# Introduction

## Scope/Context of this analysis

## High level background

Cardiovascular disease (CVD) is defined as **any serious, abnormal condition of the heart or blood vessels (arteries, veins)**. Cardiovascular disease includes coronary heart disease (CHD), stroke, peripheral vascular disease, congenital heart disease, endocarditis, and many other conditions.

We would like to know what are the factors that increases the possibility of a heart disease

# Data Acquisition, Cleansing, Transformation, Munging

The data was gathered from <https://archive.ics.uci.edu/ml/datasets/heart+disease>

This is multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia. This database includes 76 attributes, but all published studies relate to the use of a subset of 14 of them. The Cleveland database is the only one used by ML researchers to date. One of the major tasks on this dataset is to predict based on the given attributes of a patient that whether that particular person has a heart disease or not and other is the experimental task to diagnose and find out various insights from this dataset which could help in understanding the problem more.

**The dataset was created by: -**

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.

Data Information

|  |  |
| --- | --- |
| Database: | Instances: |
| Cleveland | 303 |
| Hungarian: | 294 |
| Switzerland: | 123 |
| Long Beach VA | 200 |

**Data Dictionary:**

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Name | Description | Type |
| 1 | age | Age of patient | Numeric |
| 2 | sex | Sex of patient | Boolean (1 = male 0 =female) |
| 3 | cp | Type of chest pain | Numeric value: 1 – Typical Angina  2 – Atypical Angina  3 – Non-Angina Pain  4 – asymptomatic |
| 4 | trestbps | Resting Blood Pressure in mm Hg when being admitted to the hospital | Numeric |
| 5 | chol | Serum cholestoral in mg/dl | Numeric |
| 6 | fbs | Fasting blood sugar > 120 mg/dl | Boolean (1 = true; 0 = false) |
| 7 | restecg | Resting electrocardiographic measurement | Numeric value:  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 |
| 8 | thalach | Maximum heart rate achieved | Numeric |
| 9 | exang | Exercise induced angina | Boolean ((1 = yes; 0 = no) |
| 10 | oldpeak | ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more [here](https://litfl.com/st-segment-ecg-library/)) |  |
| 11 | slope | the slope of the peak exercise ST segment | 1 – Upsloping  2 – Flat  3 – Downsloping |
| 12 | ca | number of major vessels | Numeric (0-3) colored by flourosopy |
| 13 | thal | A blood disorder called thalassemia | Numeric 3 = normal; 6 = fixed defect; 7 = reversable defect |
| 14 | target | Heart disease | Boolean 0 = no 1 =yes |

# Data Pre-Processing

With str and summarizeColumns (Table 1), we noticed the following anomalies:

The target feature, target had a cardinality of 5, which should be 2 since target was desiganted as the binary target feature. Ten of the 14 features contained missing values. Notably, the features Slope, CA and Thal had 309, 611 and 486 missing values, respectively.

# trestbps (resting blood pressure) and chol (serum cholestrol) contained several data entries with values of 0 which are not possible for these diagnostic tests.

Fourthly, in the target column clinicians had graded patients as either having no heart disease (value of 0) or displaying various degrees of heart disease (values 1 to 4).

