Indian Liver Patient Records

Introduction

We download and use the data set ILPD (Indian Liver Patient Dataset) from web site kaggle (https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)) This data set contients the Indian patients that have a liver diseas and not liver, and our goal is predict status of liver patients or not

Data Analisys

The data set is composed from 11 variable:
1. Age : Age of the patient
2. Sex: Gender of the patient
3. TB: Total Bilirubin
4. DB: Direct Bilirubin
5. Alkphos: Alkaline Phosphotase
6. Alamine: Alamine Aminotransferase
7. Aspartate: Aspartate Aminotransferase
8. TP: Total Protiens
9. ALB: Albumin
10. A_G_Ratio: Ratio Albumin and Globulin Ratio
11. Diesease: Selector field used to split the data into two sets (labeled by the experts)

we see the first rows of data set:

```
head(df_IndianPatient)
```

```
##   Age    Sex   TB  DB Alkphos Alamine Aspartate  TP ALB A_G_Ratio Disease
## 1  62   Male 10.9 5.5     699      64       100 7.5 3.2      0.74       1
## 2  62   Male  7.3 4.1     490      60        68 7.0 3.3      0.89       1
## 3  58   Male  1.0 0.4     182      14        20 6.8 3.4      1.00       1
## 4  72   Male  3.9 2.0     195      27        59 7.3 2.4      0.40       1
## 5  46   Male  1.8 0.7     208      19        14 7.6 4.4      1.30       1
## 6  26 Female  0.9 0.2     154      16        12 7.0 3.5      1.00       1
```

Most variabiles are type integer or numerical but only the variable “Sex” is a factor

```
str(df_IndianPatient)
```

```
## 'data.frame':    578 obs. of  11 variables:
##  $ Age      : int  62 62 58 72 46 26 29 17 55 57 ...
##  $ Sex      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 1 1 2 2 2 ...
##  $ TB       : num  10.9 7.3 1 3.9 1.8 0.9 0.9 0.9 0.7 0.6 ...
##  $ DB       : num  5.5 4.1 0.4 2 0.7 0.2 0.3 0.3 0.2 0.1 ...
##  $ Alkphos  : int  699 490 182 195 208 154 202 202 290 210 ...
##  $ Alamine  : int  64 60 14 27 19 16 14 22 53 51 ...
##  $ Aspartate: int  100 68 20 59 14 12 11 19 58 59 ...
##  $ TP       : num  7.5 7 6.8 7.3 7.6 7 6.7 7.4 6.8 5.9 ...
##  $ ALB      : num  3.2 3.3 3.4 2.4 4.4 3.5 3.6 4.1 3.4 2.7 ...
##  $ A_G_Ratio: num  0.74 0.89 1 0.4 1.3 1 1.1 1.2 1 0.8 ...
##  $ Disease  : num  1 1 1 1 1 1 1 0 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int  209 241 253 312
##   ..- attr(*, "names")= chr  "209" "241" "253" "312"
```

