      130  250   0       0     187     0     3.5     3  0    0
## 4:  41   0  2      130  204   0       2     172     0     1.4     1  0    0
## 5:  56   1  2      120  236   0       0     178     0     0.8     1  0    0
## 6:  62   0  4      140  268   0       2     160     0     3.6     3  2    0
##    target
## 1:      1
## 2:      1
## 3:      0
## 4:      0
## 5:      0
## 6:      1
```

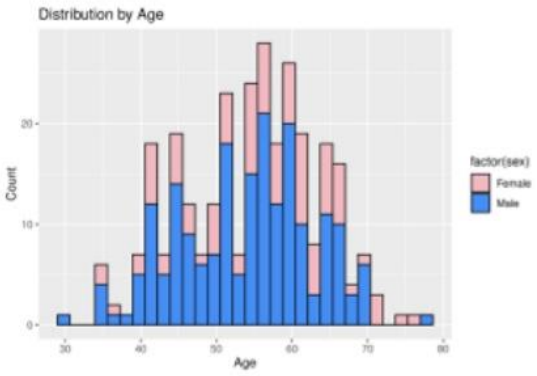
```
summary(heart_data.dt)
```

```
##       age             sex               cp            trestbps
##  Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0
##  1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0
##  Median :56.00   Median :1.0000   Median :3.000   Median :130.0
##  Mean   :54.51   Mean   :0.6757   Mean   :3.166   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0
##       chol             fbs            restecg          thalach
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.0
##  Median :243.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :247.4   Mean   :0.1419   Mean   :0.9932   Mean   :149.6
##  3rd Qu.:276.2   3rd Qu.:0.0000   3rd Qu.:2.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##      exang           oldpeak          slope             ca
##  Min.   :0.0000   Min.   :0.000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.800   Median :2.000   Median :0.0000
##  Mean   :0.3277   Mean   :1.051   Mean   :1.598   Mean   :0.6791
##  3rd Qu.:1.0000   3rd Qu.:1.600   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.200   Max.   :3.000   Max.   :3.0000
##       thal            target
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.8345   Mean   :0.4628
##  3rd Qu.:2.0000   3rd Qu.:1.0000
##  Max.   :2.0000   Max.   :1.0000
```

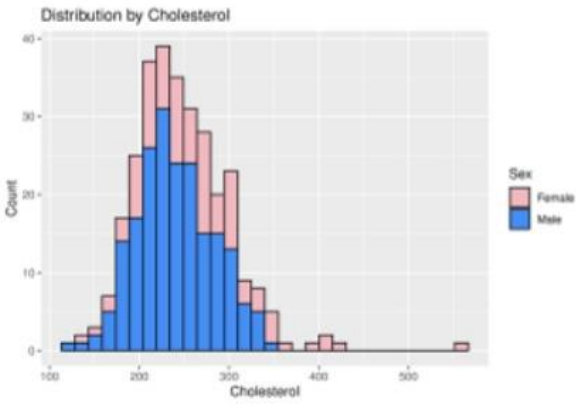
4. EXPLORATORY DATA ANALYSIS (EDA)

4.1 Graphical Analysis

1. **Histogram for Age:**



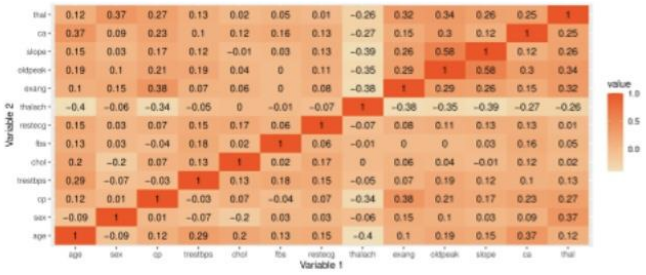
2. **Histogram for Cholesterol:**



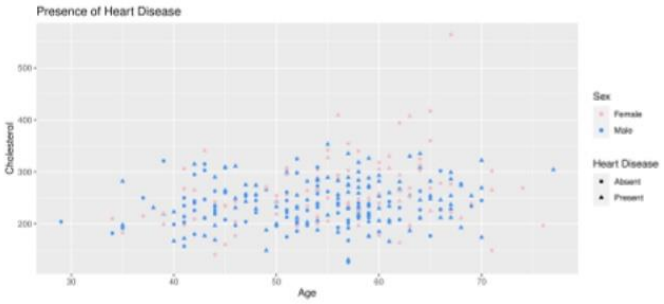
4.2 Correlation heatmap between variables:

We used a correlation heatmap to plot out the strongest positive and negative correlations. As indicated, "slope" and "oldpeak" have the strongest correlation, with a positive 0.58. "thalach" and *age" have the lowest correlation, with a negative 0.4.

Which variables are highly correlated?



4.3 Analyzing relationships between cholesterol and age:



As we can see in the scatterplot, age and cholesterol do not appear to have a significant correlation with Heart Disease. In the top right corner, we can see an elderly female with high cholesterol, but who does not have Heart Disease. On the other hand, on the bottom left, we can see a young male with low cholesterol, who has who has Heart Disease. Hence, we cannot derive any relationship between these variables and Heart Disease.

5. DATA MODELLING

5.1 Splitting the data:

First, we divided the dataset into two parts: training dataset and validation dataset. We allocated 80% of the dataset for the training dataset and the remaining 20% of the dataset for the validation dataset.

5.2 Logistic Regression:

It extends the idea of Linear Regression to the situation where the outcome variable is categorical. It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Sex, cp, trestbps, ca, thalach , exang, slope are significant variables.