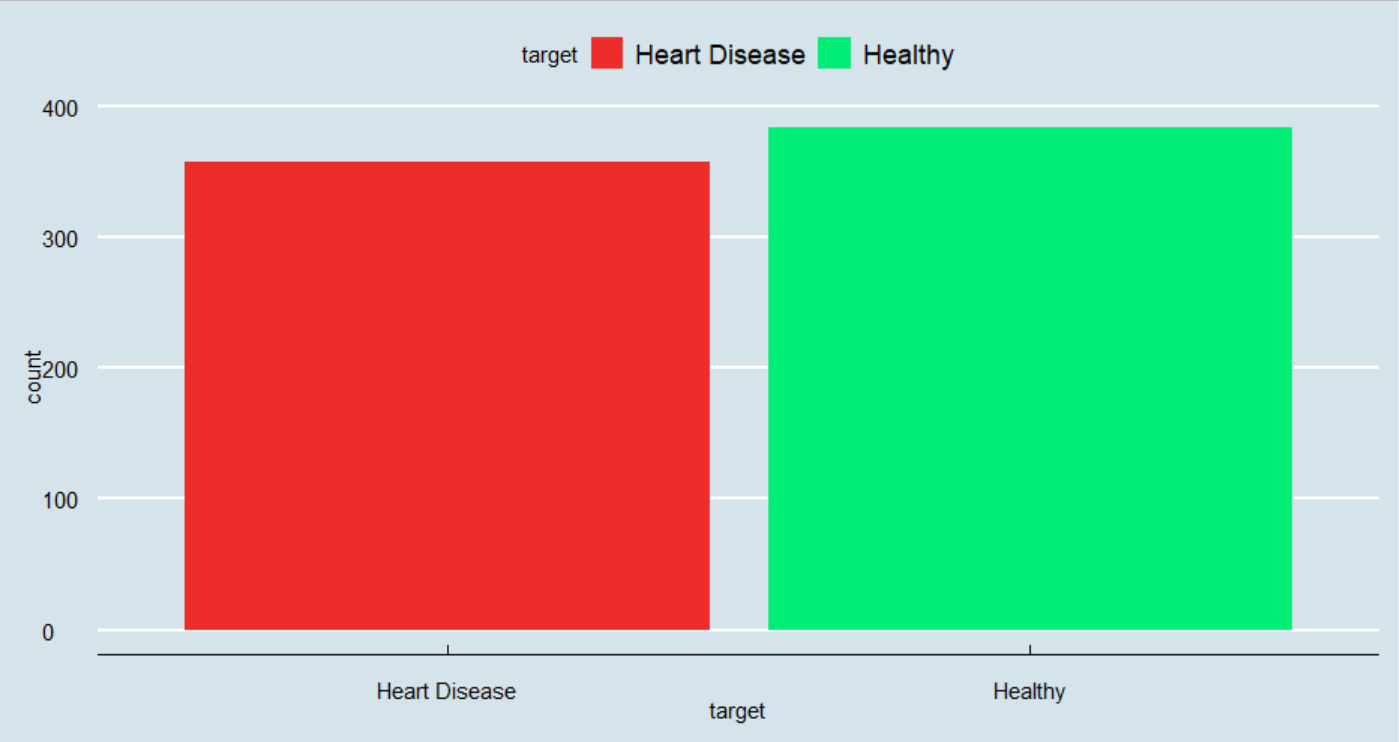
# We chose to group the data into 2 categories of ‘no heart disease’ (value of 0) and ‘displaying heart disease’ (value of 1) so it became binary.

It was noted that a higher proportion of people were diagnosed with heart disease (Table2). Therefore, we may require additional parameter-tuning in building models to cater for such an unbalanced class.

Summary of the data   
Text

Description automatically generated

# Data Visualization



Histogram of all the diseases vs non diseases patient. In the given data, we seem to have more number of healthy people than people with heart disease

Chart, histogram

Description automatically generated

Heart diseases seems to be irrelevant of ages. Both healthy and Heart disease patients seem to follow the same age factor

Chart, histogram

Description automatically generated

No major difference in Rest ECG for Healthy and Heart Disease patients

Chart, histogram

Description automatically generated

More Heart Disease patients seem to have between 200 and 250 mg/dl

Chart, histogram

Description automatically generated

Heart Disease patients have higher maximum heart rate than healthy patients   
Graphical user interface, text

Description automatically generated with medium confidence

Almost all of the patients who have Heart Disease have 0 major vessels as observed by Fluoroscopy

Graphical user interface

Description automatically generated with medium confidence

.

