Health Disease Prediction

Azeez O. Baki

25/12/2021

## Setting Working Directory, Importing tidyverse library and Loading Data

## Exploring the Data

glimpse(data)

## Rows: 303  
## Columns: 14  
## $ ï..age <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5~  
## $ sex <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1~  
## $ cp <int> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0~  
## $ trestbps <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1~  
## $ chol <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2~  
## $ fbs <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~  
## $ restecg <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1~  
## $ thalach <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1~  
## $ exang <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~  
## $ oldpeak <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0~  
## $ slope <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1~  
## $ ca <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0~  
## $ thal <int> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~  
## $ target <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~

str(data)

## 'data.frame': 303 obs. of 14 variables:  
## $ ï..age : int 63 37 41 56 57 57 56 44 52 57 ...  
## $ sex : int 1 1 0 1 0 1 0 1 1 1 ...  
## $ cp : int 3 2 1 1 0 0 1 1 2 2 ...  
## $ trestbps: int 145 130 130 120 120 140 140 120 172 150 ...  
## $ chol : int 233 250 204 236 354 192 294 263 199 168 ...  
## $ fbs : int 1 0 0 0 0 0 0 0 1 0 ...  
## $ restecg : int 0 1 0 1 1 1 0 1 1 1 ...  
## $ thalach : int 150 187 172 178 163 148 153 173 162 174 ...  
## $ exang : int 0 0 0 0 1 0 0 0 0 0 ...  
## $ oldpeak : num 2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...  
## $ slope : int 0 0 2 2 2 1 1 2 2 2 ...  
## $ ca : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ thal : int 1 2 2 2 2 1 2 3 3 2 ...  
## $ target : int 1 1 1 1 1 1 1 1 1 1 ...

## Checking for Missing values

sum(is.na(data))

## [1] 0

sum(is.null(data))

## [1] 0

## Number of Rows and columns

nrow(data)

## [1] 303

ncol(data)

## [1] 14

colnames(data)

## [1] "ï..age" "sex" "cp" "trestbps" "chol" "fbs"   
## [7] "restecg" "thalach" "exang" "oldpeak" "slope" "ca"   
## [13] "thal" "target"

## Summary of the Data

summary(data)

## ï..age sex cp trestbps   
## Min. :29.00 Min. :0.0000 Min. :0.000 Min. : 94.0   
## 1st Qu.:47.50 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:120.0   
## Median :55.00 Median :1.0000 Median :1.000 Median :130.0   
## Mean :54.37 Mean :0.6832 Mean :0.967 Mean :131.6   
## 3rd Qu.:61.00 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:140.0   
## Max. :77.00 Max. :1.0000 Max. :3.000 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.:133.5   
## Median :240.0 Median :0.0000 Median :1.0000 Median :153.0   
## Mean :246.3 Mean :0.1485 Mean :0.5281 Mean :149.6   
## 3rd Qu.:274.5 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:166.0   
## Max. :564.0 Max. :1.0000 Max. :2.0000 Max. :202.0   
## exang oldpeak slope ca   
## Min. :0.0000 Min. :0.00 Min. :0.000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.00 1st Qu.:1.000 1st Qu.:0.0000   
## Median :0.0000 Median :0.80 Median :1.000 Median :0.0000   
## Mean :0.3267 Mean :1.04 Mean :1.399 Mean :0.7294   
## 3rd Qu.:1.0000 3rd Qu.:1.60 3rd Qu.:2.000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :6.20 Max. :2.000 Max. :4.0000   
## thal target   
## 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

## Data Transformation

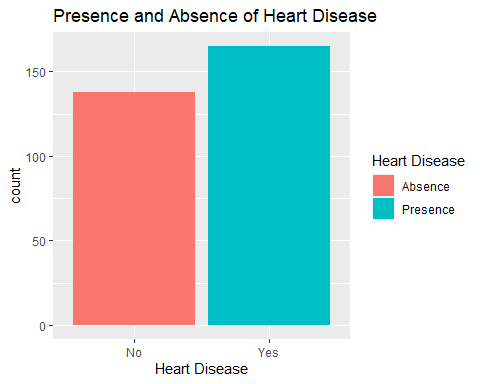
This is done for easy visualization of the data when needed. Although, it is not necessary to do.

data2 <- data %>%   
 mutate(sex = if\_else(sex==1, "male", "female"),  
 fbs = if\_else(fbs==1, ">120", "<=120"),  
 exang = if\_else(exang==1,"yes", "no"),  
 cp = if\_else(cp==1,"Atypical Angina",  
 if\_else(cp==2,"Non-Anginal pain", "Asymptomatic")),  
 restecg = if\_else(restecg==0,"Normal",  
 if\_else(restecg==1, "Abnormality", "Probable or Definite")),  
 slope = as.factor(slope),  
 ca = as.factor(ca),  
 thal =as.factor(thal),  
 target = if\_else(target== 1, "Yes", "No")  
 ) %>%   
 mutate\_if(is.character,as.factor) %>%   
 dplyr::select(target, sex, fbs, exang, cp, restecg, slope, ca, thal, everything())

