Indian Liver Patients

Octavio M.

11 June 2019

General Overview

This project is part of the Professional Certification of HarvardX: Data Science. The main objective of this to analize one data base and use some ML. First we are going to start with a short introducction, then the given

dataset will be prepared and get ready for a data analysis that will be carried out to acomplish the main goal and develop a machine learning (ML) algorithm that will help us to analize the Indian Liver Patients. After that the results will be explained and it will help to make some conclusions.

Introduction

This project will examine data from liver patients especially concentrating on the relations between a list of key liver indicators, age, gender and then try to use them to predict liver disease.

Here is important to know that if we detect **early signs of liver disease**, we can save a lot of lifes. We know that the models have limitations and some errors, but they help us a lot.

In this project, the possibility of find early signs of liver disease with the variables we said before can help to decrease costs and help to improve the quality of life.

Data

The Liver dataset is automatically downloaded

```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-proje
ct.org")
if(!require(caTools)) install.packages("caTools", repos = "http://cran.us.r-project.o
rg")
if(!require(pscl)) install.packages("pscl", repos = "http://cran.us.r-project.org")

liver_data <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/002
25/Indian%20Liver%20Patient%20Dataset%20(ILPD).csv",
    header = FALSE)
colnames(liver_data) <- c("Age", "Sex", "Tot_Bil", "Dir_Bil", "Alkphos", "Alamine",
    "Aspartate", "Tot_Prot", "Albumin", "A_G_Ratio", "Disease")
liver_data$Sex <- (ifelse(liver_data$Sex == "Male", "M", "F"))  #made shorter
liver_data$Disease <- as.numeric(ifelse(liver_data$Disease == 2, 0, 1))  #converted t
o zeros and ones
```

Analysis

Data

First of all we need to know a little bit of our data set. Down are the first rows of the *Liver Data* subset. The subset contain eleven variables **Age**, **Total_Bilirubin**, **Gender**, **Direct_Bilirubin**, **Alkaline_Phosphotase**,**Alamine_Aminotransferase**,**Aspartate_Aminotransferase**,**Total_Protiens**,**Albumin**, **Albumin_and_Globulin_Ratio** and **Dataset**. Each row represent a single patient.

```
##   Age Sex Tot_Bil Dir_Bil Alkphos Alamine Aspartate Tot_Prot Albumin
## 1  65   F     0.7     0.1     187      16        18      6.8     3.3
## 2  62   M    10.9     5.5     699      64       100      7.5     3.2
## 3  62   M     7.3     4.1     490      60        68      7.0     3.3
## 4  58   M     1.0     0.4     182      14        20      6.8     3.4
## 5  72   M     3.9     2.0     195      27        59      7.3     2.4
## 6  46   M     1.8     0.7     208      19        14      7.6     4.4
##   A_G_Ratio Disease
## 1      0.90       1
## 2      0.74       1
## 3      0.89       1
## 4      1.00       1
## 5      0.40       1
## 6      1.30       1
```

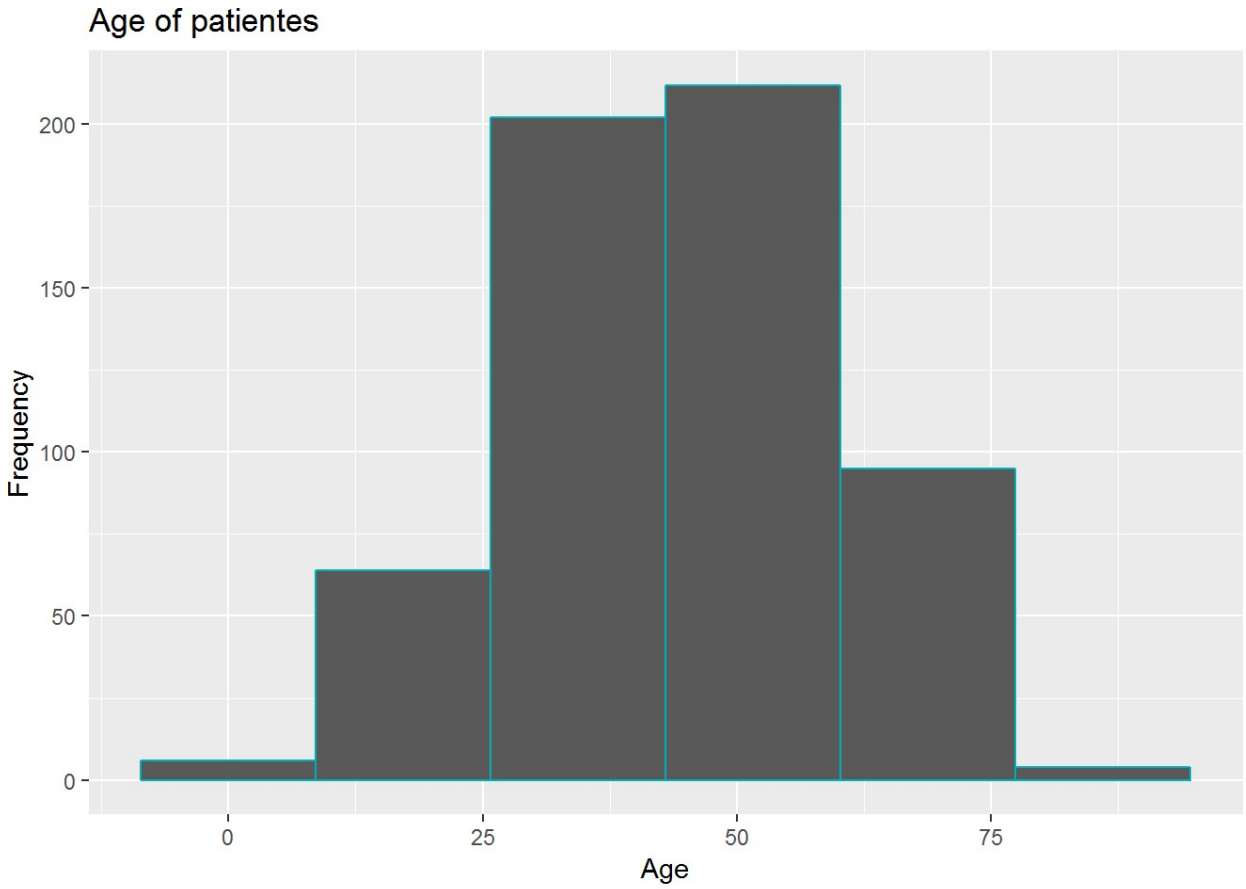
A summary of the Data:

```
##       Age            Sex               Tot_Bil          Dir_Bil       
##  Min.   : 4.00   Length:583         Min.   : 0.400   Min.   : 0.100  
##  1st Qu.:33.00   Class :character   1st Qu.: 0.800   1st Qu.: 0.200  
##  Median :45.00   Mode  :character   Median : 1.000   Median : 0.300  
##  Mean   :44.75                      Mean   : 3.299   Mean   : 1.486  
##  3rd Qu.:58.00                      3rd Qu.: 2.600   3rd Qu.: 1.300  
##  Max.   :90.00                      Max.   :75.000   Max.   :19.700  
##                                                                      
##     Alkphos          Alamine          Aspartate         Tot_Prot     
##  Min.   :  63.0   Min.   :  10.00   Min.   :  10.0   Min.   :2.700  
##  1st Qu.: 175.5   1st Qu.:  23.00   1st Qu.:  25.0   1st Qu.:5.800  
##  Median : 208.0   Median :  35.00   Median :  42.0   Median :6.600  
##  Mean   : 290.6   Mean   :  80.71   Mean   : 109.9   Mean   :6.483  
##  3rd Qu.: 298.0   3rd Qu.:  60.50   3rd Qu.:  87.0   3rd Qu.:7.200  
##  Max.   :2110.0   Max.   :2000.00   Max.   :4929.0   Max.   :9.600  
##                                                                     
##     Albumin        A_G_Ratio         Disease      
##  Min.   :0.900   Min.   :0.3000   Min.   :0.0000  
##  1st Qu.:2.600   1st Qu.:0.7000   1st Qu.:0.0000  
##  Median :3.100   Median :0.9300   Median :1.0000  
##  Mean   :3.142   Mean   :0.9471   Mean   :0.7136  
##  3rd Qu.:3.800   3rd Qu.:1.1000   3rd Qu.:1.0000  
##  Max.   :5.500   Max.   :2.8000   Max.   :1.0000  
##                  NA's   :4
```

