Indian Liver Patient Records

Wilson Moreno

1/8/2020

# Introduction

The present work wishes to predict if a patient in India is going to suffer from liver disease, by means of some indicators, for example: Age, Gender, Alkaline\_Phosphography, etc. The result of a categorical variable. Different types of approaches are used such as: using all variables, through a correlation matrix and by specific variables.

# Methods/Analysis

### Step 0: require package

if (!require(package)) install.packages('psych', repos = "http://cran.us.r-project.org")  
if (!require(package)) install.packages('knitr', repos = "http://cran.us.r-project.org")  
if (!require(package)) install.packages('ggplot2', repos = "http://cran.us.r-project.org")  
library(knitr)  
library(ggplot2)  
library(psych)  
library(caret)

### **Step 1:** Load the data base

database <- read.csv("Data/indian\_liver\_patient.csv")

### **Step 2:** Exploratory Data Analysis

#### Summay Statistics

The following table shows the descriptive statistics for all the variables in the database.

round(data.frame( describeBy(database, digits= 2)),1)

## Warning in describeBy(database, digits = 2): no grouping variable requested

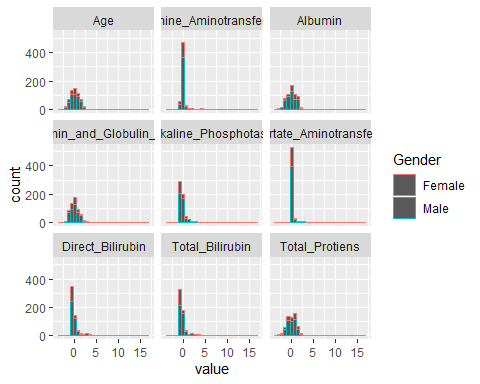
## vars n mean sd median trimmed mad min max  
## Age 1 583 44.7 16.2 45.0 44.8 17.8 4.0 90.0  
## Gender\* 2 583 1.8 0.4 2.0 1.8 0.0 1.0 2.0  
## Total\_Bilirubin 3 583 3.3 6.2 1.0 1.7 0.4 0.4 75.0  
## Direct\_Bilirubin 4 583 1.5 2.8 0.3 0.7 0.3 0.1 19.7  
## Alkaline\_Phosphotase 5 583 290.6 242.9 208.0 238.4 74.1 63.0 2110.0  
## Alamine\_Aminotransferase 6 583 80.7 182.6 35.0 43.9 22.2 10.0 2000.0  
## Aspartate\_Aminotransferase 7 583 109.9 288.9 42.0 56.8 31.1 10.0 4929.0  
## Total\_Protiens 8 583 6.5 1.1 6.6 6.5 1.0 2.7 9.6  
## Albumin 9 583 3.1 0.8 3.1 3.1 0.9 0.9 5.5  
## Albumin\_and\_Globulin\_Ratio 10 579 0.9 0.3 0.9 0.9 0.3 0.3 2.8  
## Dataset 11 583 1.3 0.5 1.0 1.2 0.0 1.0 2.0  
## range skew kurtosis se  
## Age 86.0 0.0 -0.6 0.7  
## Gender\* 1.0 -1.2 -0.6 0.0  
## Total\_Bilirubin 74.6 4.9 36.7 0.3  
## Direct\_Bilirubin 19.6 3.2 11.2 0.1  
## Alkaline\_Phosphotase 2047.0 3.7 17.5 10.1  
## Alamine\_Aminotransferase 1990.0 6.5 50.0 7.6  
## Aspartate\_Aminotransferase 4919.0 10.5 149.1 12.0  
## Total\_Protiens 6.9 -0.3 0.2 0.0  
## Albumin 4.6 0.0 -0.4 0.0  
## Albumin\_and\_Globulin\_Ratio 2.5 1.0 3.2 0.0  
## Dataset 1.0 0.9 -1.1 0.0

# tmp <- describeBy(database,  
# group = database$Gender,  
# digits= 1)

#### Visualization

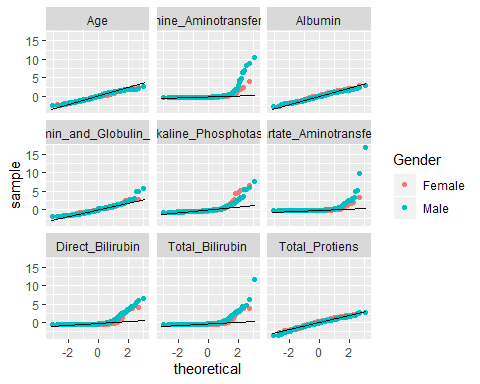
As can be seen in the visual analysis of the data, they were divided into 2 groups by gender, in all the graphs the rates are much higher in women than in men, and tend to follow the same distribution. Variables: Age, Albumin, Albumin\_and\_Globulin\_Ratio and Total\_Protiens, are suspected to follow a normal distribution.

