Logbook

Keimpe Dijkstra

14-9-2022

Contents

Cirrhosis prediction

The dataset

The dataset consists of 424 Primary biliary cholangitis patients who where on a trial for D-penicillamine which was placebo controlled. 312 of the instances contain complete data as these patients have participated the trial, the remaining individuals did not participate in the trial but consented to have basic measurements taken.

Loading the data

Reading the data used for the project.

```
data <- read.csv("./data/cirrhosis.csv")
```

Codebook

```
codebook <- read.csv("./data/codebook.csv", header = T, sep = ";", row.names = 1)
kable(codebook)
```

	unit	type	description
ID		int	unique identifier
N_Days	days	int	number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986
Status		str	status of the patient C (censored), CL (censored due to liver tx), or D (death)
Drug		boolean	type of drug D-penicillamine or placebo
Age	days	int	age in days
Sex		str	Â M (male) or F (female)
Ascites		boolean	presence of ascites N (No) or Y (Yes), the accumulation of fluid in the peritoneal cavity
Hepatomegaly		boolean	presence of hepatomegaly N (No) or Y (Yes), an enlarged liver
Spiders		boolean	presence of spiders N (No) or Y (Yes), spider veins are swollen, twisted veins that usually appear on the legs
Edema		str	Â presence of edema N (no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy)
Bilirubin	mg/dl	float	serum bilirubin
Cholesterol	mg/dl	float	serum cholesterol
Albumin	gm/dl	float	albumin
Copper	ug/day	float	urine copper
Alk_Phos	U/liter	float	alkaline phosphatase
SGOT	U/ml	float	serum glutamic-oxaloacetic transaminase, SGOT is a protein made by liver cells
Tryglicerides	mg/dl	float	triglicerides, a type of fat (lipid) found in your blood
Platelets	ml/1000	float	platelets per cubic, the cells that circulate within our blood and bind together when they recognize damaged blood vessels.
Prothrombin	seconds	float	prothrombin time in seconds, a test to evaluate blood clotting
Stage		int	histologic stage of disease (1, 2, 3, or 4)

Research question

"Can the stage of cirrhosis be determined with basic measurements and non-invasive testing of the blood and urine using machine learning?".

Setup of the project

For this project we constructed a machine-learning (ML) algorithm that tries to determine the stage, and thus the severity of cirrhosis in PBC patients. This will involve supervised learning where the class variable is the stage of the disease

Data exploration and preprocessing

Basic statistics

Numeric data

Summary

First off we will create a summary for all numeric attributes to have a look at their distribution.

```
summary(data[,c(2,5,11,12:19)])
```

```
##      N_Days          Age          Bilirubin       Cholesterol
##  Min.   :  41   Min.   : 9598   Min.   : 0.300   Min.   : 120.0
##  1st Qu.:1093   1st Qu.:15644   1st Qu.: 0.800   1st Qu.: 249.5
##  Median :1730   Median :18628   Median : 1.400   Median : 309.5
##  Mean   :1918   Mean   :18533   Mean   : 3.221   Mean   : 369.5
##  3rd Qu.:2614   3rd Qu.:21273   3rd Qu.: 3.400   3rd Qu.: 400.0
##  Max.   :4795   Max.   :28650   Max.   :28.000   Max.   :1775.0
##                                                  NA's   :134
##     Albumin          Copper          Alk_Phos           SGOT
##  Min.   :1.960   Min.   :  4.00   Min.   :  289.0   Min.   : 26.35
##  1st Qu.:3.243   1st Qu.: 41.25   1st Qu.:  871.5   1st Qu.: 80.60
##  Median :3.530   Median : 73.00   Median : 1259.0   Median :114.70
##  Mean   :3.497   Mean   : 97.65   Mean   : 1982.7   Mean   :122.56
##  3rd Qu.:3.770   3rd Qu.:123.00   3rd Qu.: 1980.0   3rd Qu.:151.90
##  Max.   :4.640   Max.   :588.00   Max.   :13862.4   Max.   :457.25
##                  NA's   :108      NA's   :106       NA's   :106
##  Tryglicerides      Platelets      Prothrombin
##  Min.   : 33.00   Min.   : 62.0   Min.   : 9.00
##  1st Qu.: 84.25   1st Qu.:188.5   1st Qu.:10.00
##  Median :108.00   Median :251.0   Median :10.60
##  Mean   :124.70   Mean   :257.0   Mean   :10.73
##  3rd Qu.:151.00   3rd Qu.:318.0   3rd Qu.:11.10
##  Max.   :598.00   Max.   :721.0   Max.   :18.00
##  NA's   :136      NA's   :11      NA's   :2
```