Age

A big part of the patiens is **in the range of age of 25 and 62 years old**.

```
liver_data %>%
ggplot(aes(Age)) +
geom_histogram(bins = 6, color = "#00AFBB") +
xlab("Age") +
ylab("Frequency") +
ggtitle("Age of patientes")
```



As we can see below, there aro more male than female, in relative terms **77.8% of Male patients are diseased**

```
liver_data %>%
group_by(Sex,Disease)%>%
   summarise (n = n())
```

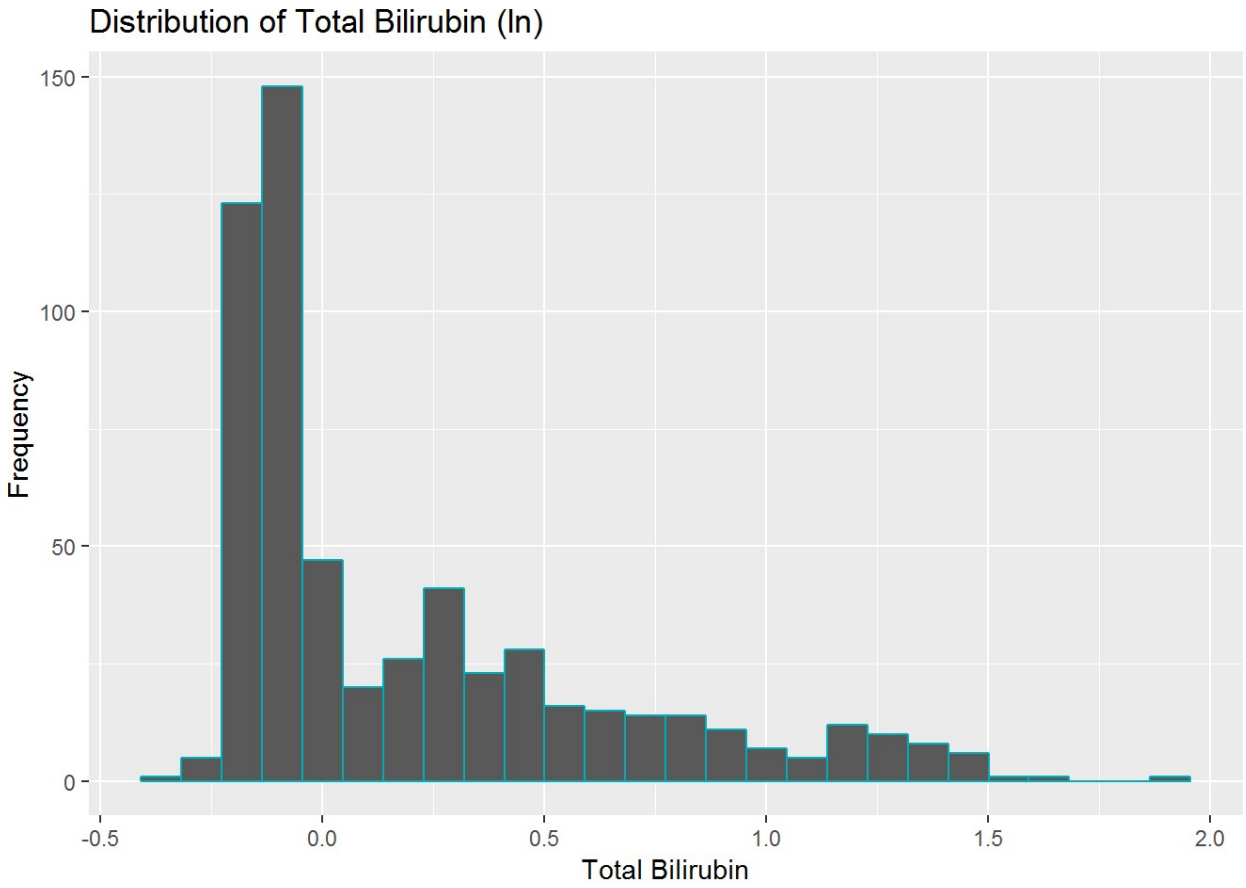
Sex <chr>	**Disease** <dbl>	**n** <int>
F	0	50
F	1	92
M	0	117
M	1	324