## Data Visualization

To check what is going on with the data. we will be using ggplot to plot bar chart

ggplot(data2, aes(x=target, fill=target))+  
 geom\_bar()+  
 xlab("Heart Disease")+  
 ylab("count")+  
 ggtitle("Presence and Absence of Heart Disease")+  
 scale\_fill\_discrete(name = "Heart Disease", labels = c("Absence", "Presence"))



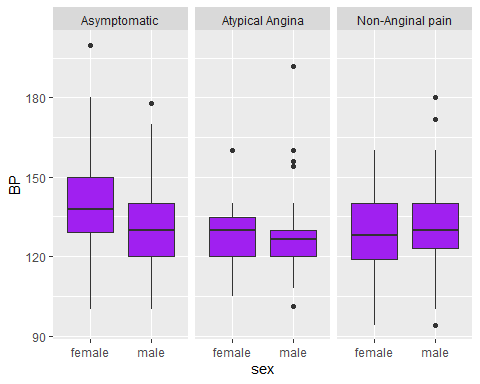
Since the plot above shows high number of people with heart disease, we can check the proportions.

prop.table(table(data2$target))

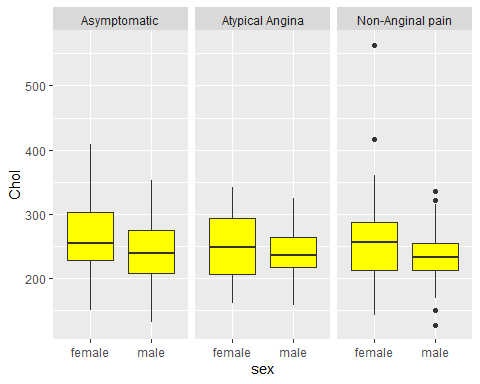
##   
## No Yes   
## 0.4554455 0.5445545

## Comparing blood pressure across the chest pain

data2 %>%   
 ggplot(aes(x=sex, y=trestbps))+  
 geom\_boxplot(fill = "purple")+  
 xlab("sex")+  
 ylab("BP")+  
 facet\_grid(~cp)



data2 %>%   
 ggplot(aes(x=sex, y=chol))+  
 geom\_boxplot(fill = "yellow")+  
 xlab("sex")+  
 ylab("Chol")+  
 facet\_grid(~cp)



## Correlation

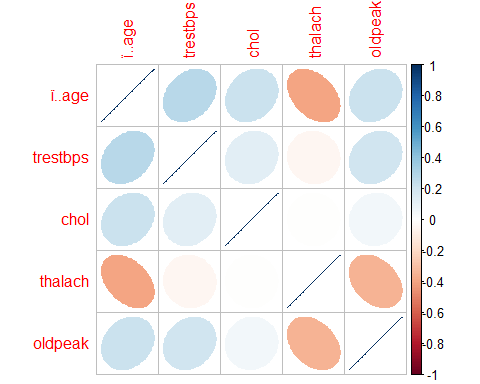
library(corrplot)

## corrplot 0.90 loaded

cor\_heart = cor(data2[, 10:14], method = "pearson")  
cor\_heart

## ï..age trestbps chol thalach oldpeak  
## ï..age 1.0000000 0.27935091 0.213677957 -0.398521938 0.21001257  
## trestbps 0.2793509 1.00000000 0.123174207 -0.046697728 0.19321647  
## chol 0.2136780 0.12317421 1.000000000 -0.009939839 0.05395192  
## thalach -0.3985219 -0.04669773 -0.009939839 1.000000000 -0.34418695  
## oldpeak 0.2100126 0.19321647 0.053951920 -0.344186948 1.00000000

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

 ## Building Model

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

heart\_tr = trainControl(method = "repeatedcv", number = 10, repeats = 5)  
model\_rf = train(target~.,data2, method="rf", trControl = heart\_tr)  
model\_rf

## Random Forest   
##   
## 303 samples  
## 13 predictor  
## 2 classes: 'No', 'Yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 5 times)   
## Summary of sample sizes: 273, 272, 273, 272, 273, 272, ...   
## Resampling results across tuning parameters:  
##   
## mtry Accuracy Kappa   
## 2 0.8156085 0.6280794  
## 11 0.7799881 0.5567421  
## 21 0.7699407 0.5359248  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was mtry = 2.

## To Predict the heart disease from the model using the data2

h\_disease= predict(model\_rf, data2)   
confusionMatrix(data2$target, h\_disease)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 130 8  
## Yes 3 162  
##   
## Accuracy : 0.9637   
## 95% CI : (0.936, 0.9817)  
## No Information Rate : 0.5611   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.9266   
##   
## Mcnemar's Test P-Value : 0.2278   
##   
## Sensitivity : 0.9774   
## Specificity : 0.9529   
## Pos Pred Value : 0.9420   
## Neg Pred Value : 0.9818   
## Prevalence : 0.4389   
## Detection Rate : 0.4290   
## Detection Prevalence : 0.4554   
## Balanced Accuracy : 0.9652   
##   
## 'Positive' Class : No   
##