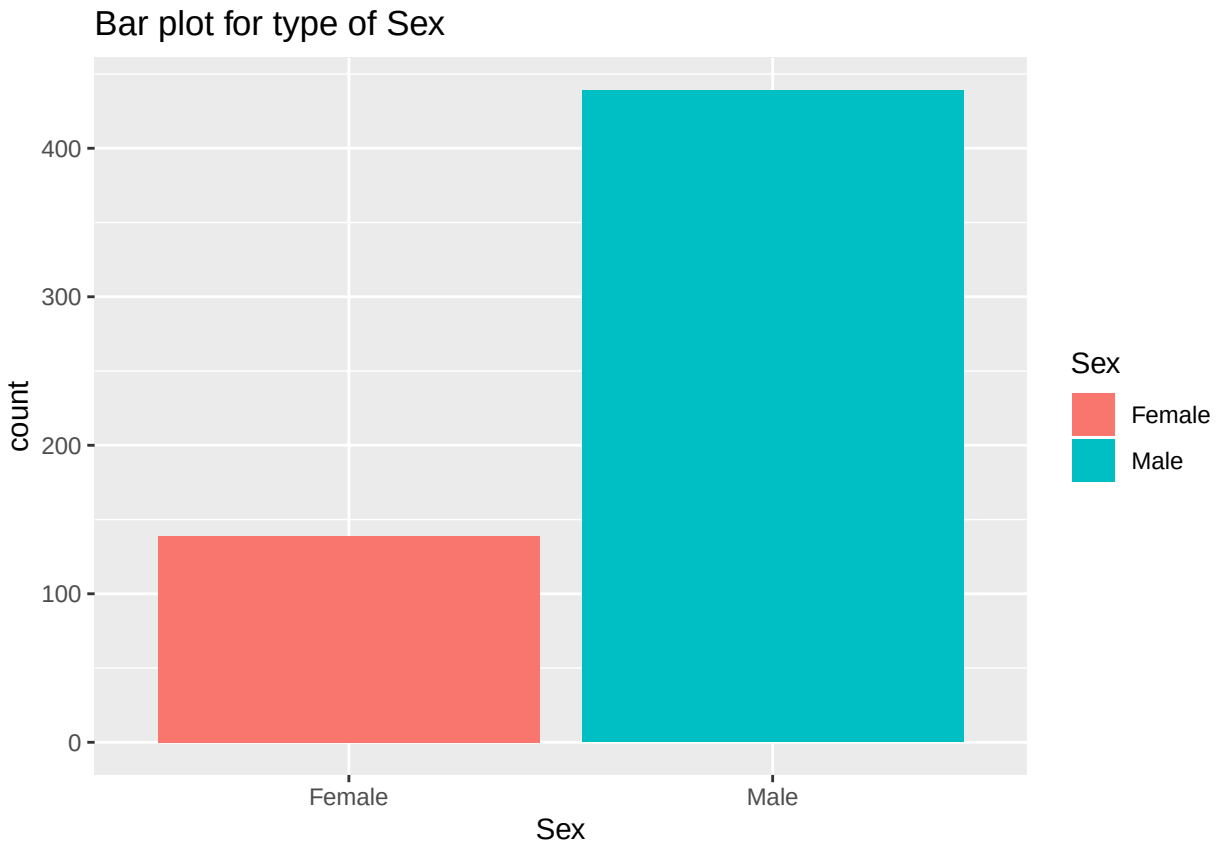
We see the summary of data set and their distribution

```r
summary(df_IndianPatient)
```

```
##       Age            Sex            TB              DB
##  Min.   : 4.00   Female:139   Min.   : 0.40   Min.   : 0.100
##  1st Qu.:33.00   Male  :439   1st Qu.: 0.80   1st Qu.: 0.200
##  Median :45.00                Median : 1.00   Median : 0.300
##  Mean   :44.75                Mean   : 3.32   Mean   : 1.497
##  3rd Qu.:58.00                3rd Qu.: 2.60   3rd Qu.: 1.300
##  Max.   :90.00                Max.   :75.00   Max.   :19.700
##     Alkphos          Alamine          Aspartate            TP
##  Min.   :  63.0   Min.   :  10.00   Min.   :  10.0   Min.   :2.700
##  1st Qu.: 175.2   1st Qu.:  23.25   1st Qu.:  25.0   1st Qu.:5.800
##  Median : 208.5   Median :  35.00   Median :  42.0   Median :6.600
##  Mean   : 291.5   Mean   :  81.24   Mean   : 110.6   Mean   :6.481
##  3rd Qu.: 298.0   3rd Qu.:  61.00   3rd Qu.:  87.0   3rd Qu.:7.200
##  Max.   :2110.0   Max.   :2000.00   Max.   :4929.0   Max.   :9.600
##       ALB          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 :0.9400   Median :1.0000
##  Mean   :3.138   Mean   :0.9471   Mean   :0.7145
##  3rd Qu.:3.800   3rd Qu.:1.1000   3rd Qu.:1.0000
##  Max.   :5.500   Max.   :2.8000   Max.   :1.0000
```