4 rows

Total Bilirubin

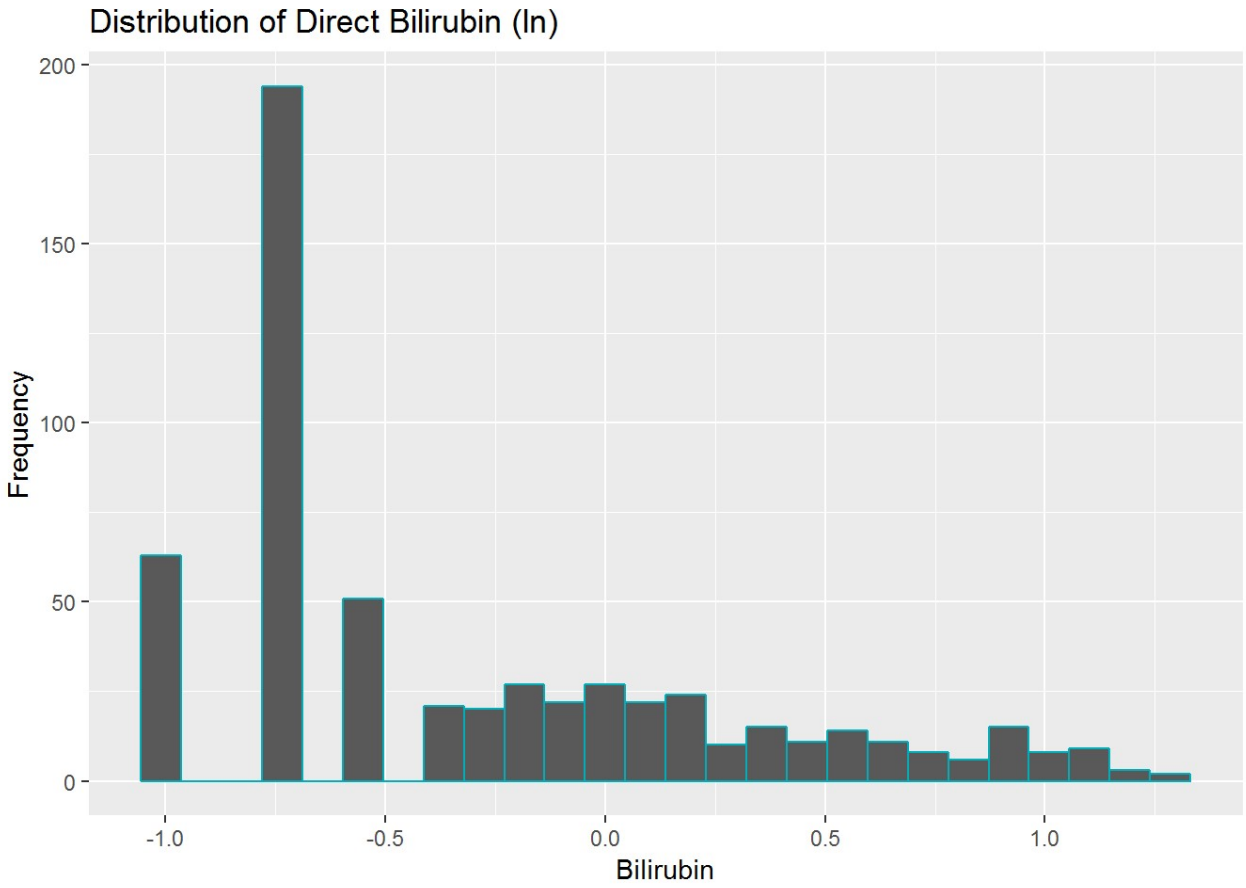
Bilirubin is a prodcut of the catabolism, is one of the substance that the liver have the job to filter. Elevated levels of this can be a hint of liver disease, it cause the change of color in the skin (yellow).

```
liver_data %>%
ggplot(aes(log10(Tot_Bil))) +
geom_histogram(bins = 26, color = "#00AFBB") +
xlab("Total Bilirubin") +
ylab("Frequency") +
ggtitle("Distribution of Total Bilirubin (ln)")
```



