Prediction of Liver Disease using ML Algorithms

R Kannan (1948019)

04/03/2020

Introduction:

In this fast growing world, we see various changes in day-to-day life. The Population of the world is drastically increasing. India being one of the most densely populated country, we see many new health issues coming up. One such health issue is “The Liver Disease”. Through our Analysis and Interpretation, we would be trying to predict/classify a person if he is suffering from a liver disease or not.

liv=read.csv(file.choose())  
head(liv)

## Age Gender Total\_Bilirubin Direct\_Bilirubin Alkaline\_Phosphotase  
## 1 65 Female 0.7 0.1 187  
## 2 62 Male 10.9 5.5 699  
## 3 62 Male 7.3 4.1 490  
## 4 58 Male 1.0 0.4 182  
## 5 72 Male 3.9 2.0 195  
## 6 46 Male 1.8 0.7 208  
## Alamine\_Aminotransferase Aspartate\_Aminotransferase Total\_Protiens  
## 1 16 18 6.8  
## 2 64 100 7.5  
## 3 60 68 7.0  
## 4 14 20 6.8  
## 5 27 59 7.3  
## 6 19 14 7.6  
## Albumin Albumin\_and\_Globulin\_Ratio Dataset  
## 1 3.3 0.90 1  
## 2 3.2 0.74 1  
## 3 3.3 0.89 1  
## 4 3.4 1.00 1  
## 5 2.4 0.40 1  
## 6 4.4 1.30 1

tail(liv)

## Age Gender Total\_Bilirubin Direct\_Bilirubin Alkaline\_Phosphotase  
## 578 32 Male 12.7 8.4 190  
## 579 60 Male 0.5 0.1 500  
## 580 40 Male 0.6 0.1 98  
## 581 52 Male 0.8 0.2 245  
## 582 31 Male 1.3 0.5 184  
## 583 38 Male 1.0 0.3 216  
## Alamine\_Aminotransferase Aspartate\_Aminotransferase Total\_Protiens  
## 578 28 47 5.4  
## 579 20 34 5.9  
## 580 35 31 6.0  
## 581 48 49 6.4  
## 582 29 32 6.8  
## 583 21 24 7.3  
## Albumin Albumin\_and\_Globulin\_Ratio Dataset  
## 578 2.6 0.90 1  
## 579 1.6 0.37 2  
## 580 3.2 1.10 1  
## 581 3.2 1.00 1  
## 582 3.4 1.00 1  
## 583 4.4 1.50 2

Dataset:

This data set contains 416 liver patient records and 167 non liver patient records and 11 features (classes) collected from North East of Andhra Pradesh, India. The “Dataset” column is a class label used to divide groups into liver patient (liver disease) or not (no disease). This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age “90”.

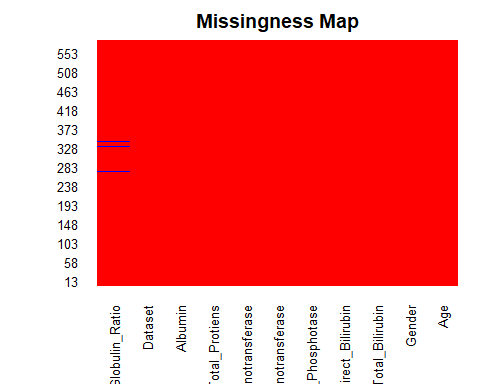
library(Amelia)

## Warning: package 'Amelia' was built under R version 3.6.3

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.7.6, built: 2019-11-24)  
## ## Copyright (C) 2005-2020 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

missmap(liv, col=c("blue", "red"), legend=FALSE)



liver=na.omit(liv)

**Interpretation:**

From the above plot it is inferred that there are four missing values in the dataset and it is treated by omitting the rows that contains missing values as the number of missing values in this dataset doesn’t impact a lot while building suitable models.

