# Indian Liver Patient Record Project

Alejandro Jiménez

May 25, 2020

## Introduction

Every day each it's vital to identify early a disease, this could significate the survival of a patient. Thanks to medical exams, the doctor can analyze various factor and determine if a patient has a risk of a disease. In order to attempt helping the doctors to make a faster analysis, there can be a data analysis to create a model that can predict a certain disease and help doctor to make a quicker treatment.

This work analyze the "Indian Liver Patient Records" extracted from Kaggle to attempt building an algorithm that helps determine if a patient has a liver disease with a good accuracy. The methods used in the analysis were: K-means, logistic regression, LDA, QDA, Loess, kNN, Random Forest and ensemble all this methods. This methods were used first not considering gender and later consider gender to see if there was improvement in the accuracy. Finally there was an analysis with a neural network to see if there was a better result than the other methods.

### Data Preparation

As mentioned before this data was extracted from Kaggle. The file (cvs format) was previously downloaded in a .zip file and extracted. The next code shows the packages used in the process and the data stored in the "patients" variable.

```
#Install packages (if necessary) and load them
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(matrixStats)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(neuralnet)) install.packages("caret", repos = "http://cran.us.r-project.org")

#Indian Liver Patient Records dataset:
  # https://www.kaggle.com/uciml/indian-liver-patient-records?select=indian_liver_patient.csv
  # Data previously downloaded in a .zip file and extract the dataset.

#Inspecting the file, it seems that it have column names.
#.csv files, .R and .Rmd must be treated as a R project to load correctly the files
#Load dataset
patients <- read_csv("indian_liver_patient.csv")
```

Inspecting the data we see there are eleven variables, ten being numeric and Gender a character:

```
##                       Age                    Gender
##                 "numeric"                 "character"
##           Total_Bilirubin           Direct_Bilirubin
##                 "numeric"                   "numeric"
##       Alkaline_Phosphotase   Alamine_Aminotransferase
```

```
##                   "numeric"                   "numeric"
## Aspartate_Aminotransferase             Total_Protiens
##                   "numeric"                   "numeric"
##                     Albumin Albumin_and_Globulin_Ratio
##                   "numeric"                   "numeric"
##                     Dataset
##                   "numeric"
```

Observing the classes and variables, we rearrange the order of the data to put the "Dataset" in the first column since this is our outcome and change its name to "Liver" for easier recognition. Then change the class of "Liver" and "Gender" into factor and finally change NA's into zeros. We see our data in the next format after this changes:

```
## $Liver
## integer(0)
##
## $Gender
## integer(0)
##
## $Age
## integer(0)
##
## $Total_Bilirubin
## integer(0)
##
## $Direct_Bilirubin
## integer(0)
##
## $Alkaline_Phosphotase
## integer(0)
##
## $Alamine_Aminotransferase
## integer(0)
##
## $Aspartate_Aminotransferase
## integer(0)
##
## $Total_Protiens
## integer(0)
##
## $Albumin
## integer(0)
##
## $Albumin_and_Globulin_Ratio
## [1] 210 242 254 313
```

```
##                      Liver                      Gender
##                   "factor"                    "factor"
##                        Age             Total_Bilirubin
##                   "numeric"                   "numeric"
##           Direct_Bilirubin       Alkaline_Phosphotase
##                   "numeric"                   "numeric"
##   Alamine_Aminotransferase Aspartate_Aminotransferase
##                   "numeric"                   "numeric"
##             Total_Protiens                     Albumin
```