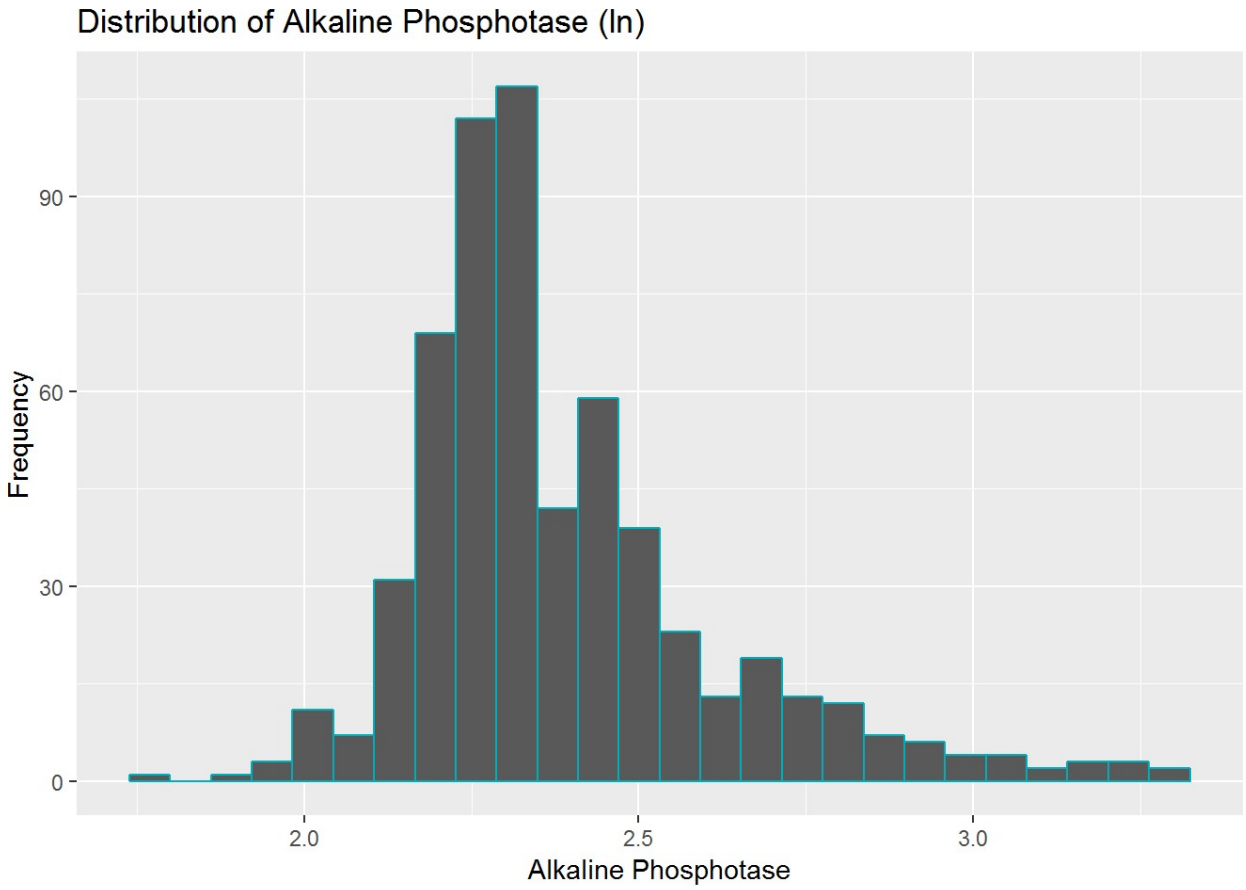
Direct Bilirubin

```
liver_data %>%
ggplot(aes(log10(Dir_Bil))) +
geom_histogram(bins = 26, color = "#00AFBB") +
xlab("Bilirubin") +
ylab("Frequency") +
ggtitle("Distribution of Direct Bilirubin (ln)")
```



Alkaline Phosphotase With this meause we can estimate, in general, **the liver health**. **More means disease**.

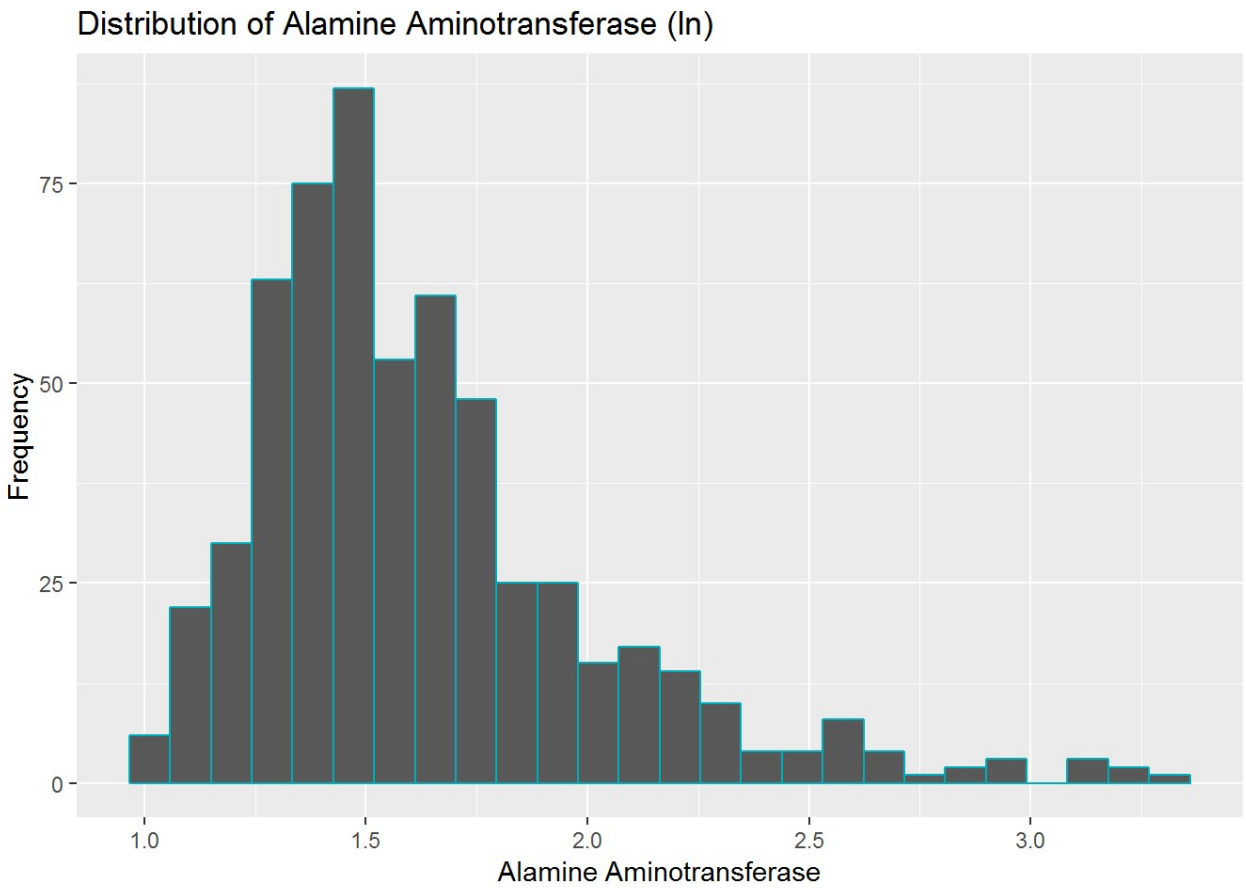
```
liver_data %>%
ggplot(aes(log10(Alkphos))) +
geom_histogram(bins = 26, color = "#00AFBB") +
xlab("Alkaline Phosphotase") +
ylab("Frequency") +
ggtitle("Distribution of Alkaline Phosphotase (ln)")
```