More females have Heart Disease

Chart, bar chart

Description automatically generated

More Heart Disease patients have chest pain type Atypical Angina or Non – Angina Pain 2

Chart, bar chart

Description automatically generated with medium confidence

No difference in fasting blood sugar

Chart, bar chart

Description automatically generated

Patients with Rest ECG 1 have more Heart Diseases

Chart, bar chart

Description automatically generated

Patients with no exercise induced angina have more Heart Disease

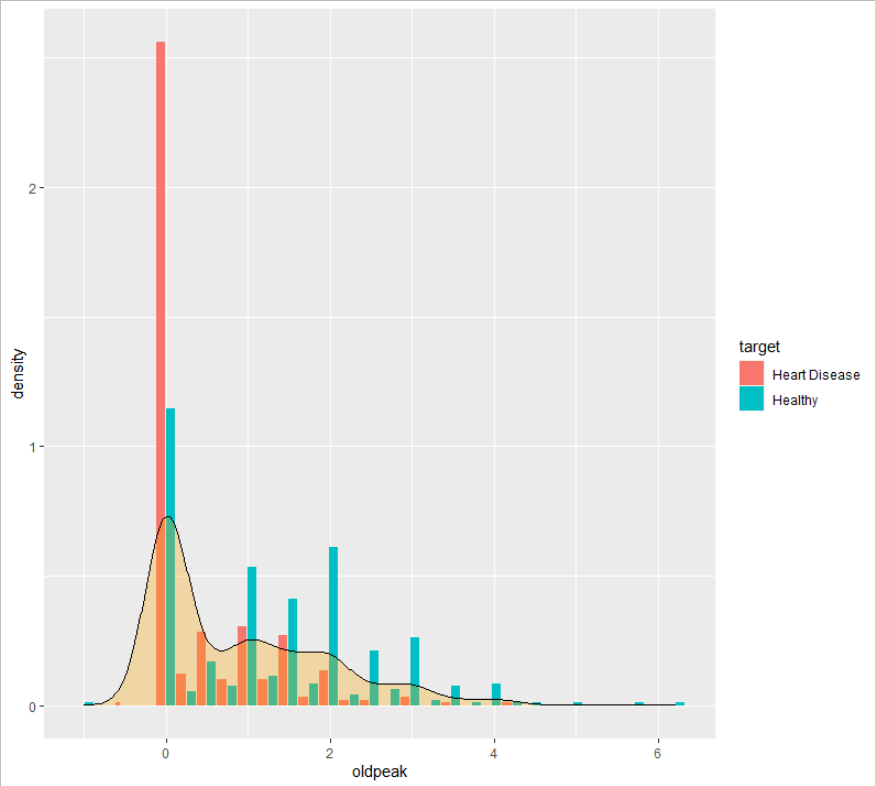
Chart, bar chart

Description automatically generated

Peak exercise ST Slope 2 (flat)have more Heart Disease

Timeline

Description automatically generated



Fixed defect thalasemia has more Heart Disease

Chart, bar chart

Description automatically generated

Comparison of Cholesterol across pain type

Chart, box and whisker chart

Description automatically generated

Representation of Cholesterol levelChart, scatter chart

Description automatically generated

Corellation plotting Chart, bubble chart

Description automatically generated

# Business Questions:

## What age groups are more vulnerable to heart disease?

Based on the data above, there is no specific age group that can be considered to be vulnerable to age diseases.

## What are the cholesterol levels that contribute toward heart diseases?

Cholesterol level between 200 and 250 mg/dl seems to contribute more towards a heart disease

## Which sex is more at risk for heart diseases?

Based on the ratio women are more at risk than compared to men

## Predict heart disease with Logistic Regression and Linear Regression

Classification uses logistic regression

Data division into train and test set fitting a full logistic model for classification

Text

Description automatically generated

Among the full.mod and step.model the second one is best in terms of accuracy.

Logistic regression diagnostics:

#the graph shows that linearity is maintained for continuous variable (not discrete variable like,ca,cp,sex,exang,etc)

Graphical user interface

Description automatically generated

The plot of cook\_sd shows that 184 index haves higher cook sd

Timeline

Description automatically generated with low confidence

Checking for the standard residuals error

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

The above model is a final model which gives us an accuracy of 84.74%

# RCODE

library(ggplot2)

library(caret)

library(GGally)

library(ggthemes)

library(broom)

library(dplyr)

library(bindrcpp)

library(caTools)

library(rattle)

library(RColorBrewer)

library(nnet)

library(rpart.plot)

library(ggplot2)

library(`c`ar)

library(dplyr)

library(lattice)

library(tidyr)

library(caret)

library(MASS)

library(broom)

library(ROCR)

library(corrplot)

library(ggcorrplot)

cleve <- read.csv('C:/r-final-data/processed-cleavland.csv', na = "?", stringsAsFactors = FALSE, header = FALSE)

hung <- read.csv('C:/r-final-data/processed-hungarian.csv', na = "?", stringsAsFactors = FALSE, header = FALSE)

swiss <- read.csv('C:/r-final-data/processed-switzerland.csv', na = "?", stringsAsFactors = FALSE, header = FALSE)

va <- read.csv('C:/r-final-data/processed-va.csv', na = "?", stringsAsFactors = FALSE, header = FALSE)

h <- rbind(cleve, hung, swiss, va)