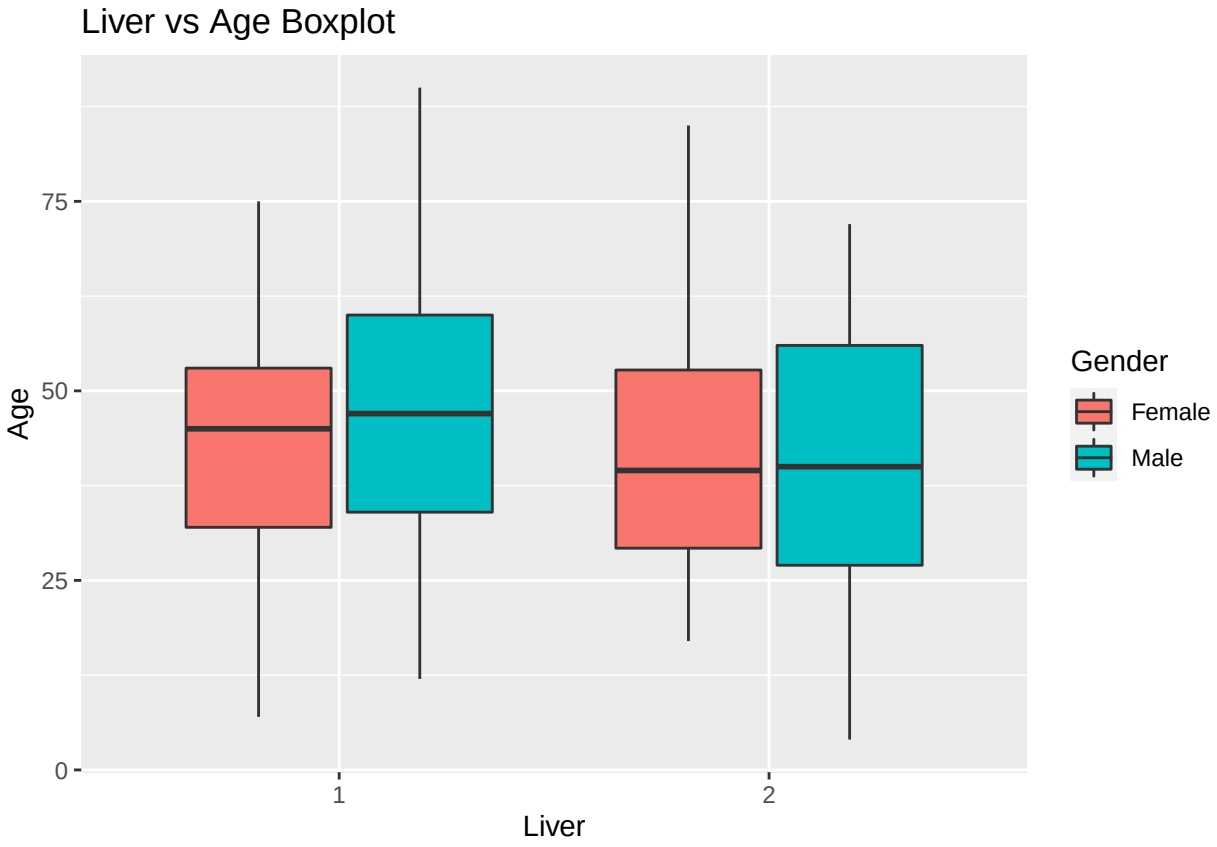
```
##                   "numeric"                   "numeric"
## Albumin_and_Globulin_Ratio
##                   "numeric"
```