Alamine Aminotransferase

Natural component in the liver. Is tested in a liver panel.

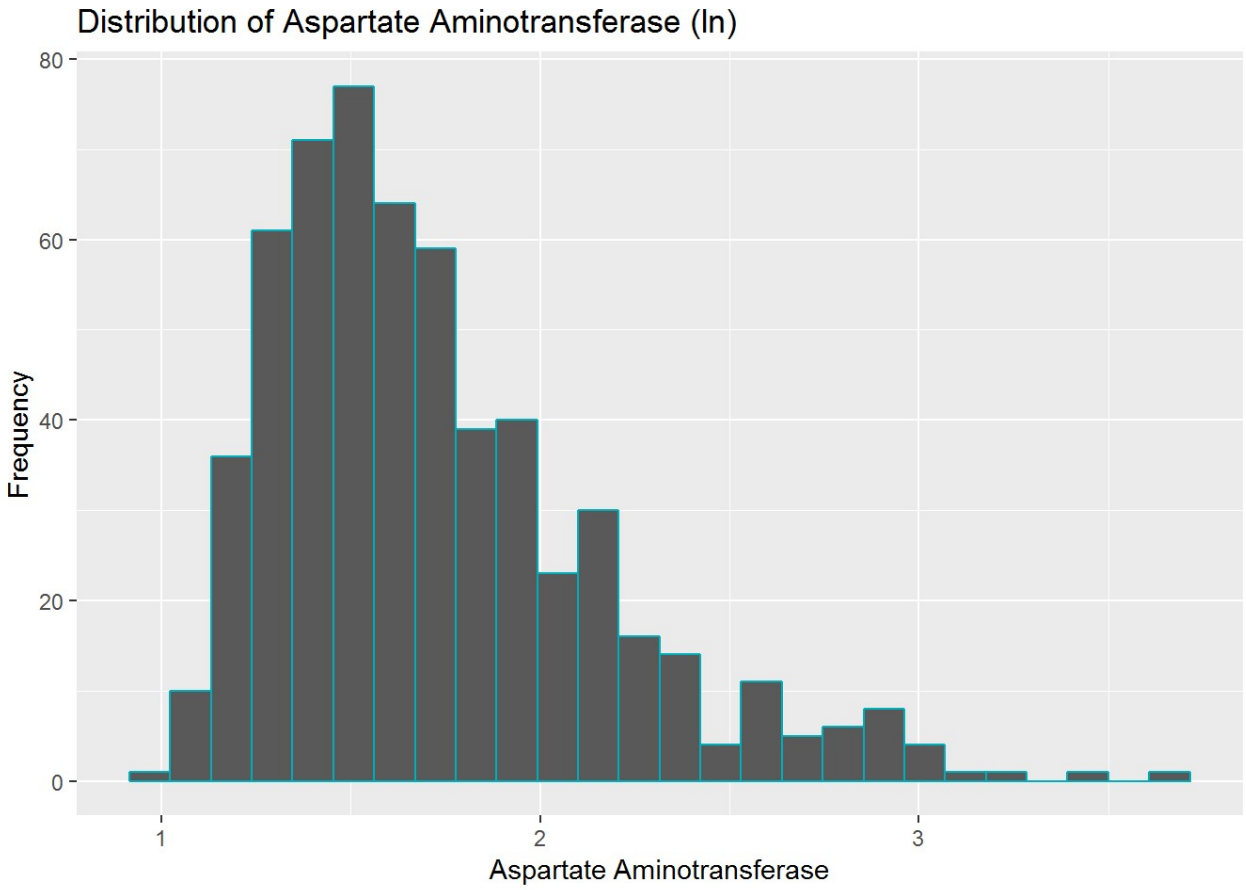
```
liver_data %>%
ggplot(aes(log10(Alamine))) +
geom_histogram(bins = 26, color = "#00AFBB") +
xlab("Alamine Aminotransferase") +
ylab("Frequency") +
ggtitle("Distribution of Alamine Aminotransferase (ln)")
```



Aspartate Aminotransferase

Natural component in the liver. Is tested in a liver panel.