# Data Cleaning and Transformation

# With str and summarizeColumns (Table 1), we noticed the following anomalies:

#

# The target feature, target had a cardinality of 5, which should be 2 since target was desiganted as the binary target feature.

# Ten of the 14 features contained missing values. Notably, the features Slope, CA and Thal had 309, 611 and 486 missing values, respectively.

# trestbps (resting blood pressure) and chol (serum cholestrol) contained several data entries with values of 0 which are not possible for these diagnostic tests.

names(h) <- c('age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target')

str(h)

summarizeColumns(h) %>% knitr::kable( caption = 'Feature Summary before Data Preprocessing')

# Fourthly, in the target column clinicians had graded patients as either having no heart disease (value of 0) or displaying various degrees of heart disease (values 1 to 4).

# We chose to group the data into 2 categories of ‘no heart disease’ (value of 0) and ‘displaying heart disease’ (value of 1) so it became binary.

#

# It was noted that a higher proportion of people were diagnosed with heart disease (Table2). Therefore, we may require additional parameter-tuning in building models to cater for such an unbalanced class.

h$target[h$target == 2] <- 1

h$target[h$target == 3] <- 1

h$target[h$target == 4] <- 1

h <- h[!is.na(h$trestbps),]

h <- h[!is.na(h$chol),]

h <- h[!is.na(h$restecg),]

h <- h[!is.na(h$oldpeak),]

h <- h[!is.na(h$fbs),]

h <- [!is.na(h$thal),]

h <- [!is.na(h$slope,)]

sapply(h[sapply(h, is.character)], table)

h[, sapply(h, is.character)] <- lapply( h[, sapply(h, is.character )], factor)

summarizeColumns(h) %>% kable( caption = 'Feature Summary before Data Preprocessing' )

heart\_data2 <- na.omit(h)

h$sex<-as.factor(h$sex)

h$cp<-as.factor(h$cp)

h$fbs<-as.factor(h$fbs)

h$exang<-as.factor(h$exang)

h$restecg<-as.factor(h$restecg)

h$slope<-as.factor(h$slope)

h$thal<-as.factor(h$thal)

h$target<-as.factor(h$target)

str(h)

levels(h$sex)[levels(h$sex)==0] <- "Female"

levels(h$sex)[levels(h$sex)==1] <- "Male"

levels(h$fbs)[levels(h$fbs)==0] <- "Fasting Blood Sugar <= 120"

levels(h$fbs)[levels(h$fbs)==1] <- "Fasting Blood Sugar > 120"

levels(h$thal)[levels(h$thal)==3] <- "Normal"

levels(h$thal)[levels(h$thal)==6] <- "Fixed Defect"

levels(h$thal)[levels(h$thal)==7] <- "Reversible Defect Thalassemia"

levels(h$target)[levels(h$target)==0] <- "Heart Disease"

levels(h$target)[levels(h$target)==1] <- "Healthy"

levels(h$exang)[levels(h$exang)==1] <- "Exercise Induced Angina"

levels(h$exang)[levels(h$exang)==0] <- "No Exercise Induced Angina"

levels(h$cp)[levels(h$cp)==1] <- "Typical Angina"

levels(h$cp)[levels(h$cp)==2] <- "Atypical Angina"

levels(h$cp)[levels(h$cp)==3] <- "Non-Angina Pain"

levels(h$cp)[levels(h$cp)==4] <- "asymptomatic"

levels(h$restecg)[levels(h$restecg)==0] <- "Rest ECG 0"

levels(h$restecg)[levels(h$restecg)==1] <- "Rest ECG 1"

levels(h$restecg)[levels(h$restecg)==2] <- "Rest ECG 2"

levels(h$slope)[levels(h$slope)==1] <- "Upsloaping"

levels(h$slope)[levels(h$slope)==2] <- "Flat"

levels(h$slope)[levels(h$slope)==3] <- "Downsloping"

sum(is.na(h))

summary(h)

graphColor <- c("firebrick2","springgreen2")