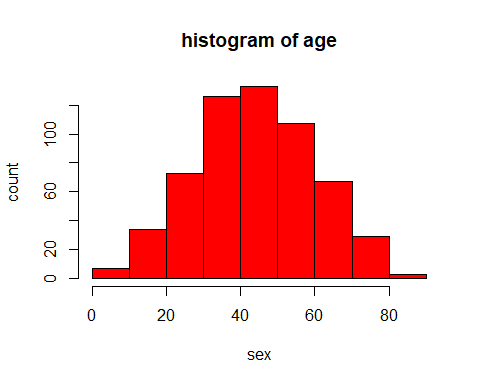
summary(liver)

## Age Gender Total\_Bilirubin Direct\_Bilirubin  
## Min. : 4.00 Female:140 Min. : 0.400 Min. : 0.100   
## 1st Qu.:33.00 Male :439 1st Qu.: 0.800 1st Qu.: 0.200   
## Median :45.00 Median : 1.000 Median : 0.300   
## Mean :44.78 Mean : 3.315 Mean : 1.494   
## 3rd Qu.:58.00 3rd Qu.: 2.600 3rd Qu.: 1.300   
## Max. :90.00 Max. :75.000 Max. :19.700   
## Alkaline\_Phosphotase Alamine\_Aminotransferase Aspartate\_Aminotransferase  
## Min. : 63.0 Min. : 10.00 Min. : 10.0   
## 1st Qu.: 175.5 1st Qu.: 23.00 1st Qu.: 25.0   
## Median : 208.0 Median : 35.00 Median : 42.0   
## Mean : 291.4 Mean : 81.13 Mean : 110.4   
## 3rd Qu.: 298.0 3rd Qu.: 61.00 3rd Qu.: 87.0   
## Max. :2110.0 Max. :2000.00 Max. :4929.0   
## Total\_Protiens Albumin Albumin\_and\_Globulin\_Ratio  
## Min. :2.700 Min. :0.900 Min. :0.3000   
## 1st Qu.:5.800 1st Qu.:2.600 1st Qu.:0.7000   
## Median :6.600 Median :3.100 Median :0.9300   
## Mean :6.482 Mean :3.139 Mean :0.9471   
## 3rd Qu.:7.200 3rd Qu.:3.800 3rd Qu.:1.1000   
## Max. :9.600 Max. :5.500 Max. :2.8000   
## Dataset   
## Min. :1.000   
## 1st Qu.:1.000   
## Median :1.000   
## Mean :1.285   
## 3rd Qu.:2.000   
## Max. :2.000

**Interpretation:**

The above is the summary of the model. There are two categorical variables. One is Age and the other is the target variable where 1 denotes presence of liver disease and 2 represents absence of liver disease.

hist(liver$Age,main="histogram of age",xlab="sex",ylab="count",col='red')



**Interpretation:**

The above histogram tells that most of the age category of the patients fall under 40-50 and the lowest being above 80. This indicates that there are more people affected by liver disease in the age category of 40-50 compared to other categories. And also the age data follows normal distribution therefore mean=median(approximately).

gender=table(liver$Gender)  
barplot(gender,main="Barplot of Gender",xlab = "Gender",ylab="count",col="skyblue")

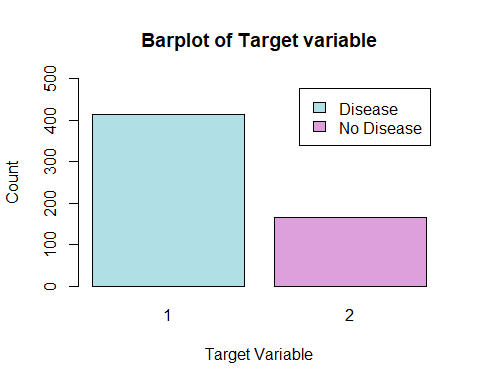
**Interpretation:**

The above barplot explains that most of the data records in this dataset belongs to male with male count being 439 and female count being 140.

target=table(liver$Dataset)  
target

##   
## 1 2   
## 414 165

barplot(target,main="Barplot of Target variable",xlab="Target Variable",ylab="Count",legend=c("Disease","No Disease"),col=c("powderblue","plum"),ylim = c(0,500))



**Interpretation:**

From the above barplot it is inferred that more than 400 people are affected by liver disease out of 579 people.