```
## [1] "Summary of  the Logistic Regression Model"

##
## Call:
## glm(formula = target ~ ., family = "binomial", data = train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8171  -0.5359  -0.1926   0.4157   2.2818
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.807764   3.321557  -1.146  0.25164
## age         -0.020362   0.027361  -0.744  0.45675
## sex          1.327736   0.549625   2.416  0.01570 *
## cp           0.525923   0.208812   2.519  0.01178 *
## trestbps     0.021971   0.011453   1.918  0.05506 .
## chol         0.003446   0.004294   0.803  0.42220
## fbs         -0.799407   0.640307  -1.248  0.21186
## restecg      0.235885   0.207367   1.138  0.25532
## thalach     -0.027953   0.011914  -2.346  0.01896 *
## exang        1.114523   0.479347   2.325  0.02007 *
## oldpeak      0.264658   0.233532   1.133  0.25709
## slope        0.405958   0.403248   1.007  0.31407
## ca           1.300778   0.297202   4.377 0.000012 ***
## thal         0.675913   0.229062   2.951  0.00317 **
## ---
```

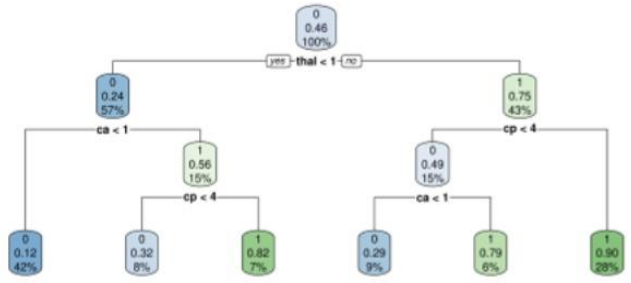
From the output, we can conclude that ca is the most significant variable among all the other variables.

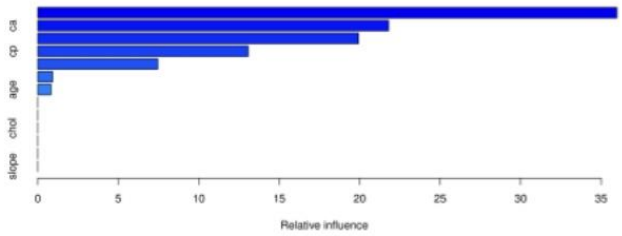
The way we interpret these coefficients is as follows. Considering thalach (maximum heart rate achieved) as example, if it goes up by one unit, the odds of having Heart Disease goes down by 2.8%.

5.3 Decision Tree:

Decision Tree is the most Powerful and popular tool for classification and prediction. Decision Tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label.

From the Decision Tree, we get the following Rule with the most percentage cover of cases. When thal < 1 & ca < 1 THEN CLASS = 0 and this rule covers 42% of cases.






```
##              var    rel.inf
## thal         thal 35.9797241
## ca             ca 21.8011640
## thalach   thalach 19.9336106
## cp             cp 13.0744268
## exang       exang  7.4605795
## oldpeak   oldpeak  0.9263317
## age           age  0.8241632
## sex           sex  0.0000000
## trestbps trestbps  0.0000000
## chol         chol  0.0000000
## fbs           fbs  0.0000000
## restecg   restecg  0.0000000
## slope       slope  0.0000000
```

6. PERFORMANCE EVALUATION

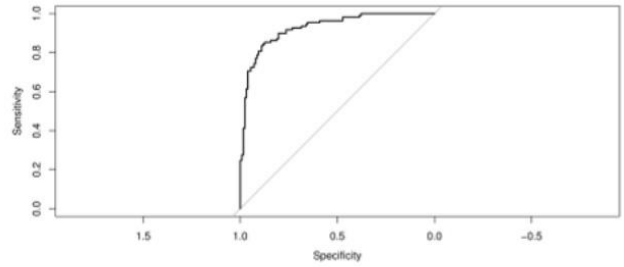
The performance of a regression model can be understood by knowing the error rate of the predictions made by the model. You can also measure the performance by knowing how well your regression line fit the dataset and knowing the accuracy of such models.

Confusion Matrix:

Confusion matrix is a measurement that used to represent the performance of a classification model by recording the sources of errors: false positives and false negatives. We use confusion matrix to depict the accuracy of the training data.

7. ROC CURVE

7.1 ROC for Logistic Regression

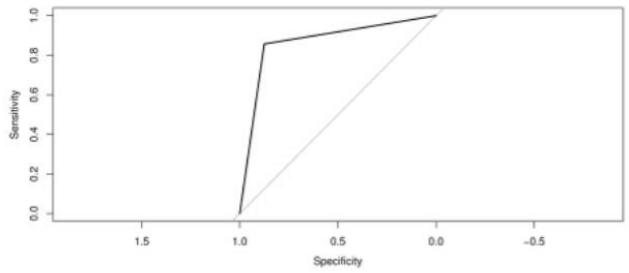