#Number of observations: Healthy and Heart Disease cases

ggplot(h,aes(target, fill=target)) +

geom\_bar(stat="count") +

theme\_economist() +

scale\_fill\_manual(values=graphColor)

#Heart Disease is uniformly spread out across Age

ggplot(h,aes(age, fill=target)) +

geom\_histogram(aes(y=..density..),breaks=seq(0, 80, by=1), color="grey17") +

geom\_density(alpha=.1, fill="black")+

facet\_wrap(~target, ncol=1,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor)+

xlab("Age") +

ylab("Density / Count") +

ggtitle("Age Histogram")

#No major difference in Rest ECG for Healthy and Heart Disease patients

ggplot(h,aes(trestbps, fill=target)) +

geom\_histogram(aes(y=..density..),breaks=seq(90, 200, by=10), color="grey17") +

geom\_density(alpha=.1, fill="black")+

facet\_wrap(~target, ncol=1,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor) +

xlab("Resting Blood Pressure (in mm Hg on admission to the hospital") +

ylab("Density / Count") +

ggtitle("Rest ECG Histogram")

#More Heart Disease patients seem to have between 200 and 250 mg/dl

ggplot(h,aes(chol, fill=target)) +

geom\_histogram(breaks=seq(100, 600, by=25), color="grey17") +

geom\_density(alpha=.1, fill="black")+

facet\_wrap(~target, ncol=1,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor) +

xlab("Serum Cholestoral in mg/dl") +

ylab("Density / Count") +

ggtitle("Cholestoral Histogram")

#Heart Disease patients have higher maximum heart rate than healthy patients

ggplot(h,aes(thalach, fill=target)) +

geom\_histogram(aes(y=..density..),breaks=seq(70, 205, by=10), color="grey17") +

geom\_density(alpha=.1, fill="black")+

facet\_wrap(~target, ncol=1,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor)+

xlab("Maximum Heart Rate Achieved") +

ylab("Density / Count") +

ggtitle("Max Heart Rate Histogram")

#More Heart Disease patients have ST depression of 0.1

ggplot(h,aes(oldpeak, fill=target)) +

geom\_histogram(aes(y=..density..),breaks=seq(0, 7, by=0.1), color="grey17") +

geom\_density(alpha=.1, fill="black")+

facet\_wrap(~target, ncol=1,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor) +

ggtitle("ST Depression Histogram") +

xlab("ST Depression Induced by Exercise Relative to Rest") +

ylab("Density / Count")

#Almost all of the patients who have Heart Disease have 0 major vessels as observed by Fluroscopy

ggplot(h,aes(ca, fill=target)) +

geom\_histogram(aes(y=..density..),breaks=seq(0, 5, by=1), color="grey17") +

geom\_density(alpha=.1, fill="black")+

facet\_wrap(~target, ncol=1,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor) +

ggtitle("No. Major Vessels Histogram") +

xlab("Number of Major Vessels (0-3) Colored by Flourosopy") +

ylab("Density / Count")

#More females have Heart Disease

ggplot(h,aes(target, fill=target)) +

geom\_bar(stat="count") +

facet\_wrap(~sex, ncol=2,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor)

#More Heart Disease patients have chest pain type 1 or 2

ggplot(h,aes(target, fill=target)) +

geom\_bar(stat="count") +

facet\_wrap(~cp, ncol=2,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor)

#No difference in fasting blood sugar

ggplot(h,aes(target, fill=target)) +

geom\_bar(stat="count") +

facet\_wrap(~fbs, ncol=2,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor)

#Patients with Rest ECG 1 have more Heart Diseases

ggplot(h,aes(target, fill=target)) +

geom\_bar(stat="count") +

facet\_wrap(~restecg, ncol=3,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor)