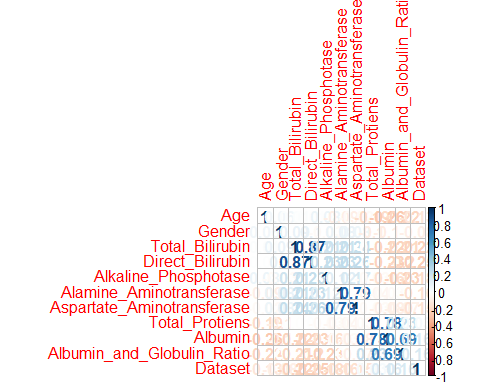
library(corrplot)

## Warning: package 'corrplot' was built under R version 3.6.2

## corrplot 0.84 loaded

liver[,2]= sapply(liver[,2],switch,"Male"=1,"Female"=2)

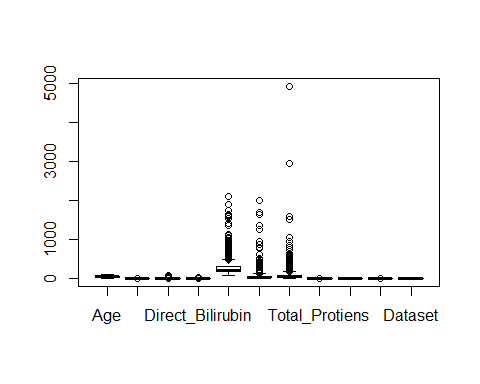
ger.  
## Consider using 'switch(as.character( \* ), ...)' instead.  
cor=cor(liver)  
corrplot(cor,method="number")



**Interpretation:**

The above is the correlation matrix of the dataset and it is evident that none of the features are highly correlated with the target variable with highest being Direct\_Bilirubin at 0.25.

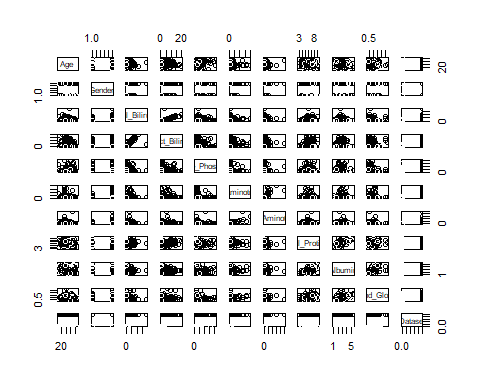
liver[,11]= sapply(liver[,11],switch,"1"=1,"2"=0)  
boxplot(liver)



**Interpretation:**

The above boxplot indicates that there are more number of outliers in the dataset especially in the features Alkaline\_Phosphotase, Alamaine\_Aminotransferse, Aspartate\_Aminotransferase. Outliers are not removed in this because a large portion of the dataset are outliers and treatment of this will impact in the model. Influence of outliers are checked after building the model, if found influencing then it will be treated.

pairs(liver, col=liver$Dataset)



**Interpretation:**

The above is scatterplot matrix of the dataset and it is evident that most of the features are not associated with the target variable.

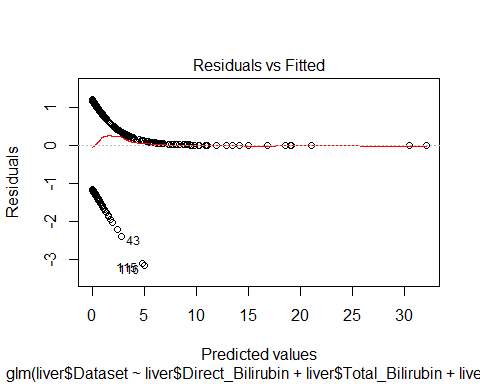
glm.fit <- glm(liver$Dataset ~ liver$Direct\_Bilirubin + liver$Total\_Bilirubin +liver$Alkaline\_Phosphotase +liver$Alamine\_Aminotransferase +liver$Aspartate\_Aminotransferase, family = binomial)