scale\_database <- database %>% select(-Gender,-Dataset) %>% scale() %>% as.data.frame() %>%   
 cbind(Gender = database$Gender)  
  
database.gathered <- scale\_database %>% as.data.frame() %>%   
 gather(key = "variable", value = "value", - Gender)  
  
ggplot(data = database.gathered , mapping = aes(x = value, color = Gender)) +  
 geom\_histogram() +  
 facet\_wrap(facets = vars(variable ))



This graph certifies the suspicion that the aforementioned variables follow a normal distribution.

ggplot(data = database.gathered , mapping = aes(sample = value, color = Gender)) +  
 stat\_qq() + stat\_qq\_line(color = "black") +  
 facet\_wrap(facets = vars(variable ))



### **Step 3:** Split the database in training and testing

#### Split the database in training and testing

set.seed(755)  
test\_index <- createDataPartition(y = database$Dataset, times = 1,  
 p = 0.2, list = FALSE)  
  
train\_set <- database[-test\_index,]  
test\_set <- database[test\_index,]  
  
  
RMSE <- function(true\_ratings, predicted\_ratings){  
 sqrt(mean((true\_ratings - predicted\_ratings)^2))  
}

How will the logistic regression algorithm be used **glm ()** you have to modify the response variable **Data set** to values of 0 if the person does not suffer the disease and 1 if a person suffers the disease and convert the gender variable into a factor.

# Modify the response variable.  
train\_set$Dataset[train\_set$Dataset == 1] = 1  
train\_set$Dataset[train\_set$Dataset == 2] = 0  
test\_set$Dataset[test\_set$Dataset == 1] = 1  
test\_set$Dataset[test\_set$Dataset == 2] = 0  
  
# convert the variable into factor  
train\_set$Gender <- as.factor(train\_set$Gender)  
test\_set$Gender <- as.factor(test\_set$Gender)

### **Step 4:** Choose the Model and train the model with the training base

For the following approach some models are used, which are a function of the predictive variables to choose.

#### Model 1: Using all the variables

mol\_1 <- glm(Dataset ~. , data = train\_set, family = binomial() )

#### Model 2: Variables + correlation

round(cor(database[,-2]),2)

## Age Total\_Bilirubin Direct\_Bilirubin  
## Age 1.00 0.01 0.01  
## Total\_Bilirubin 0.01 1.00 0.87  
## Direct\_Bilirubin 0.01 0.87 1.00  
## Alkaline\_Phosphotase 0.08 0.21 0.23  
## Alamine\_Aminotransferase -0.09 0.21 0.23  
## Aspartate\_Aminotransferase -0.02 0.24 0.26  
## Total\_Protiens -0.19 -0.01 0.00  
## Albumin -0.27 -0.22 -0.23  
## Albumin\_and\_Globulin\_Ratio NA NA NA  
## Dataset -0.14 -0.22 -0.25  
## Alkaline\_Phosphotase Alamine\_Aminotransferase  
## Age 0.08 -0.09  
## Total\_Bilirubin 0.21 0.21  
## Direct\_Bilirubin 0.23 0.23  
## Alkaline\_Phosphotase 1.00 0.13  
## Alamine\_Aminotransferase 0.13 1.00  
## Aspartate\_Aminotransferase 0.17 0.79  
## Total\_Protiens -0.03 -0.04  
## Albumin -0.17 -0.03  
## Albumin\_and\_Globulin\_Ratio NA NA  
## Dataset -0.18 -0.16  
## Aspartate\_Aminotransferase Total\_Protiens Albumin  
## Age -0.02 -0.19 -0.27  
## Total\_Bilirubin 0.24 -0.01 -0.22  
## Direct\_Bilirubin 0.26 0.00 -0.23  
## Alkaline\_Phosphotase 0.17 -0.03 -0.17  
## Alamine\_Aminotransferase 0.79 -0.04 -0.03  
## Aspartate\_Aminotransferase 1.00 -0.03 -0.09  
## Total\_Protiens -0.03 1.00 0.78  
## Albumin -0.09 0.78 1.00  
## Albumin\_and\_Globulin\_Ratio NA NA NA  
## Dataset -0.15 0.04 0.16  
## Albumin\_and\_Globulin\_Ratio Dataset  
## Age NA -0.14  
## Total\_Bilirubin NA -0.22  
## Direct\_Bilirubin NA -0.25  
## Alkaline\_Phosphotase NA -0.18  
## Alamine\_Aminotransferase NA -0.16  
## Aspartate\_Aminotransferase NA -0.15  
## Total\_Protiens NA 0.04  
## Albumin NA 0.16  
## Albumin\_and\_Globulin\_Ratio 1 NA  
## Dataset NA 1.00