#Patients with no exercise induced angina have more Heart Disease

ggplot(h,aes(target, fill=target)) +

geom\_bar(stat="count") +

facet\_wrap(~exang, ncol=1,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor)

#Peak excercise ST Slope 2(Flat) have more Heart Disease

ggplot(na.omit(h),aes(target, fill=target)) +

geom\_bar(stat="count") +

facet\_wrap(~slope, ncol=1,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor)

#Fixed defect thalasemia has more Heart Disease

ggplot(na.omit(h),aes(target, fill=target)) +

geom\_bar(stat="count") +

facet\_wrap(~thal, ncol=2,scale="fixed") +

theme\_economist() +

scale\_fill\_manual(values=graphColor)

ggplot(h, aes(x=sex,y=chol))+

geom\_boxplot(fill="#D55E00")+

xlab("Sex")+

ylab("Chol")+

facet\_grid(~cp)

ggplot(h, aes(x=age,y=chol,color=sex, size=chol))+

geom\_point(alpha=0.7)+xlab("Age") +

ylab("Cholestoral")+

guides(fill = guide\_legend(title = "Gender"))

cor\_heart <- cor(heart\_data2[,10:14])

cor\_heart

corrplot(cor\_heart, method = "ellipse", type="upper",)

#classification :use logistic regression

#data division into train and test set

set.seed(123)

train.index <- heart\_data2$target %>% createDataPartition(p =0.8,list =F)

train.data <- heart\_data2[train.index,]

test.data <- heart\_data2[-train.index,]

#fitting a full logistic model for classification

full.mod <- glm(target~.,data =train.data,family =binomial)

summary(full.mod)

#checking the model accuracy

prob <- full.mod %>% predict(test.data,type ="response")

predicted.class1 <- ifelse(prob>0.5,1,0)

mean(predicted.class1==test.data$target)

#accuracy =0.86

#stepwise logistic regression in R

step.model <- full.mod %>% stepAIC(trace = F)

summary(step.model)

prob.step <- step.model %>% predict(test.data,type ="response")

predicted.class2 <- ifelse(prob.step>0.5,1,0)

mean(predicted.class2==test.data$target)

#accuracy =0.83

#among the full.mod and step.model the secand one is best interms of accuracy.

#logistic regression diognostic

#taking only quantitative variable

model\_check <- glm(target~.,data =heart\_data2,family = binomial)

prob.check <- predict(model\_check,type ="response")

my\_data <- heart\_data2 %>% select\_if(is.numeric)

predictors <- colnames(my\_data)

my\_data <- my\_data%>% mutate(logit = log(prob.check/(1-prob.check))) %>%

gather(key = "predictors",value = "predicted.value",-logit)

#plotting the graph for cheking the linearity

ggplot(my\_data,aes(x =logit,y =predicted.value)) + geom\_point() + geom\_smooth(method ="loess") + theme\_classic() + theme\_bw()+facet\_wrap(~predictors,scale ="free\_y")

#the graph shows that linearity is maintained for continous variable (not discrete variable like,ca,cp,sex,exang,etc)

plot(step.model,which=4,id.n =3)

#the plot of cook\_sd shows that 102 index haves higher cook\_sd

#checking for the standarg residuals error

model.data <- augment(step.model) %>% mutate(index =1:n())

model.data %>% top\_n(3,.cooksd)

#in these case the standard residuals error<3 so that can't be considered as a influential point

#checking for multi-collinearity

car::vif(step.model)

#no multi-collinearity presents .

#choosing the best cut-off probabillity value to the model

res <- predict(step.model,type ="response")

ROCR\_Pred <- prediction(res,train.data$target)

ROCR\_perf <- performance(ROCR\_Pred,"tpr","fpr")

plot(ROCR\_perf,colorize=T,print.cutoffs.at =seq(0.1,by =0.1))

#from the graph the cut-off value = 0.6.

#here is the final model

final.model <- glm(target~.,data =train.data,family =binomial) %>%stepAIC(trash =FALSE)

prob.final <- predict(final.model,test.data,type ="response")

predicted.class\_final <- ifelse(prob.final>0.6,1,0)

mean(predicted.class\_final==test.data$target)

#the model gives a accuracy of 84.74% and used for prediction