# Data Analysis

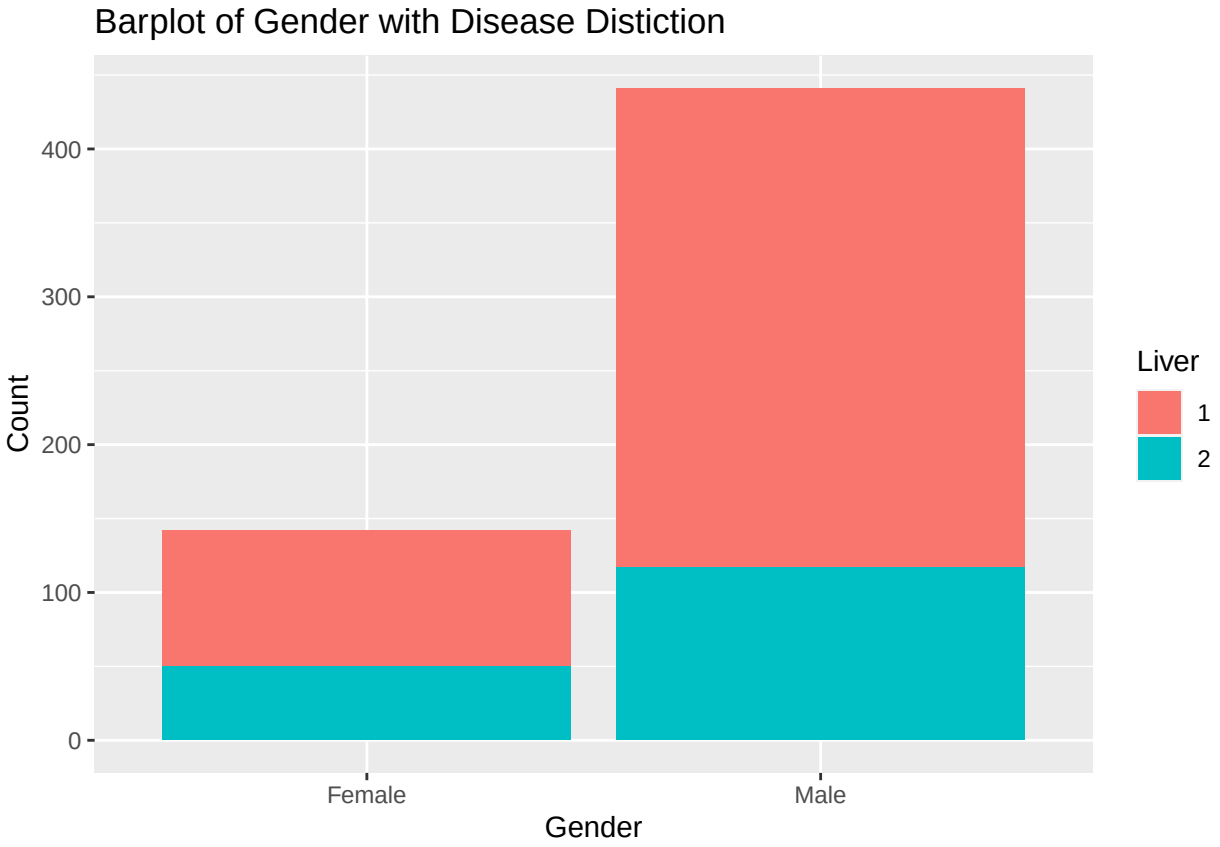
##Data Exploration First we identify with value in "Liver" indicates a disease:

```
##   1   2
## 416 167
```

According to the page, there are 416 patients with liver disease. Factor 1 corresponds to having liver disease and 2 patients who don't. Then we see the distribution of patients through age with liver disease and gender distinction. We observe that there's no a clear difference of patients with and without disease by their age.