From the correlation matrix it can be observed that there is a high degree of correlation between the variables \*\* Total\_Bilirubin \*\* with \*\* Direct\_Bilirubin \*\* and \*\* Albumin \*\* with \*\* Total\_Protiens \*\*, therefore, it can be removed from the database.

train\_set\_mol\_2 = train\_set %>% select(-Total\_Bilirubin,-Total\_Protiens)  
mol\_2 <- glm(Dataset ~. , data = train\_set\_mol\_2, family = binomial())

#### Model 3: Significant variables

summary(mol\_1)

##   
## Call:  
## glm(formula = Dataset ~ ., family = binomial(), data = train\_set)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.1370 -1.0787 0.4107 0.9186 1.4787   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.3869084 1.5202083 -1.570 0.11639   
## Age 0.0193951 0.0069943 2.773 0.00555 \*\*  
## GenderMale -0.0193057 0.2551699 -0.076 0.93969   
## Total\_Bilirubin -0.2239224 0.4940652 -0.453 0.65039   
## Direct\_Bilirubin 0.9564189 0.9344332 1.024 0.30606   
## Alkaline\_Phosphotase 0.0006987 0.0008039 0.869 0.38481   
## Alamine\_Aminotransferase 0.0145580 0.0058760 2.478 0.01323 \*   
## Aspartate\_Aminotransferase 0.0008123 0.0035444 0.229 0.81874   
## Total\_Protiens 0.5299709 0.4289254 1.236 0.21662   
## Albumin -0.8757233 0.8457121 -1.035 0.30044   
## Albumin\_and\_Globulin\_Ratio 0.6625556 1.2938651 0.512 0.60860   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 556.42 on 461 degrees of freedom  
## Residual deviance: 459.51 on 451 degrees of freedom  
## (4 observations deleted due to missingness)  
## AIC: 481.51  
##   
## Number of Fisher Scoring iterations: 7

From the summary of model 1, where all the variables were used, it can be seen that the significant ones where their p-value is minus 0.05 are: \*\* Age \*\* and \*\* Alamine\_Aminotransferase \*\*. Therefore, only those variables are selected.

train\_set\_mol\_3 = train\_set %>% select(Age,Alamine\_Aminotransferase)  
mol\_3 <- glm(Dataset ~. , data = train\_set, family = binomial())

### **Step 5:** Predict the possible ratings for the test base

#### Model 1: Using all the variables

y\_hat\_1 <- predict(mol\_1,test\_set,type = "response")  
glm.pred\_1 <- ifelse(y\_hat\_1 > 0.5, "1", "0")  
accuracy <- data\_frame(method="Using all the variables",  
 Accuracy = mean(glm.pred\_1 == test\_set$Dataset))

y\_hat\_2 <- predict(mol\_2,test\_set,type = "response")  
glm.pred\_2 <- ifelse(y\_hat\_2 > 0.5, "1", "0")  
accuracy <- bind\_rows(accuracy,data\_frame(method="Variables + correlation",  
 Accuracy = mean(glm.pred\_2 == test\_set$Dataset)))

y\_hat\_3 <- predict(mol\_3,test\_set,type = "response")  
glm.pred\_3 <- ifelse(y\_hat\_3 > 0.5, "1", "0")  
accuracy <- bind\_rows(accuracy,data\_frame(method="Significant variables",  
 Accuracy = mean(glm.pred\_3 == test\_set$Dataset)))

### **Step 6:** Accuracy

accuracy

## # A tibble: 3 x 2  
## method Accuracy  
## <chr> <dbl>  
## 1 Using all the variables 0.769  
## 2 Variables + correlation 0.744  
## 3 Significant variables 0.769

# Results

As more relevant results in the exploratory analysis it can be observed that the distribution in men and women for the different variables is similar, this allows us to think that gender is not a relevant variable to take into account, that it can be justified because it does not It is significant in the analysis of the third model. On the basis of the second graph, it can also be observed that certain variables are distributed normally, which is good, if one wishes to make univariate predictions of them. After partitioning the data in training and testing, to be later modeled by means of different approaches, it is evident that through the precision that model 1 and model 3 are equal, take into account that model 3 It only consists of 2 variables (**Age** and **Alamine\_Aminotransferase**) to achieve this accuracy, therefore, they are the most important characteristics to know and predict if a patient will suffer from a liver problem.

# Conclusion

From the researched literature, decision trees or Vector Support Machine could be taken as models to achieve better prediction levels.