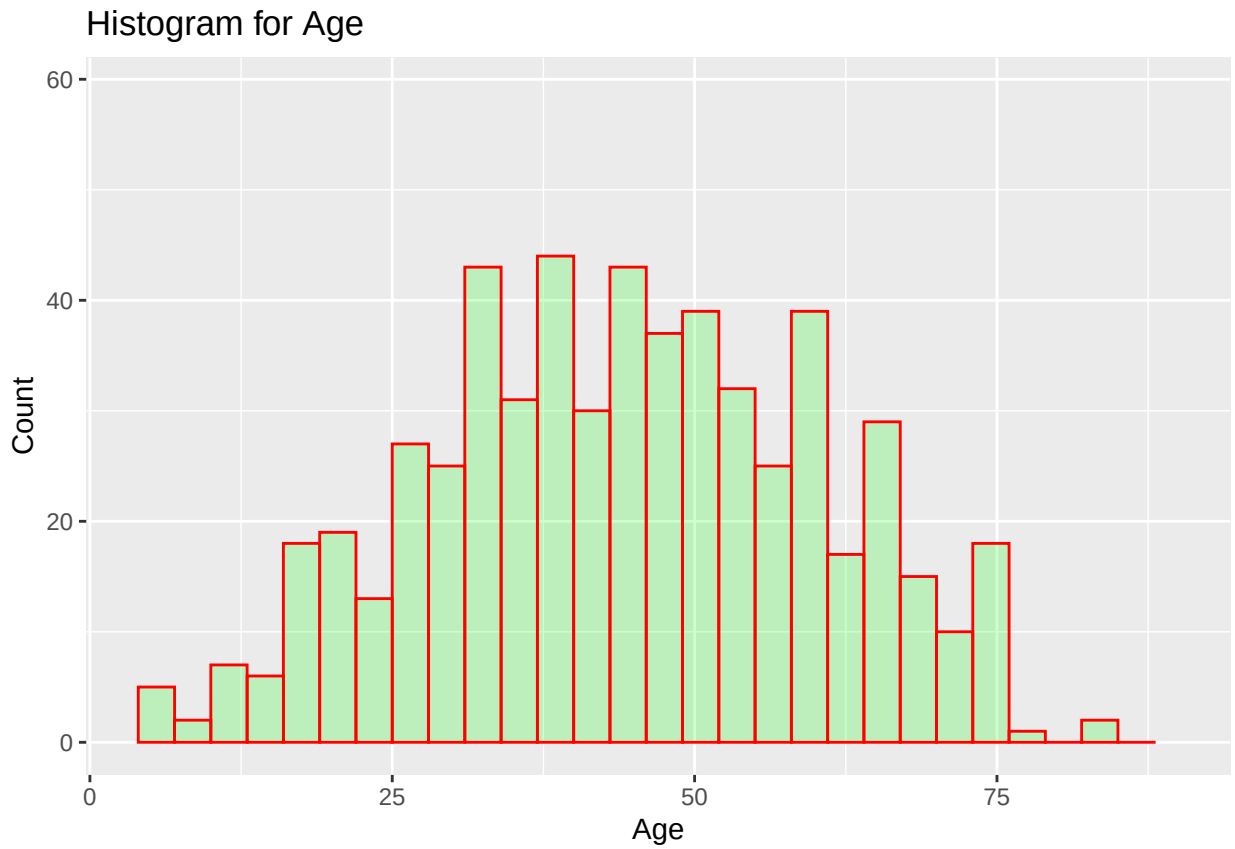
Data Visualization

The type of Sex in the Indian Patients are more of type Male than Female



The Male Patients are 441 and the Female are 141

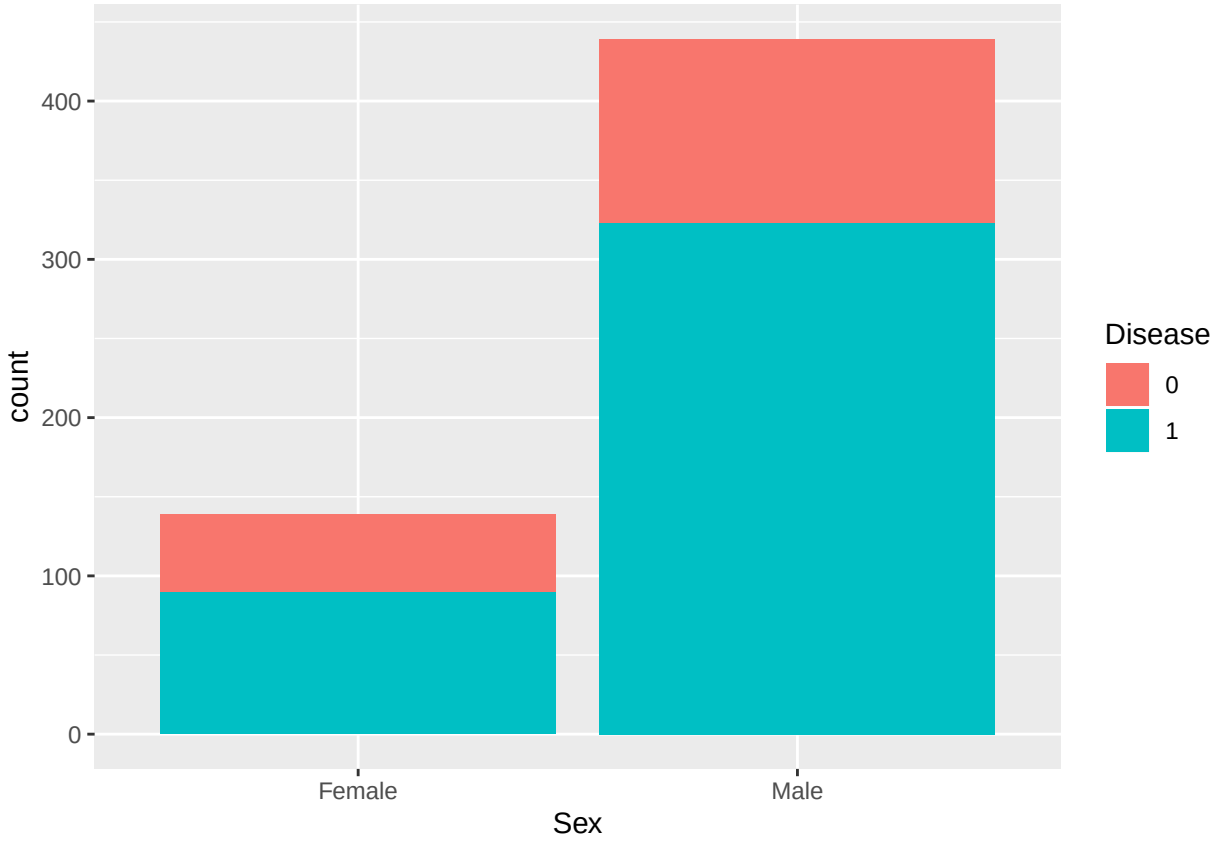