We notice a lot of missing values due to a large number of patients not being involved in the whole screening but just provided basic measurements. The extend to which these missing values will influence the accuracy of the model to be produced is large when about twenty percent of the patients miss a value. For that reason the choice has been made to for the time being not remove any instances based on missing values.

Log transformation

If we look at the summaries we see differences between maximum- and mean values in the order of hundreds. Therefore, to reduce the amount of skewness, it is wise to log-transform the data to make our analysis more valid.

```
data[,c(11,12:19)] <- log(data[,c(11,12:19)])
```

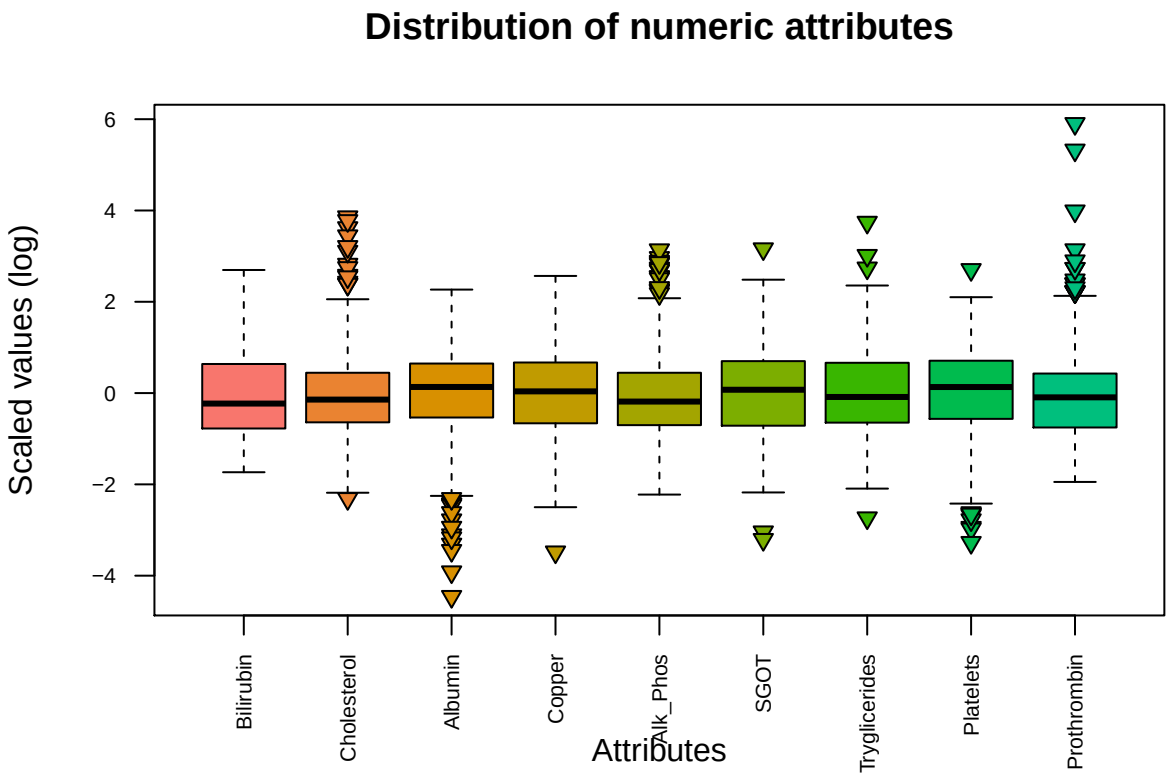
Normilization

To compare the numeric attributes with each other we first have to normalize the data, we do this by scaling. The boxplot below shows all the normalized numeric values in comparison to each other.

```
data[,c(11,12:19)] <- lapply(data[,c(11,12:19)], scale)
```