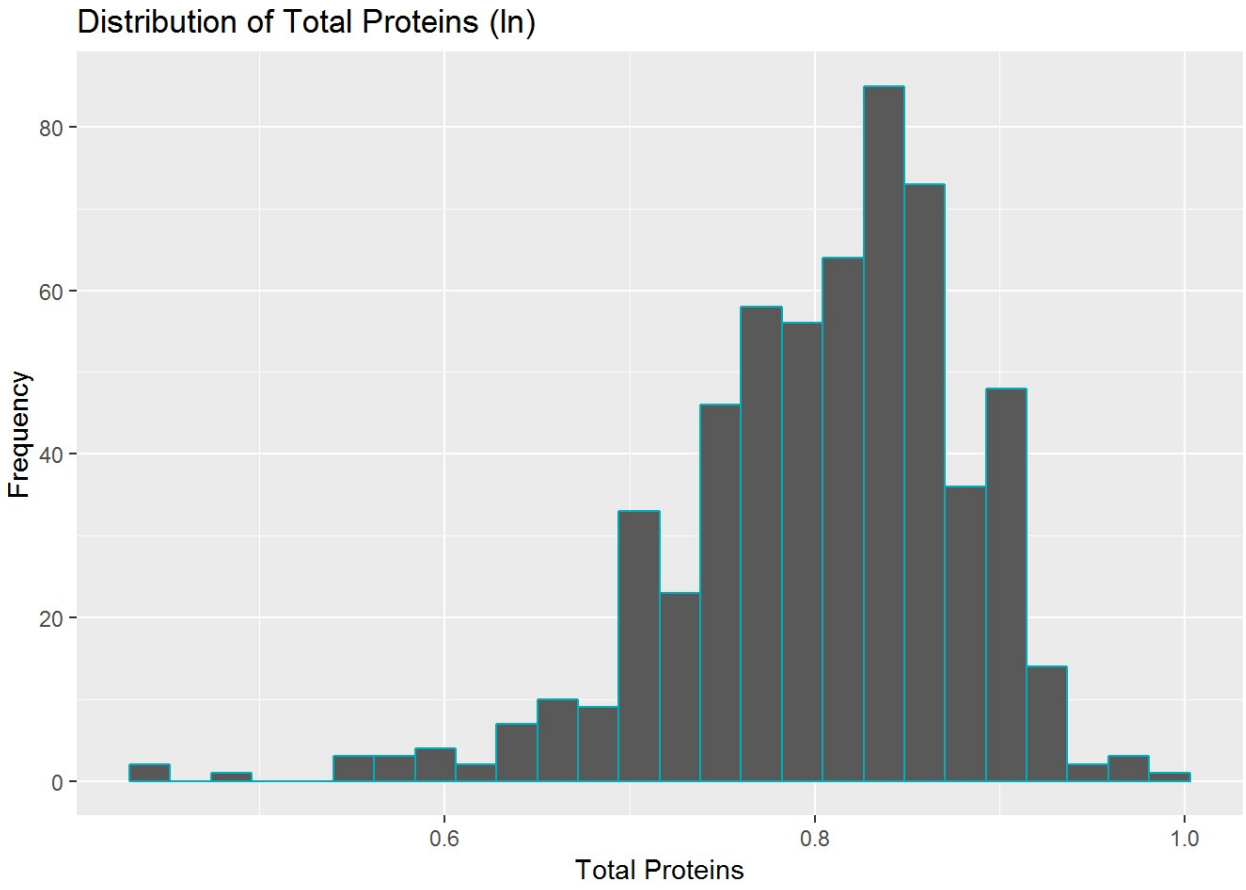
```
liver_data %>%
ggplot(aes(log10(Aspartate))) +
geom_histogram(bins = 26, color = "#00AFBB") +
xlab("Aspartate Aminotransferase") +
ylab("Frequency") +
ggtitle("Distribution of Aspartate Aminotransferase (ln)")
```



Total Proteins

Is a measure of globulin and albumin combined

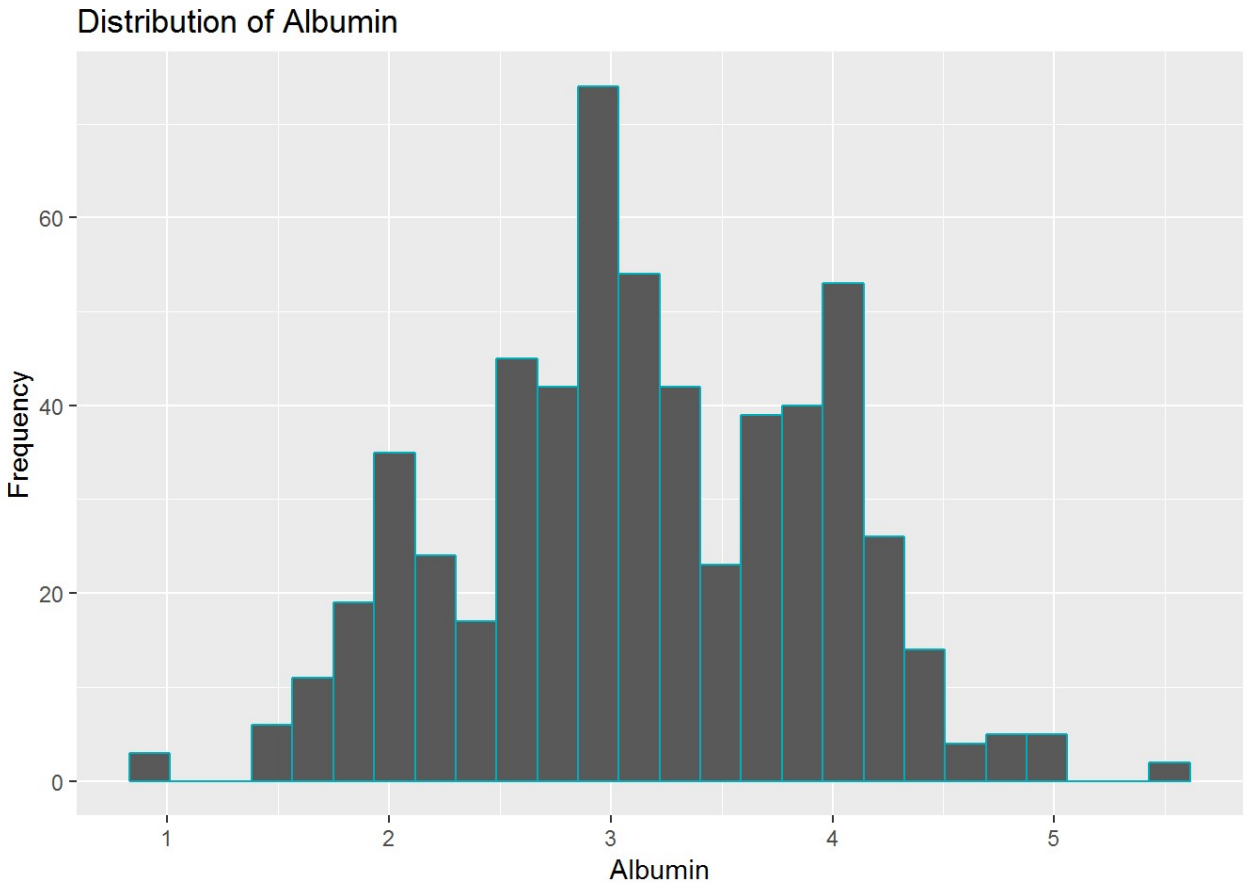
```
liver_data %>%
ggplot(aes(log10(Tot_Prot))) +
geom_histogram(bins = 26, color = "#00AFBB") +
xlab("Total Proteins") +
ylab("Frequency") +
ggtitle("Distribution of Total Proteins (ln)")
```