We see the distribution of Age in the histogram, that is similar at the gaussian.



and the age for type of sex

The Disease for type of Sex have this distribution:

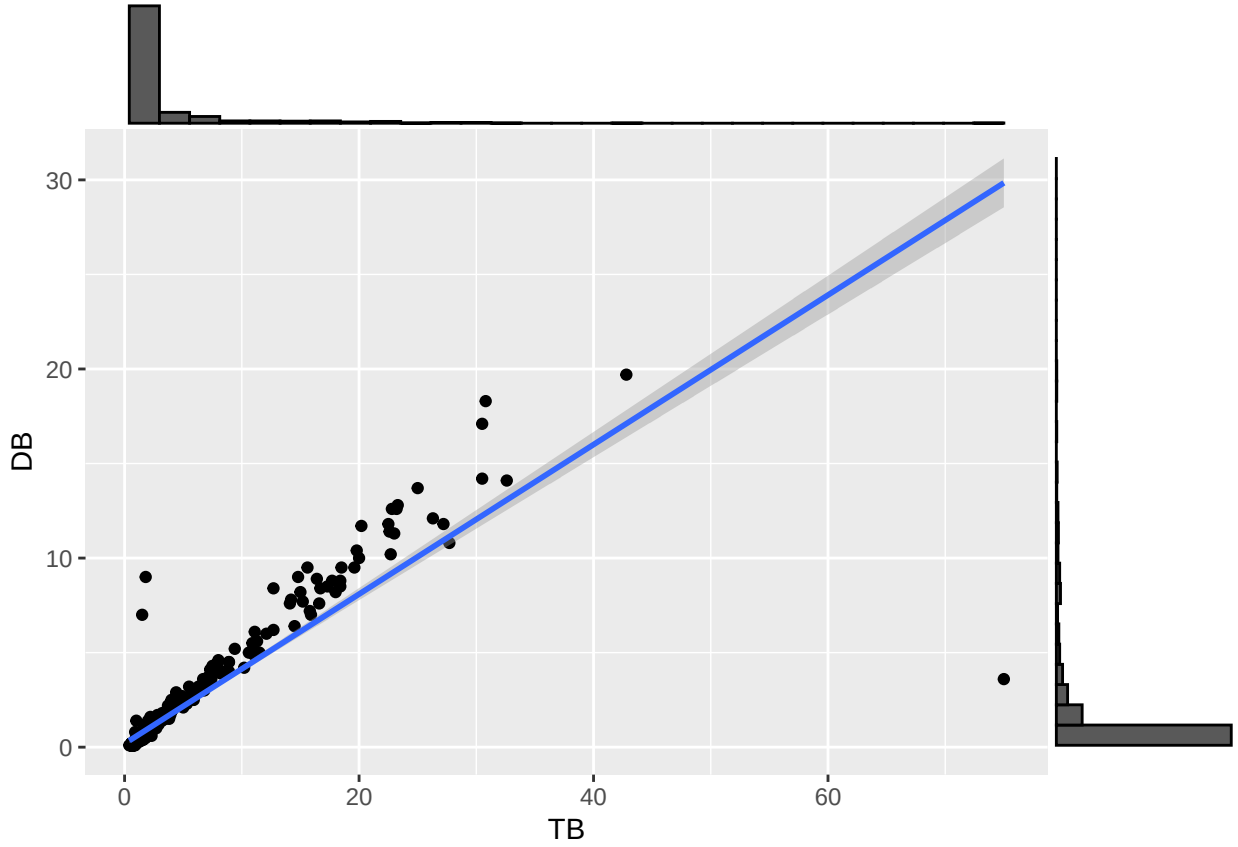
```
##
##     Female Male
##   0     49  116
##   1     90  323
```