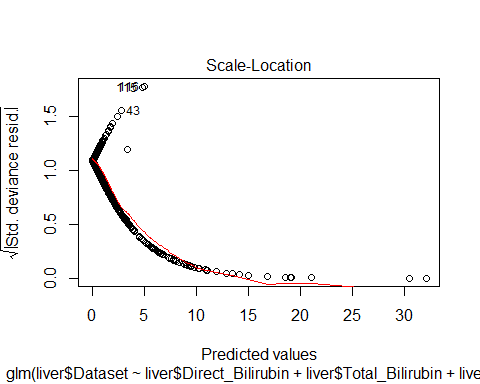
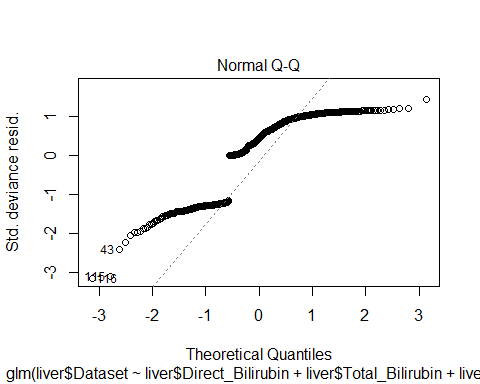
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

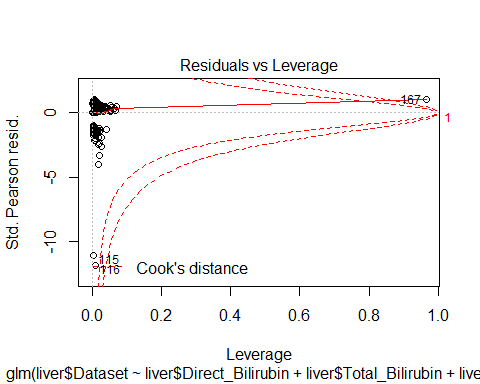
summary(glm.fit)

##   
## Call:  
## glm(formula = liver$Dataset ~ liver$Direct\_Bilirubin + liver$Total\_Bilirubin +   
## liver$Alkaline\_Phosphotase + liver$Alamine\_Aminotransferase +   
## liver$Aspartate\_Aminotransferase, family = binomial)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1435 -1.2269 0.4317 0.9558 1.2050   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.4780532 0.2229873 -2.144 0.0320 \*  
## liver$Direct\_Bilirubin 0.6084656 0.2412347 2.522 0.0117 \*  
## liver$Total\_Bilirubin 0.0078367 0.0791951 0.099 0.9212   
## liver$Alkaline\_Phosphotase 0.0014013 0.0007665 1.828 0.0675 .  
## liver$Alamine\_Aminotransferase 0.0082549 0.0046575 1.772 0.0763 .  
## liver$Aspartate\_Aminotransferase 0.0034909 0.0031800 1.098 0.2723   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 692.01 on 578 degrees of freedom  
## Residual deviance: 586.24 on 573 degrees of freedom  
## AIC: 598.24  
##   
## Number of Fisher Scoring iterations: 7

plot(glm.fit)







**Interpretation:**

A logistic regression is built for the above model and from the Z values and p-values it is evident that none of the coefficients are significant here. The AIC value indicates that it is an average model not the best one. From the residual vs fitted plot the points don’t form a straight line therefore not a good model and in normal Q-Q plot it is inferred that the data doesn’t follow normal distribution, and scale location plot indicates that the points are not evenly scattered therefore not a good model and the cook’s distance indicates that there is outlier influence in this model.

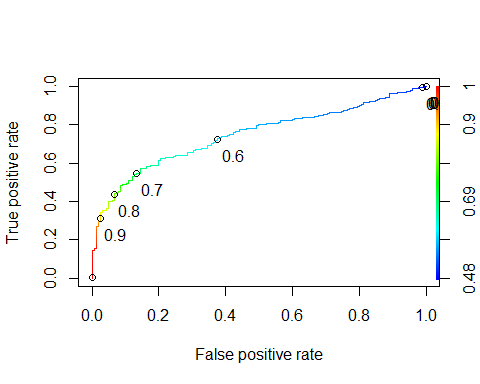
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

predictions= predict(glm.fit, type='response')  
ROCRpred= prediction(predictions, liver$Dataset)  
ROCRperf= performance(ROCRpred, measure ="tpr", x.measure ="fpr")  
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7), print.cutoffs.at = seq(0,1,0.1))



**Interpretation:**

The above ROC curve indicates that the logistic model built is not a good model as the curve is not close to 1. Therefore the accuracy of the model will be very less implying that the target variable can’t be predicted accurately. Therefore other techniques are used to increase the accuracy.