Distribution of numeric attributes

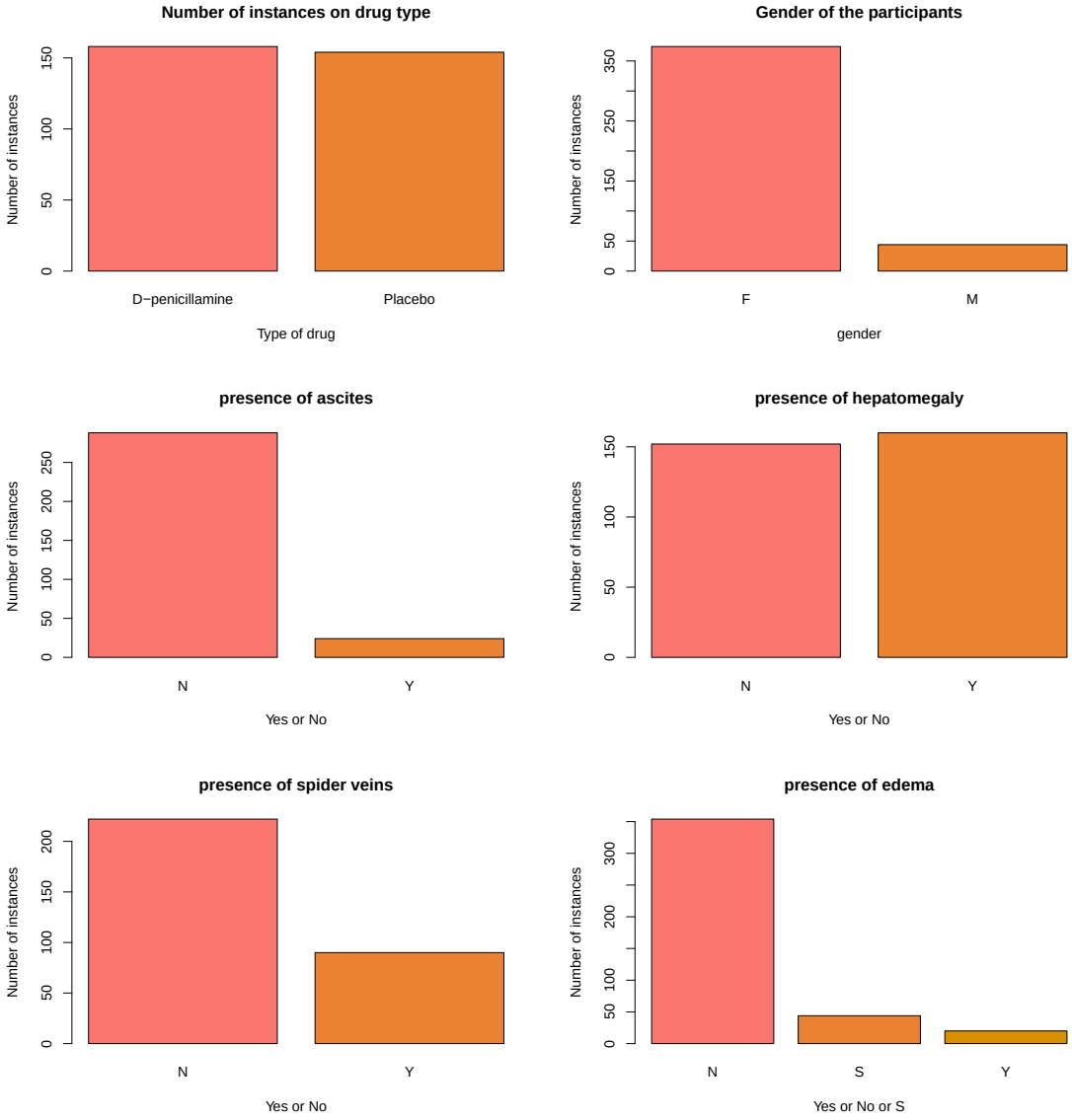
```
boxplot(data[,c(11,12:19)], las = 2, col = hue_pal()(20),outpch = 25, outbg = hue_pal()(20), cex.axis =
```



We see a number off nummbers that remain skewed with either outliers on the left or right side of the distribution. The albumin and platelet attributes seem to be skewed to the left and the alkaline phosphate and prothrombin attributes seem to be skewed to the right eventhough the data has been logscaled. This does not mean the data is of less quality, we hope for an scenario in which these outliers may help catagorize the patients in the cirrhosis stages.