There are in the data set most cases with Disease Liver Patient

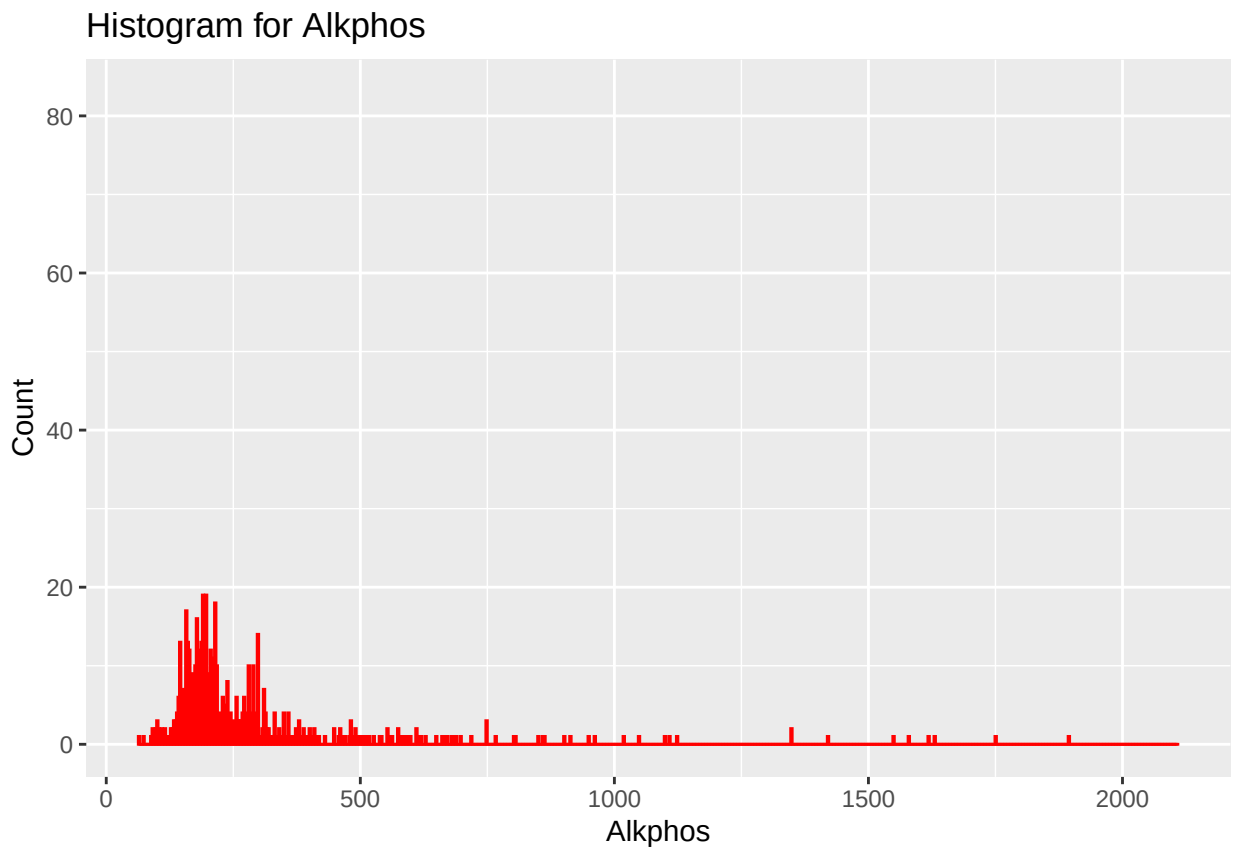
The Bilorubin is define as yellow compound that occurs in the normal catabolic pathway that breaks down heme in vertebrates. This catabolism is a necessary process in the body's clearance of waste products that arise from the destruction of aged or abnormal red blood cell (for other information to Bilirubin to see https://en.wikipedia.org/wiki/Bilirubin)