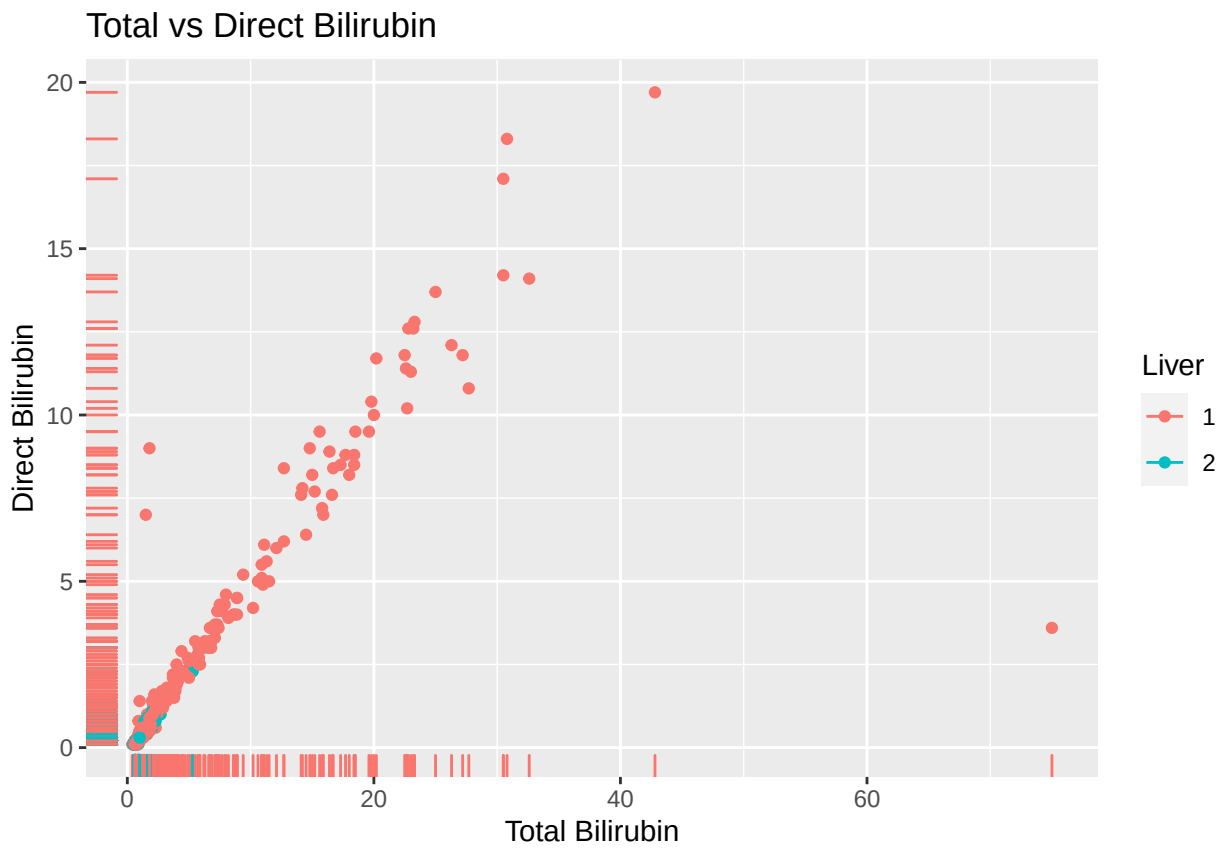



Then we examine proportion of patients with disease by gender. The bar plot and the proportions doesn't significant difference that gender is a factor for liver disease.

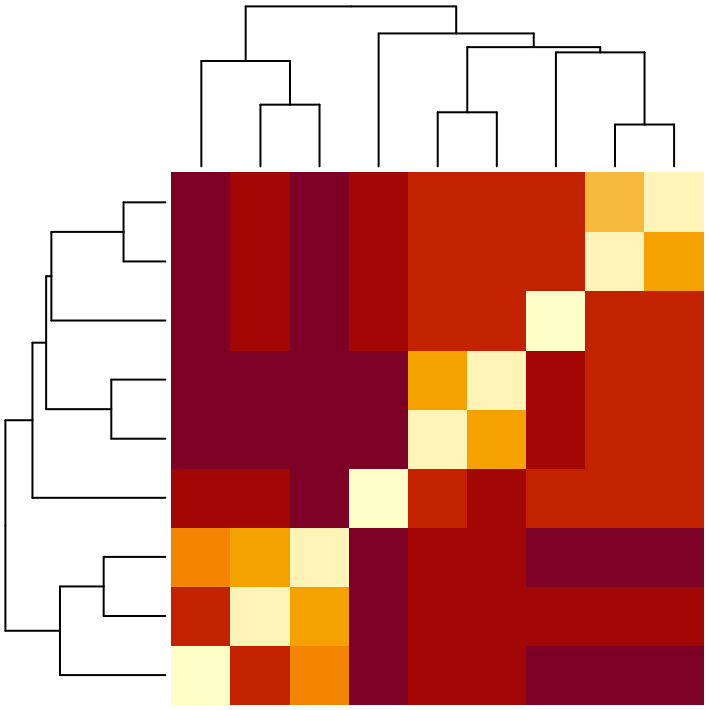



```
## # A tibble: 2 x 3
##   Gender Disease Not_Disease
##   <fct>    <dbl>       <dbl>
## 1 Female   0.648       0.352
## 2 Male     0.735       0.265
```