Nominal data

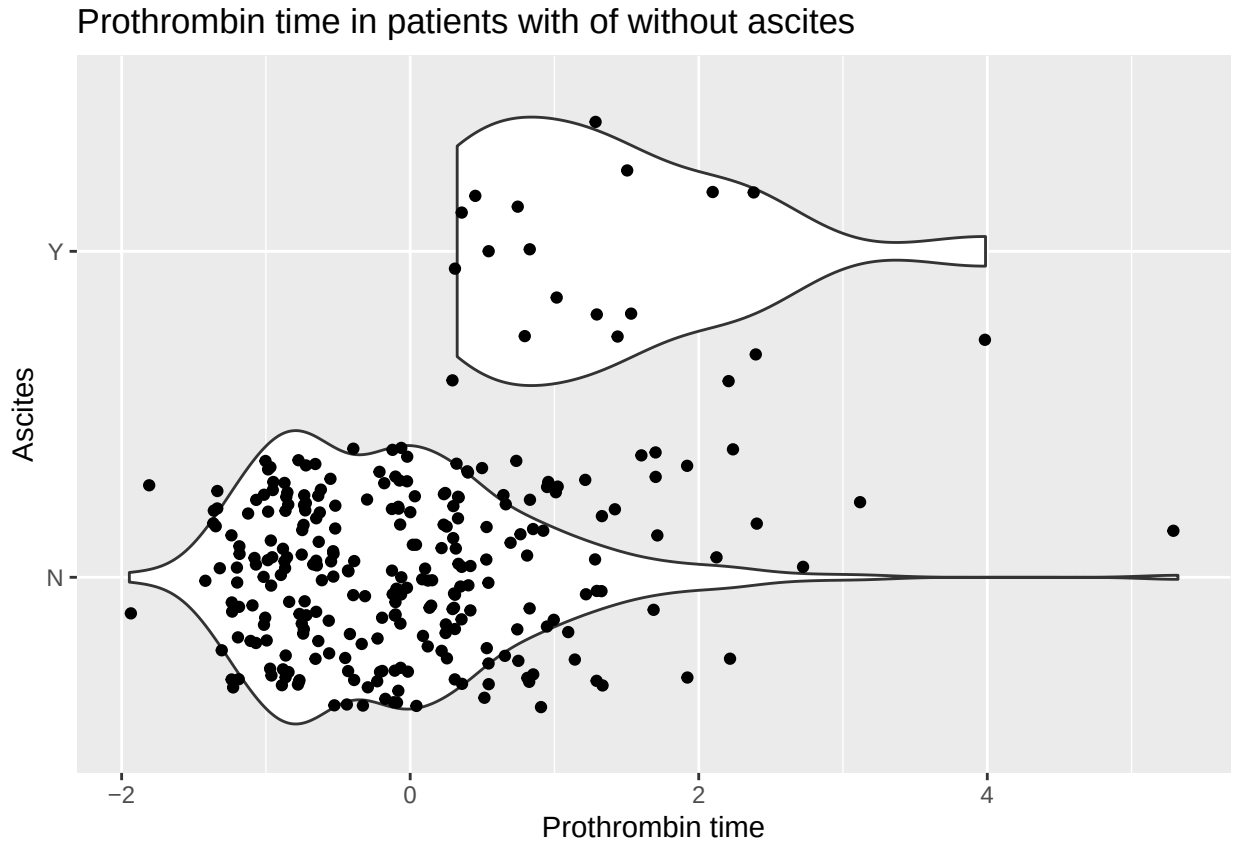
To get a sense of the nominal attributes data we create a barplot for all types.



The plots that draw our attention are the ones that have a big difference in the number of instances between groups. We notice that the patient pool primarily consist of women, and that most of the patients do not suffer from ascites or edema. As expected the difference between patients on placebo and the actual drug is about the same, just as the distribution of patients that suffer from hepatomegaly.

Correlations

Protthrombin time and the prevalence of ascites From the literature we can derive that a prolonged prothrombin time is a useful value to identify the prevalence of ascites. Therefore we plot the prothrombin time against the the presence of ascites in a violin plot.



