

Heart attack analysis and prediction using classification modelling.

By: Group A



April 11, 2022

By: Kandi Areji, Mary Bello and Trinath

Contents

[Introduction 2](#_Toc100582107)

[Methods: 2](#_Toc100582108)

[Results 7](#_Toc100582109)

[Discussion: 10](#_Toc100582110)

[APPENDIX 10](#_Toc100582111)

# Introduction

In this project we chose heart attack analysis and prediction dataset, we are interested in finding out the likelihood that a patient will have heart attack or not based on the different conditions and what kind of pain can cause a patient to have heart attack. So that we could recommend that when a patient have certain type of pain to go to the hospital immediately. The dataset contains 14 variables and 303 sample size.

We used R to provide descriptive statistics on the central tendency and the spread of the dataset of patients with different age groups ranging from 29 to 77 and had various conditions. We cleaned up our data and used three classification modelling to train, test and to predict the likelihood that the patient will have heart attack or not if he has the following predictor mentioned above.

# Methods:

In this project we used R to provide the central tendency, the spread of the data and removed the outliers. Using boxplots to view the outliers and used the coefficient formula to remove the outliers to normalize the dataset. We used KNN, CART RF models to carry out the predictions whether a patient that had the for mentioned predictor will have a heart attack or not.

Step 1: Correlation, normalization and Virtualization of the data

Correlation:

The figure below shows the overview of the correlations for all combined variables. The negative correlations show that as one variable increases the other decreases and that they vary in opposite direction. While the positive correlation shows that as one variable increases the other increases therefore, they vary in the same direction.

We looked at the correlation between the output variable and the other variables to determine which one has the strongest/highest positive and negative correlations.

Fig 1. Shows the correlations between the variables.

A close-up of a document

Description automatically generated with medium confidence

Data structure: we used the str() function to view the structure and discovered that there are categorical data that were represented as continuous variables “double”.

The figure below show the structure of the dataset.

Fig 2.

Text

Description automatically generated

The figure below shows the data after using factor() command to change it to categorical data.

Fig3

Calendar

Description automatically generated

Outliers

Dealing with outliers: using boxplot to view the variable we discovered that there were outliers and we used the interquartile range formula to remove the outlier. This resulted in changing the sample size of the data at the end and the structure of our dataset also changed.

heart <- subset(dataset,dataset$variable> (Q1 - 1.5\*IQR(dataset$variable)) & dataset$variable < (Q3+ 1.5\*IQR(dataset$variable)))

The above formular is used to remove the outliers from the following variables in the figures below.

Fig 4

boxplot(heart$thalachh) shows outlier boxplot(heart$thalachh) without outlier

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

Fig 5

boxplot(heart$trtbps) shows outliers boxplot(heart$trtbps) without outliers

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

Fig 6

boxplot(heart$chol) shows outliers boxplot(heart$chol) without outliers

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

Fig 7

boxplot(heart$oldpeak) shows outlier summary(heart$oldpeak) without outliers

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

Normalization of the dataset

Using the formular below we normalized the following 5 variables (Age, trtbps, chol, thalachh, oldpeak) because they are the predictors.

normalize <- function(x) {

num <- x - min(x)

denom <- max(x) - min(x)

return (num/denom)

}

nor <-function(x) { (x -min(x))/(max(x)-min(x)) }

we created another data set with the normalized variable and then combined them to form our normalized dataset.

Fig 8 shows the normalized data set and fig 9 shows the new dataset with 279 outcomes instead of the 303.

normalization and organization

A screenshot of a computer

Description automatically generated with low confidenceFig 8

Fig 9 normalized and organized dataset

Application, table, Excel

Description automatically generated

Descriptive statistics

The figure below shows the descriptive statistics of the dataset after cleaning the data

A screenshot of a computer

Description automatically generated with low confidenceFig 10.

# Results

From our correlation table we discovered that the highest positive correlation was the Chest pain (0.4337) and the highest negative correlation is oldpeak (-0.4306).

In the figure below 0, 1, 2, 3 which represents typical angina, atypical angina, non-anginal pain, and asymptomatic respectively.

As shown in the figure approximately 152 had heart attack while approximately 127 patients did not have heart attack. Patients with typical angina conditions has less chance of having heart attack while patients with non-anginal pain does have likelihood of having a heart attack.

Fig 11. Correlation between chest pain and the heart attack occurring

Chart, bar chart

Description automatically generated

Fig 12. In this figure we see that with oldpeak there less chances of heart attack occurring.

Chart, bar chart

Description automatically generated

As shown in the figure approximately 159 had heart attack while approximately 120 patients did not have heart attack. Patients who had heart attack did not have exercise induced angina.

Fig 13.

Chart, bar chart

Description automatically generated

# Machine learning.

Predictions

Text

Description automatically generated

Confussion Matrix

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

What answer did we find out? We found out that the knn model had the most accuracy and higgest kaapa value greater than 0.7

Did we achieve our purpose?

# Discussion:

What might the answer imply?

Why does it matter?

How does it fit in with what other researchers have done?

# APPENDIX

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Datatype | description | Values |
| Age | double | Patient’s age |  |
| sex | double | Patient’s sex | 0 = female  1 = male |
| exng | double | Exercise induced angina | 0 = no  1 = yes |
| ca | double | Number of major vessels | 0  1  2  3 |
| cp | double | Chest pain type | 0 = typical angina  1 = atypical angina  2 = non-anginal pain  3 = asymptomatic |
| trtbps | double | Resting blood sugar (mm Hg) |  |
| chol | double | Cholestorial in mg/dl fetched via BMI sensor |  |
| fbs | double | Fasting blood sugar | 1 = true  2 = false |
| Rest ecg | double | Resting ECG | 0 = normal  1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)  2 = showing probable or definite left ventricular hypertrophy by Estes' criteria |
| thalach | double | maximum heart rate achieved |  |
| target | double |  | 0= less chance of heart attack  1= more chance of heart attack |