We rappresent the relation beetwen "Total Bilurubin" and "Direct Bilirubin"

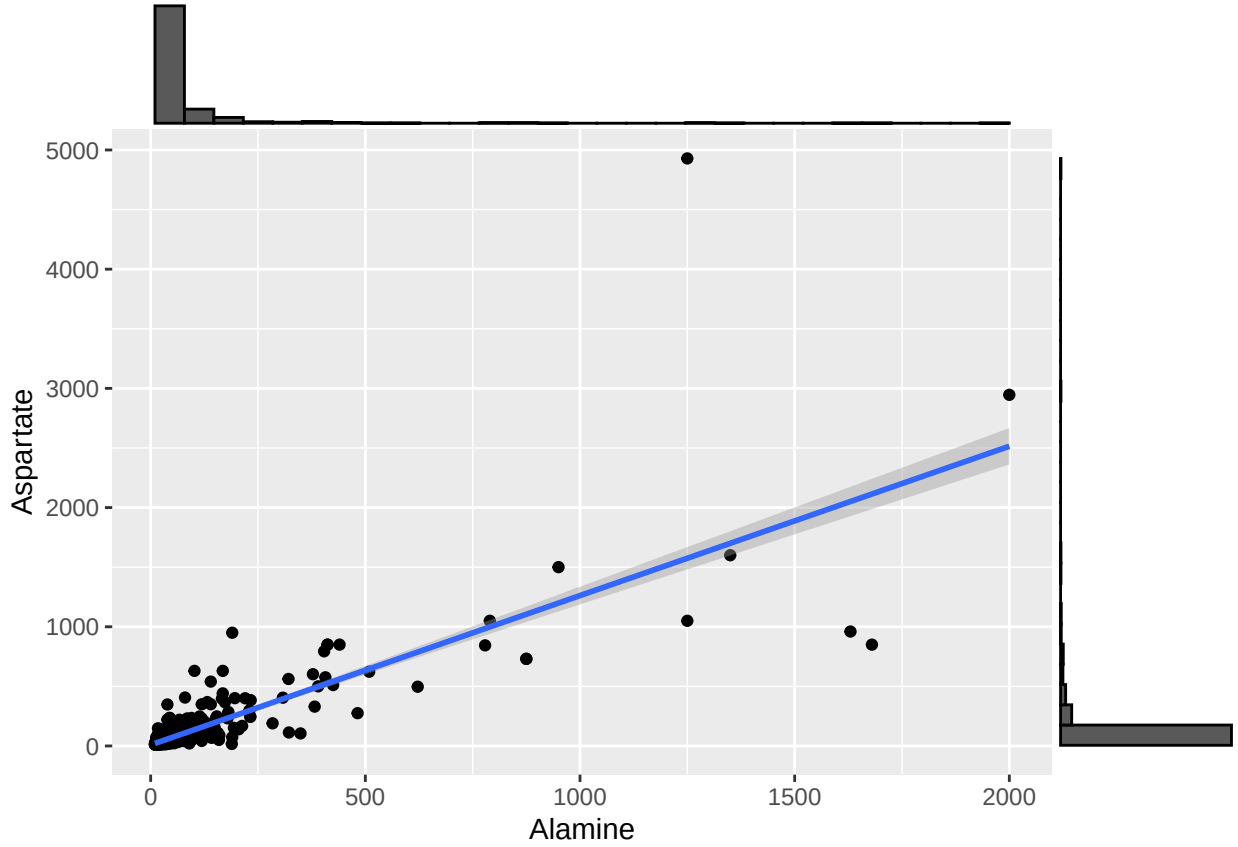


where there is a linear correlation

After we see the Alkphos, that is a homodimeric protein enzyme of 86 kilodaltons. (For other information to Alkphos to see https://en.wikipedia.org/wiki/Alkaline_phosphatase)