Albumin

Is a protein contained in the blood that gives structure to the vascular system.

```
liver_data %>%
ggplot(aes((Albumin))) +
geom_histogram(bins = 26, color = "#00AFBB") +
xlab("Albumin") +
ylab("Frequency") +
ggtitle("Distribution of Albumin")
```

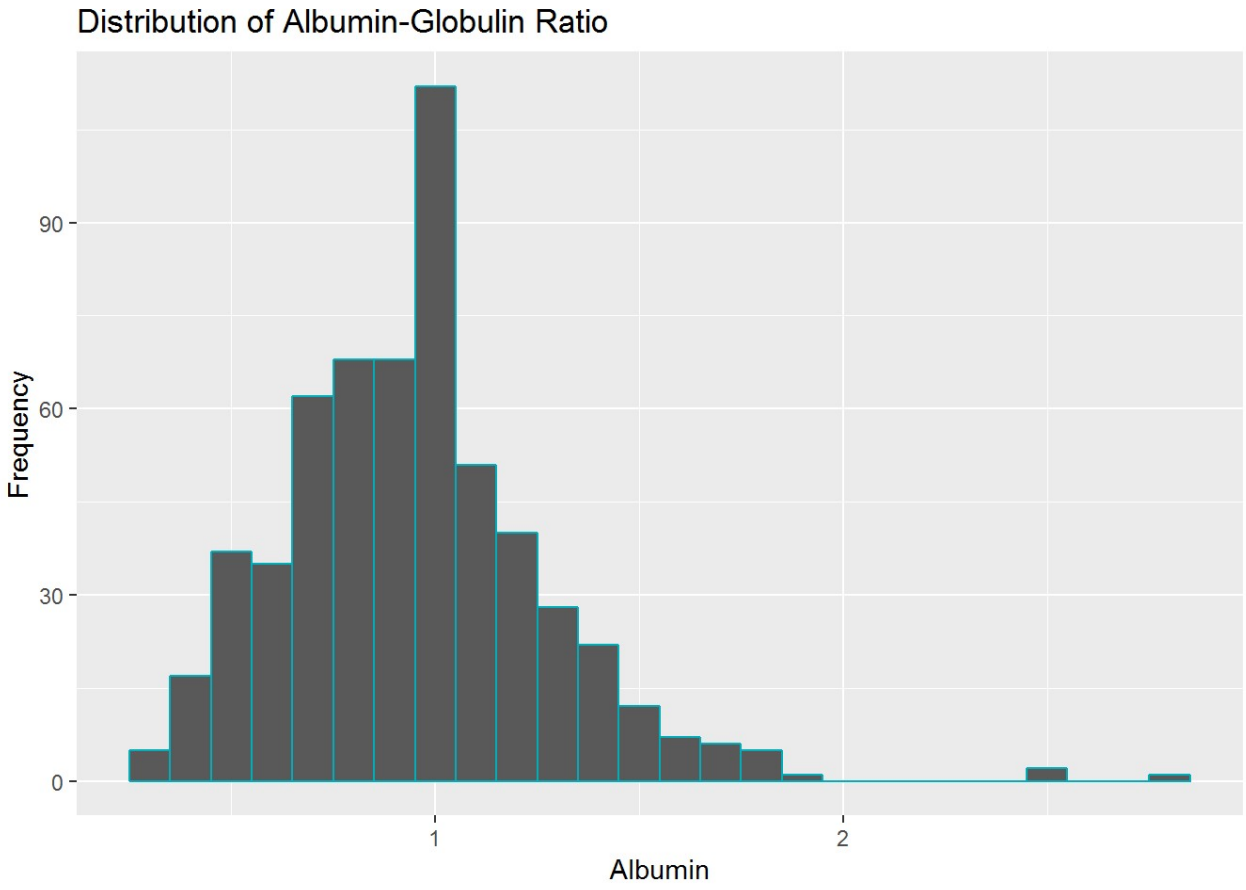


Albumin-Globulin Ratio

Is a general index of disesases.

```
liver_data %>%
ggplot(aes(A_G_Ratio)) +
geom_histogram(bins = 26, color = "#00AFBB") +
xlab("Albumin") +
ylab("Frequency") +
ggtitle("Distribution of Albumin-Globulin Ratio")
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



Model Approach

In this project we want to see if the Logistic Regression can be used on this data set to help us predict liver disease.