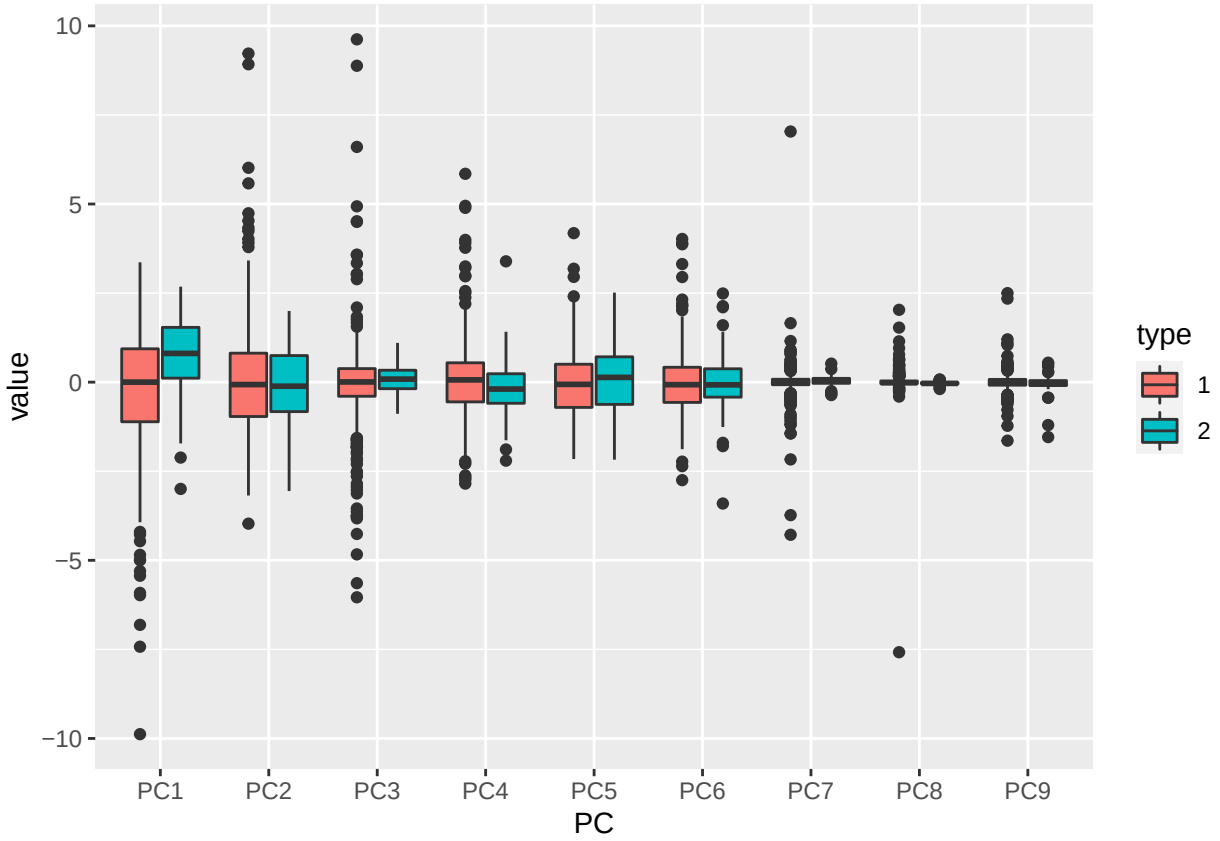
The next plot show the relation of Total and Direct Bilirubin. We observe there's a positive correlation between this variables. We also observe a major concentration of values in the left bottom size, including the values of people with no liver disease, there's no clear distinction between patients with and without disease in this graph.




Next we observe a heat map of the relation between variables, to do this first we exclude the gender feature since it's a class variable and in the previous plots we observe there's no significant difference between male and female patients. First we scale the features and obtain the distance, the next plot show a heat map of this relation between variables.



We observe that there's some correlation between some variables, but this aren't too strong to discard more features in our model. Then we run a Principal Component Analysis (PCA) to observe if some variables will help distinguish the liver condition. The next plot show the histogram all the features; we observe there's an issue with our data. Although some variable have major contribution, there's no clear distinction of patients with and without liver disease. The interquartile range of liver free disease falls into the interquartile range of patients with disease. This suspect that our fitting models won't have a good accuracy.




## Analysis without Gender

To run our analysis we split our data into test and train datasets. The proportion of the test dataset it 15% since we can't to train our models with the more data as possible since there are only 583 observations. After getting our datasets we start our analysis, we train our models and then obtain our prediction of the test dataset. Then with the confusion matrix we extract the accuracy, sensitivity, specificity and balanced accuracy. The balanced accuracy is our value of interest since this is a binary classification. All result are stored in and show in ta table to compare our models.