```
## [1] "Area Under the Curve for Decision Tree"

## Area under the curve: 0.9265

## [1] "ROC for Descision Tree"
```

7.2 ROC for Decision Tree

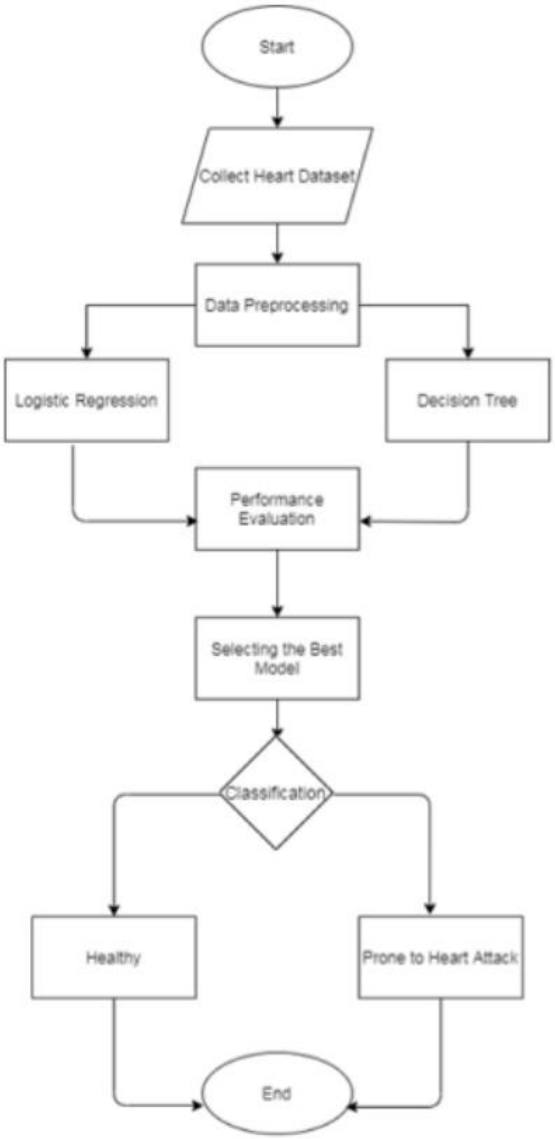



```
## [1] "Area Under the Curve for Decision Tree "

## Area under the curve: 0.8661
```

As we see here, the AUC for the training data is 0.9265 and the AUC for the validation dataset is 0.8661

8. FLOW DIAGRAM



9. CONCLUSION

We have built and compared the Logistic Regression model and the Decision Tree model and have captured the results and the performance metrics for the same. With the Logistic Regression model, we achieved an accuracy of 83.33% and for the Decision Tree we achieved an accuracy of 86.67%. Therefore, the Decision Tree Model is better for our project analysis. They also are easy to implement and interpret, and they display higher accuracy than Logistic Regression Model here. The Decision Tree Model we built helped us identify thal, ca, thalach and cp as the most important predictors of Heart Disease. This proves that age and chol are not major contributors to Heart Disease. Similar models can be built to study other prone populations and help in serving the society better with analytics and various classification techniques.

REFERENCES

[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] Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Virani, S. (2019). Heart disease and stroke statistics-2019 update: A report from the American Heart Association

[3] Fang J, Luncheon C, Ayala C. Odom E, Loustalot F. Awareness of heart attack symptoms and response among adults---United States, 2008, 2014, and 2017. MMWR. 2019; 68(5):101-6.

[4] Al can better predict risk of heart attack, cardiac death, study published in Journal of Cardiovascular Research, https://health.economictimes.indiatimes.com/news/diagnostics/ai-can-better-predict-risk-of-heart-attack-cardiac-death-study/72899878

[5] 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

BIOGRAPHIES

Darshil Shankar

Achieved First Class with Distinction in BCA (Bachelors of Computer Applications)

Neel Viradiya

Achieved First Class with Distinction in BCA (Bachelors of Computer Applications)

Prof. Chirag Prajapati

Professor at BCA department