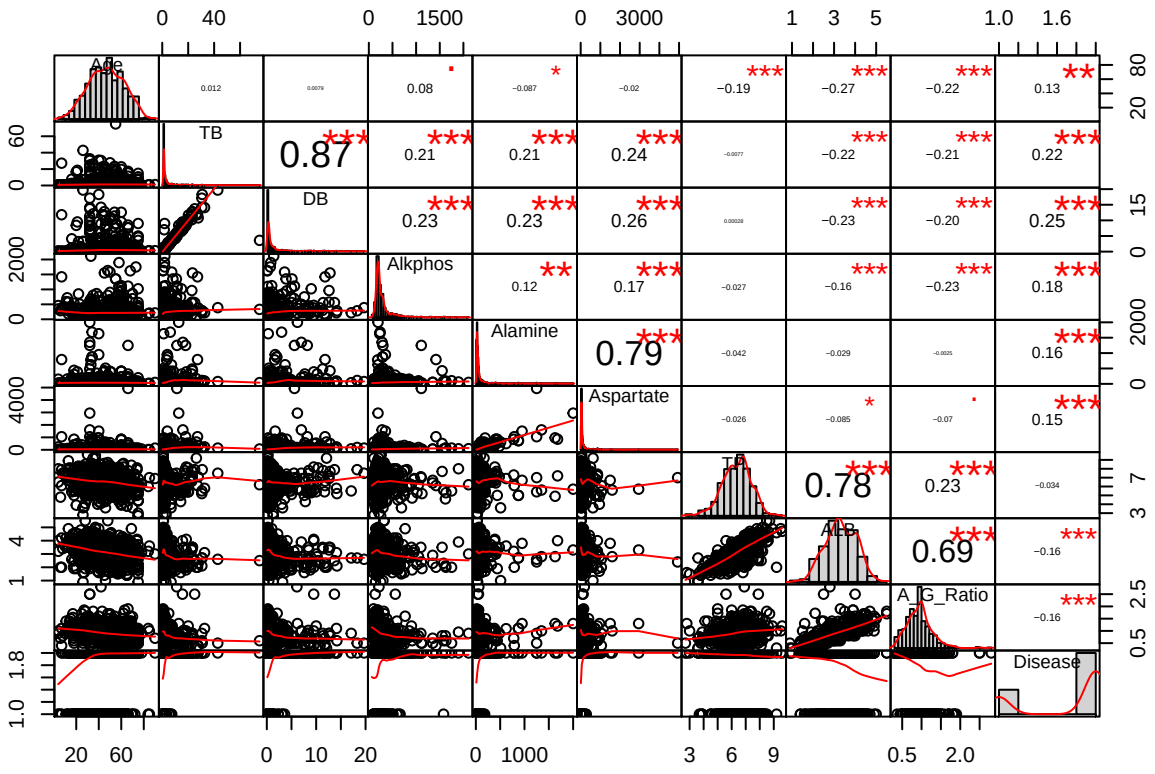


We see the ralation beetwen Alamine Aminotransferase (was formerly called serum glutamate-pyruvate transaminase (SGPT) or serum glutamic-pyruvic transaminase (SGPT)) and Aspartate Aminotransferase (is a pyridoxal phosphate (PLP)-dependent transaminase enzyme)



There is a linear correaltion beetwen two variable

Now we can see the relation beetwen all variable, in an unique plot



there are the following relations
- “TB” and “DB”"
- “Alamine” and “Aspartate”
- “TP” and “ALB”
- “ALB” and “A_G_Ratio”

that are all linear relation most significative

Model of analisys

Now create our set data to predict the model

```
set.seed(7) # for reproducibility
test_index <- createDataPartition(y = df_IndianPatient$Disease, times = 1, p = 0.7, list = FALSE)
edx <- df_IndianPatient[-test_index,]
temp <- df_IndianPatient[test_index,]
```

Logistic Regression

We applicate the logistic regression to predict a model, because the output variable assume value 0 and 1

```
fit <- glm(Disease ~ Age + Sex + TB + DB + Alkphos + Alamine + Aspartate +
    TP + ALB + A_G_Ratio, data = edx, family = binomial(link = "logit"))
```