The first one is with the k-means method. To obtain our predations we define the next function:

```
#Define Fuction to predict with K-means
predict_kmeans <- function(k, x){
  centers <- k$centers
  n_centers <- nrow(centers)
  dist_mat <- as.matrix(dist(rbind(centers, x)))
  dist_mat <- dist_mat[-seq(n_centers), seq(n_centers)]
  max.col(-dist_mat)
}
```

The methods used are: K-mean, logistic regression, LDA, QDA, Loess, kNN, Random Forest and ensemble this methods. For simplicity the results of all the analysis are shown the next table. For the ensemble method, is defined by the means of the predictions. If the mean is more than 0.5 then its prediction is 1, otherwise is 2. The next table show a low balance accuracy, the majority of the methods have a low specificity and the best prediction is given by QDA. In the result section we discuss more the values obtained.

| Method | Accuracy | Sensitivity | Specificity | Balanced_Accuracy |
|---|---|---|---|---|
| K-Means | 0.6067416 | 0.8412698 | 0.0384615 | 0.4398657 |
| Logistic Regression (glm) | 0.7303371 | 0.9523810 | 0.1923077 | 0.5723443 |
| LDA | 0.7415730 | 1.0000000 | 0.1153846 | 0.5576923 |
| QDA | 0.5617978 | 0.4126984 | 0.9230769 | 0.6678877 |
| Loess | 0.6629213 | 0.8571429 | 0.1923077 | 0.5247253 |
| kNN | 0.6741573 | 0.8730159 | 0.1923077 | 0.5326618 |
| Random Forest (rf) | 0.6404494 | 0.8095238 | 0.2307692 | 0.5201465 |
| Ensemble | 0.7303371 | 0.9523810 | 0.1923077 | 0.5723443 |

## Analysis with Gender

Even though we discard the gender feature, in this section it is incorporated due to the low balanced accuracies obtained in the last methods in attempt to improve it. To incorporate a classification variable, we define males to be -1 and female 1. The next table show the results with Genre and observe there's no improvement, actually in some methods like LDA give a worst balanced accuracy.

| Method_Gender | Accuracy | Sensitivity | Specificity | Balanced_Accuracy |
|---|---|---|---|---|
| K-Means | 0.6067416 | 0.8412698 | 0.0384615 | 0.4398657 |
| Logistic Regression (glm) | 0.7191011 | 0.9365079 | 0.1923077 | 0.5644078 |
| LDA | 0.6853933 | 0.9523810 | 0.0384615 | 0.4954212 |
| QDA | 0.5617978 | 0.4126984 | 0.9230769 | 0.6678877 |
| Loess | 0.6629213 | 0.8571429 | 0.1923077 | 0.5247253 |
| kNN | 0.7078652 | 0.9365079 | 0.1538462 | 0.5451770 |
| Random Forest (rf) | 0.6853933 | 0.8571429 | 0.2692308 | 0.5631868 |
| Ensemble | 0.7303371 | 0.9523810 | 0.1923077 | 0.5723443 |

## Neural Network

Neural Networks are a strong tool for predictions, we use the "neuralnet" package for this analysis. First we bind our Liver data with the scaled features into a data frame. Change the class of liver into numeric and for a simpler neural network change the value of 2 into 0 in the liver. This to have a single neuron as output. We make the same data partition like the other method into a test and train set. The neural net is constructed with 2 hidden layer with 10 and 5 neurons respectively.

```
#Bind x_scaled with patients$liver in a dataframe
patients_nn <- cbind(patients$Liver,as.data.frame(x_scaled))
patients_nn <- patients_nn %>% rename(Liver=`patients$Liver`)

#Change values of liver into binary and numeric
patients_nn <- patients_nn %>% mutate(Liver=as.numeric(Liver))
patients_nn$Liver[which(patients_nn$Liver==2)] <- 0

# Creating data partition
set.seed(1,sample.kind = "Rounding")
test_index <- createDataPartition(patients_nn$Liver, time=1, p=0.15,list=FALSE)
test <- as.data.frame(patients_nn[test_index,])
train <- as.data.frame(patients_nn[-test_index,])

#Neural Network method
```

```
nn <- neuralnet(Liver~Age+Total_Bilirubin+Direct_Bilirubin+Alkaline_Phosphotase+
                Alamine_Aminotransferase+Aspartate_Aminotransferase+
                Total_Protiens+Albumin+
                Albumin_and_Globulin_Ratio+num_gender,
                data=train, hidden=c(10,5),act.fct = "logistic",linear.output = FALSE)
```