We see that the patients who are not diagnosed with ascites tend to have a lower prothrombin time then their diagnosed counterparts. But we also see a lot of instances of patients with no ascites in the range of patients with ascites. This is due to the two samples being very unequal (see fig [num]), therefore we preform a welch t-test to see whether the difference is significant.

```
t.test(subset(data, Ascites == "Y", select = Prothrombin), subset(data, Ascites == "N", select = Prothr
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(data, Ascites == "Y", select = Prothrombin) and subset(data, Ascites == "N", select = P
## t = 5.8423, df = 26.678, p-value = 3.358e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7873665 1.6405446
## sample estimates:
##   mean of x   mean of y
##  1.11467731 -0.09927824
```

From the output we can see that the t test-statistic is 5.8423 and the corresponding p-value is 3.358e-06. Since this p-value is less than .05, we can reject the null hypothesis and conclude that there is a statistically significant difference in mean scores between the two groups. So the prothrombin time might be an useful indicator of the presence of acsites and then maybe even in our model for determining the stage of cirrhosis.

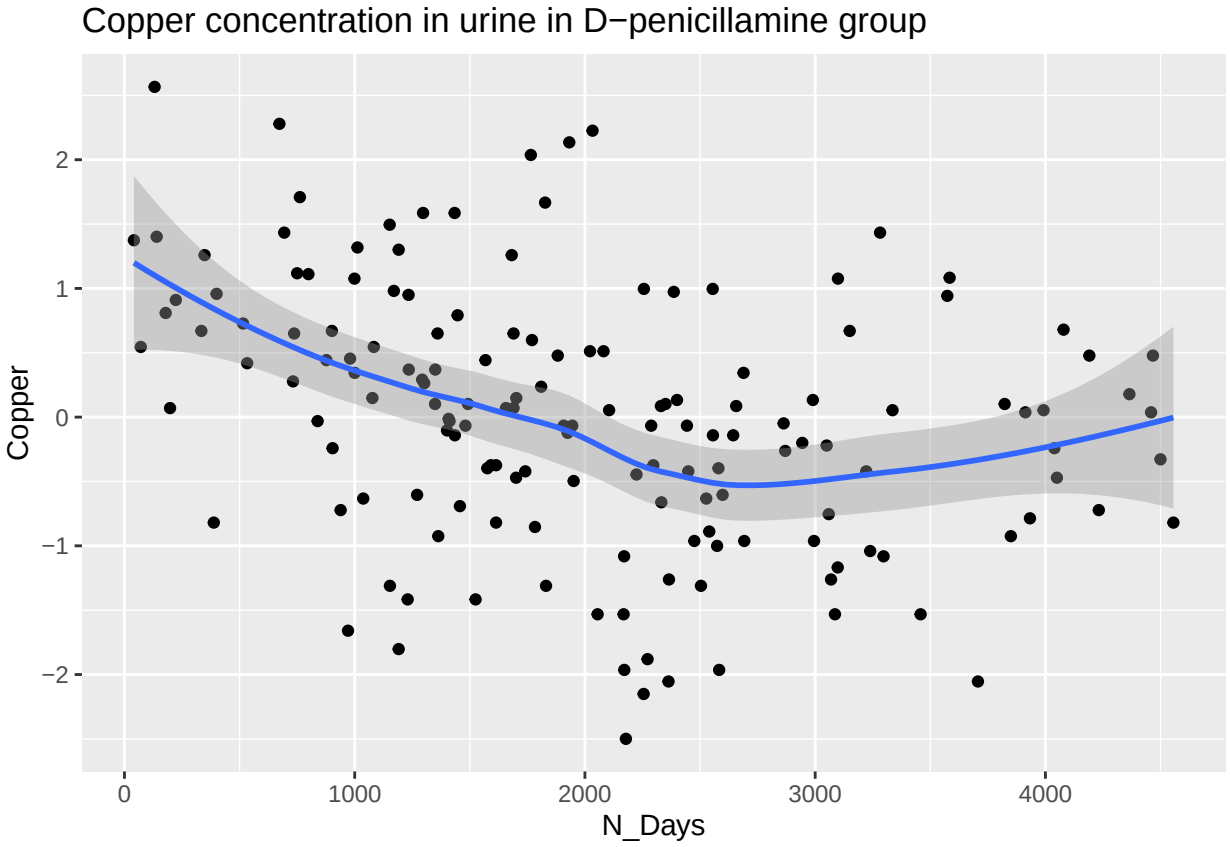
Copper and D-penicillamine

The D-penicillamine drug is a chelating agents, meaning it binds heavy metals making it possible to excrete them through urine. To see whether the drug affects the concentration of copper in urine we plotted the copper concentration against the time between registration and analysis.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