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

sample1 <- sample(2, nrow(liver),  
 replace = T,  
 prob = c(0.6,0.4))  
  
train <- liver[sample1==1,]  
test <- liver[sample1==2,]

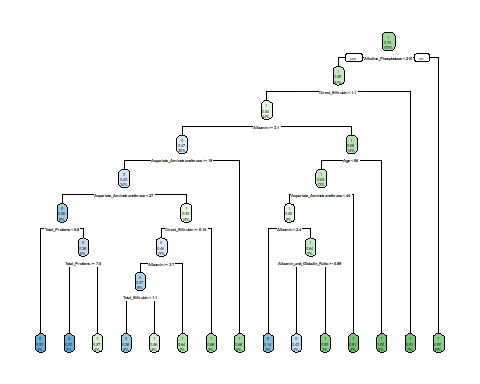
**Interpretation:**

We now split the data into training and testing data inorder to apply Decision Tree Algorithm.

library(rpart)  
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 3.6.2

fit <- rpart(Dataset~., data = train, method = 'class')  
rpart.plot(fit, extra = 106)



predict\_unseen <-predict(fit, test, type = 'class')

table\_mat <- table(test$Dataset, predict\_unseen)  
table\_mat

## predict\_unseen  
## 0 1  
## 0 22 57  
## 1 21 147

accuracy\_Test <- sum(diag(table\_mat)) / sum(table\_mat)

print(paste('Accuracy for test', accuracy\_Test))

## [1] "Accuracy for test 0.684210526315789"

**Interpretation:**

From the Decision Tree, We infer that all the samples in the test dataset haven’t been correctly classified and we’ve attained an accuracy of just 66.35% on the test data set which is very very minimal. Thus we infer that, we cannot correctly predict whether the person is suffering from a liver disease or not from our Decision Tree Model.

#Tuning the Model to Boost the Accuracy  
accuracy\_tune <- function(fit) {  
 predict\_unseen <- predict(fit, test, type = 'class')  
 table\_mat <- table(test$Dataset, predict\_unseen)  
 accuracy\_Test <- sum(diag(table\_mat)) / sum(table\_mat)  
 accuracy\_Test  
}

control <- rpart.control(minsplit = 4,  
 minbucket = round(5 / 3),  
 maxdepth = 3,  
 cp = 0)  
tune\_fit <- rpart(Dataset~., data = train, method = 'class', control = control)  
print(paste('Accuracy after Tuning', accuracy\_tune(tune\_fit)))

## [1] "Accuracy after Tuning 0.651821862348178"

**Interpretation:**

After tuning the Decision Tree, We have obtained an increased level of Accuracy which is 68.66%. Even though there is a rise in the accuracy level, we cannot consider it to be a good model as it has only 68% chance to predict the right outcome. We cannot consider a model to be a Good model unless and until it shows a minimum accuracy of 95%. If we predict it wrongly, the patient’s situation might get worse or may even end up to be life taking.

**Conclusion:**

I have done Visualization and Data Analysis on the data which is about “Liver Disease” i.e. it deals with the Indian People and classifies them as People with Liver Disease and People without Liver Disease. Also, I have fit Logistic Regression Model and Decision Tree Model.

We have inferred that our prediction accuracy rate is not very satisfying and has to be improved dramatically. This can be done by performing multiple training and testing for the dataset and also by adding more and more relevant features (columns) which could lead to better results.

I would like to caution you to take care of your liver as it is one of the most important organ in human body. Liver damage is very harmful to oneself and can be life threatening. **Be Safe, Be Healthy** and do avoid excess consumption of Alcohol cause it has a severe effect on Liver.

**THANK YOU**