Finally we make prediction with the neural network and stored in the table of methods without genre for comparison. The next code show this process.

```
#Predict Values with Neural Netwokr
nn_pred <- compute(nn,test[,2:11])
nn_pred <- ifelse(nn_pred$net.result<0.5,1,0)
#Show Results of Accuracy, Sensitivity and Specificity
cm <- confusionMatrix(data=as.factor(nn_pred),reference=as.factor(test[,1]))
results <- bind_rows(results,data.frame(Method = "Neural Network",
                                        Accuracy = cm$overall["Accuracy"],
                                        Sensitivity=cm$byClass["Sensitivity"],
                                        Specificity=cm$byClass["Specificity"],
                                        Balanced_Accuracy= cm$byClass["Balanced Accuracy"]))
```

## Results

These are the results the three methods used in the analysis. In the first table whe observe the neural network results:

| Method | Accuracy | Sensitivity | Specificity | Balanced_Accuracy |
|---|---|---|---|---|
| K-Means | 0.6067416 | 0.8412698 | 0.0384615 | 0.4398657 |
| Logistic Regression (glm) | 0.7303371 | 0.9523810 | 0.1923077 | 0.5723443 |
| LDA | 0.7415730 | 1.0000000 | 0.1153846 | 0.5576923 |
| QDA | 0.5617978 | 0.4126984 | 0.9230769 | 0.6678877 |
| Loess | 0.6629213 | 0.8571429 | 0.1923077 | 0.5247253 |
| kNN | 0.6741573 | 0.8730159 | 0.1923077 | 0.5326618 |
| Random Forest (rf) | 0.6404494 | 0.8095238 | 0.2307692 | 0.5201465 |
| Ensemble | 0.7303371 | 0.9523810 | 0.1923077 | 0.5723443 |
| Neural Network | 0.2840909 | 0.5714286 | 0.1940299 | 0.3827292 |

| Method_Gender | Accuracy | Sensitivity | Specificity | Balanced_Accuracy |
|---|---|---|---|---|
| K-Means | 0.6067416 | 0.8412698 | 0.0384615 | 0.4398657 |
| Logistic Regression (glm) | 0.7191011 | 0.9365079 | 0.1923077 | 0.5644078 |
| LDA | 0.6853933 | 0.9523810 | 0.0384615 | 0.4954212 |
| QDA | 0.5617978 | 0.4126984 | 0.9230769 | 0.6678877 |
| Loess | 0.6629213 | 0.8571429 | 0.1923077 | 0.5247253 |
| kNN | 0.7078652 | 0.9365079 | 0.1538462 | 0.5451770 |
| Random Forest (rf) | 0.6853933 | 0.8571429 | 0.2692308 | 0.5631868 |
| Ensemble | 0.7303371 | 0.9523810 | 0.1923077 | 0.5723443 |

We observe that the neural network gives the worst prediction with a balanced accuracy of 0.3827292, this implies that our constructed neural network it wrong or this isn't the right approach for this analysis. The best balanced accuracy and specificity are from QDA this indicates that each class are normally distributed

and it is corroborated with LDA which gives a sensitivity of one. We observe than genre isn't really a factor that helps improve our predictions since it gives the same results in the methods with higher balanced accuracy. The higher accuracy is given by the ensemble of the methods, with a value of 0.7303371, this isn't high enough to declare a good prediction and we can ignore the balanced accuracy since we have more data of patients with liver disease.

# Conclusion

We can declare that this methods aren't the right approach since the data of patients with and without liver disease are too similar. There's no clear distinction between the outcomes, so this methods can correctly identify results. The results of the algorithms applied can't be considered to be implemented in practical situations since its accuracy is too low.

The next thing to do is to explore more deeply linear regression analysis since the results given by QDA and LDA indicates that the class have a normal distribution and this could help improve the predictions. Another thing to help improve the accuracy could be collect more data or add more features than can help distinguish the differences between patients with and without liver disease.