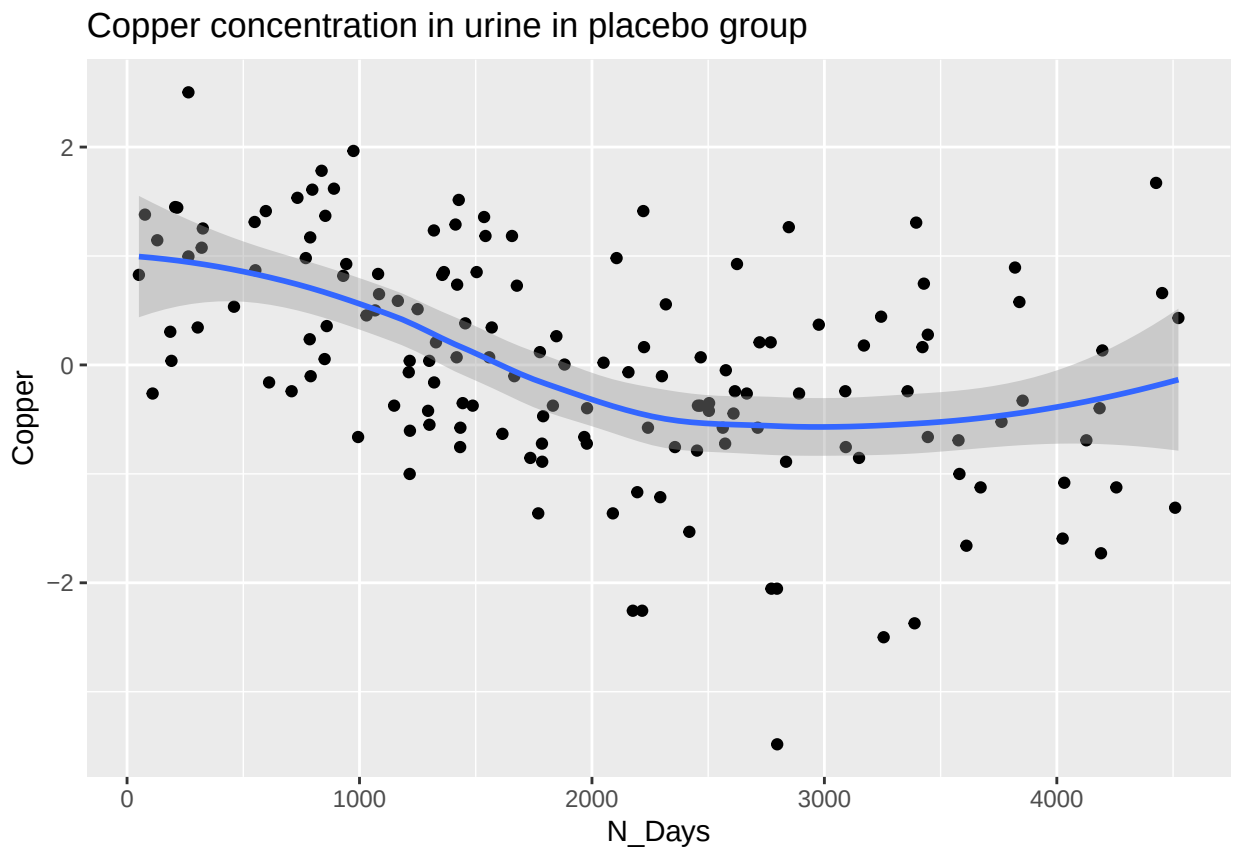
```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
## Removed 1 rows containing missing values (geom_point).
```