```
set.seed(455)
liver_data$Splits <- sample.split(liver_data, SplitRatio = 0.7)  #Index
liver_data <- liver_data %>% mutate_each(funs(log), -Age, -Sex, -Albumin, -A_G_Rati
o,
    -Disease, -Splits)
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## please use list() instead
##
## # Before:
## funs(name = f(.)
##
## # After:
## list(name = ~f(.))
## This warning is displayed once per session.
```

```
train <- liver_data[liver_data$Splits == TRUE, ]  #training index
test <- liver_data[liver_data$Splits == FALSE, ]  #test indexes
```

Summary of training

```
summary(train)
```

```
##       Age            Sex                Tot_Bil           Dir_Bil
##  Min.   : 4.00   Length:371         Min.   :-0.6931   Min.   :-2.3026
##  1st Qu.:33.00   Class :character   1st Qu.:-0.2231   1st Qu.:-1.6094
##  Median :45.00   Mode  :character   Median : 0.0000   Median :-1.2040
##  Mean   :44.29                      Mean   : 0.4635   Mean   :-0.6463
##  3rd Qu.:57.00                      3rd Qu.: 0.9555   3rd Qu.: 0.2624
##  Max.   :90.00                      Max.   : 4.3175   Max.   : 2.9806
##
##     Alkphos          Alamine         Aspartate         Tot_Prot
##  Min.   :4.143   Min.   :2.303   Min.   :2.485   Min.   :1.281
##  1st Qu.:5.185   1st Qu.:3.219   1st Qu.:3.258   1st Qu.:1.758
##  Median :5.371   Median :3.611   Median :3.761   Median :1.872
##  Mean   :5.518   Mean   :3.794   Mean   :4.010   Mean   :1.856
##  3rd Qu.:5.697   3rd Qu.:4.151   3rd Qu.:4.500   3rd Qu.:1.974
##  Max.   :7.654   Max.   :7.396   Max.   :8.503   Max.   :2.262
##
##     Albumin         A_G_Ratio         Disease         Splits
##  Min.   :1.400   Min.   :0.3000   Min.   :0.0000   Mode:logical
##  1st Qu.:2.600   1st Qu.:0.7000   1st Qu.:0.0000   TRUE:371
##  Median :3.100   Median :0.9000   Median :1.0000
##  Mean   :3.138   Mean   :0.9474   Mean   :0.7197
##  3rd Qu.:3.800   3rd Qu.:1.1000   3rd Qu.:1.0000
##  Max.   :5.500   Max.   :2.8000   Max.   :1.0000
##                  NA's   :1
```

Logistic Model

```
fit <- glm(Disease ~ Age + Sex + Tot_Bil + Dir_Bil + Alkphos + Alamine + Aspartate +
    Tot_Prot + Albumin + A_G_Ratio, data = train, family = binomial(link = "logit"))
```

Coefficients:

```
summary(fit)
```

```
##
## Call:
## glm(formula = Disease ~ Age + Sex + Tot_Bil + Dir_Bil + Alkphos +
##     Alamine + Aspartate + Tot_Prot + Albumin + A_G_Ratio, family = binomial(link
= "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4587  -0.9377   0.3171   0.8345   1.7431
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -15.855900   4.120022  -3.848 0.000119 ***
## Age           0.021706   0.008757   2.479 0.013190 *
## SexM         -0.400726   0.326646  -1.227 0.219901
## Tot_Bil       0.427679   0.743300   0.575 0.565035
## Dir_Bil       0.171838   0.485116   0.354 0.723174
## Alkphos       0.946313   0.385611   2.454 0.014125 *
## Alamine       0.937354   0.321930   2.912 0.003595 **
## Aspartate     0.198097   0.288925   0.686 0.492943
## Tot_Prot      5.326228   2.391509   2.227 0.025938 *
## Albumin      -1.455345   0.764960  -1.903 0.057104 .
## A_G_Ratio     1.890692   1.222460   1.547 0.121953
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 439.54  on 369  degrees of freedom
## Residual deviance: 341.15  on 359  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 363.15
##
## Number of Fisher Scoring iterations: 6
```

Pseudo R-Square and Log-Likelihoods: Usign the McFadden **R2** as a measure, the model explains just 22.5% of the disease classification.

```
pR2(fit)
```

```
##          llh      llhNull           G2     McFadden         r2ML
## -170.5744495 -220.0989199   99.0489409    0.2250101    0.2348626
##         r2CU
##    0.3375943
```

Now we will use the Coefficient of Discrimination

```
Test_Predictions <- data.frame(Probability = predict(fit, test, type = "response"))
Test_Predictions$Prediction <- ifelse(Test_Predictions > 0.5, 1, 0)
Test_Predictions$Disease <- test$Disease
accuracy <- mean(Test_Predictions$Disease == Test_Predictions$Prediction, na.rm = TRU
E)
disease <- Test_Predictions$Probability[which(Test_Predictions$Disease == 1)]
non <- Test_Predictions$Probability[which(Test_Predictions$Disease == 0)]
Coef_Desc <- mean(disease, na.rm = TRUE) - mean(non, na.rm = TRUE)
print(accuracy)
```

```
## [1] 0.6889952
```

The Coefficient of Discrimination is=

```
print(Coef_Desc)
```

```
## [1] 0.1420897
```

The accuracy of our model is=

```
print(accuracy)
```

```
## [1] 0.6889952
```

The model accuracy tell us the time our model made the right prediction, meaning taht the **prediction was right 69% of the times**.

Final considerations and further discussion

With the construction of the Pseudo R values (22.5%) that is really low for a predictor, i consider it will not help us at all to predict the disease and only will be a waste of time use it . But given it high accuracy of 69%, we can add more and more variables and values to the data set and could be use in a further test for the liver disease.

This model can be use for early detection, not for diagnostic.