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

```
summary(fit)
```

```
##
```

```
## Call:
## glm(formula = Disease ~ Age + Sex + TB + DB + Alkphos + Alamine +
##     Aspartate + TP + ALB + A_G_Ratio, family = binomial(link = "logit"),
##     data = edx)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1745  -0.8103   0.3822   0.8873   1.4105
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.0177456  2.0581058  -1.952  0.05092 .
## Age          0.0334366  0.0120715   2.770  0.00561 **
## SexMale     -0.2586197  0.4545964  -0.569  0.56942
## TB           0.2064170  0.5715530   0.361  0.71799
## DB          -0.0534980  0.9996706  -0.054  0.95732
## Alkphos     -0.0009868  0.0010467  -0.943  0.34582
## Alamine      0.0312391  0.0128641   2.428  0.01517 *
## Aspartate    0.0005161  0.0066522   0.078  0.93816
## TP           0.6285583  0.5040662   1.247  0.21241
## ALB         -0.9720075  0.9692552  -1.003  0.31594
## A_G_Ratio    1.2497371  1.5006172   0.833  0.40495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 205.54  on 171  degrees of freedom
## Residual deviance: 162.28  on 161  degrees of freedom
## AIC: 184.28
##
## Number of Fisher Scoring iterations: 8
```

Calculate the accuracy of model, that is

```
Edx_Predictions <- data.frame(Probability = predict(fit, edx, type = "response"))
Edx_Predictions$Prediction <- ifelse(Edx_Predictions > 0.5, 1, 0)
Edx_Predictions$Disease <- edx$Disease
accuracy <- mean(Edx_Predictions$Disease == Edx_Predictions$Prediction, na.rm = TRUE)
tot_accuracy<-data.frame(Model="Logistic Regression",Value=accuracy)

accuracy
```

```
## [1] 0.7848837
```

Backward in Logistic Regression

Now we applicate the backward for select the best significative variabiles to rappresent the model.

```
## Subset selection object
## Call: regsubsets.formula(Disease ~ ., edx, method = "backward")
## 10 Variables  (and intercept)
##           Forced in Forced out
## Age           FALSE      FALSE
## SexMale       FALSE      FALSE
## TB            FALSE      FALSE
```

```
## DB            FALSE      FALSE
## Alkphos       FALSE      FALSE
## Alamine       FALSE      FALSE
## Aspartate     FALSE      FALSE
## TP            FALSE      FALSE
## ALB           FALSE      FALSE
## A_G_Ratio     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: backward
##          Age SexMale TB  DB  Alkphos Alamine Aspartate TP  ALB A_G_Ratio
## 1  ( 1 ) " " " "     " " "*" " "     " "     " "       " " " " " "
## 2  ( 1 ) "*" " "     " " "*" " "     " "     " "       " " " " " "
## 3  ( 1 ) "*" " "     " " "*" " "     "*"     " "       " " " " " "
## 4  ( 1 ) "*" " "     " " "*" " "     "*"     " "       "*" " " " "
## 5  ( 1 ) "*" " "     " " "*" " "     "*"     " "       "*" "*" " "
## 6  ( 1 ) "*" " "     " " "*" " "     "*"     " "       "*" "*" "*"
## 7  ( 1 ) "*" " "     " " "*" "*"     "*"     " "       "*" "*" "*"
## 8  ( 1 ) "*" "*"     " " "*" "*"     "*"     " "       "*" "*" "*"
```

The most significative variables for model are Age, DB, Alamine, Thart select the m

```
backward.model <- glm(Disease ~ Age + DB + Alamine, data = edx, family = binomial(link = "logit"))
```

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

```
summary(backward.model)[7]
```

```
## $df.residual
## [1] 168
```

and we calculate the accuracy

```
predBw<-predict(backward.model,temp,type = "response")
predicted.BW <- as.numeric(ifelse(predBw > 0.5, 1, 0))
accuracyBackward<-mean(predicted.BW==temp$Disease,na.rm=TRUE)
tot_accuracy<-rbind(tot_accuracy,data.frame(Model="Backward Logistic Regression",Value=accuracyBackward
accuracyBackward
```

```
## [1] 0.7068966
```

Random Forest

Predict a Model with approach Random Forest

```
IDLR.rf=randomForest(Disease ~ ., data = edx,
              na.action=na.exclude)
 pred<-as.numeric(predict(IDLR.rf,temp))
```

and we have the accuracy model with this value:

```
## [1] 0.6773399
```

Results

We see the results of accuracy of differents models applicate

```
tot_accuracy %>% knitr::kable()
```

Model	Value
Logistic Regression	0.7848837
Backward Logistic Regression	0.7068966
Random Forest	0.6773399

Conclusion

We see the best model to predict the Diseas is the first model where we applicate the logistic regression. This model have a good prediction for our data.