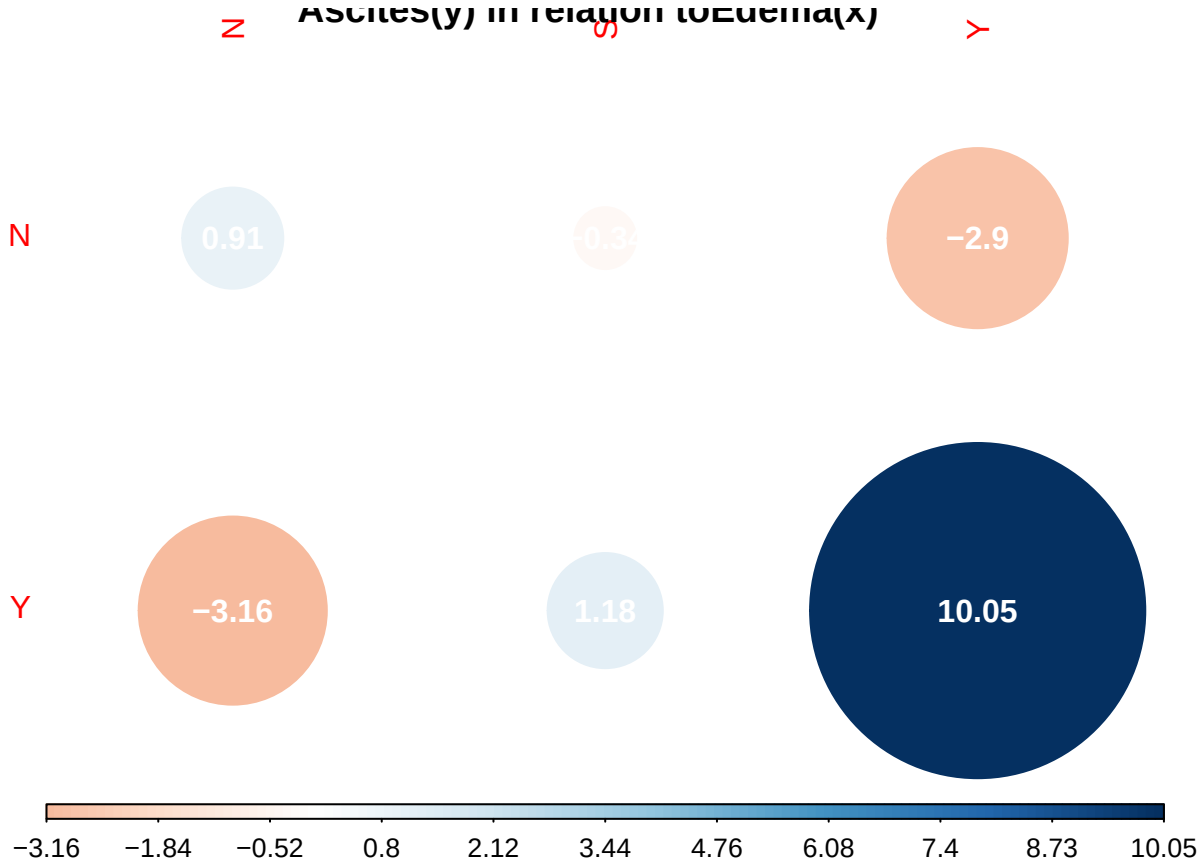


Although the data points are a bit scattered the trend seems to follow a very similar line. This is useful information because if it were to affect the concentration it might create a bias in our ML model.

Ascites and edema

```
## Warning in chisq.test(ct): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  ct
## X-squared = 121.71, df = 2, p-value < 2.2e-16
```



The plot above shows the relationship between the presence of ascites (plotted on the x-axis) and the presence of edema. When a number is below zero it means the values repel, they tend not to occur in the same instance. The opposite is true for positive numbers, these values do tend to occur together. Most noticeable is the big blue ball in the bottom right. This means when fluid builds up in the peritoneal cavity in PBC patients it usually means fluid also builds up in other parts in the body, causing edema. This is most prevalent in edema patients who did not react to treatment with diuretics. We can draw the same conclusions from the chi-squared test, returning us a p-value smaller than 2.2e-16.

Class variables

Missing class variables

Since the project uses supervised learning techniques, the rows without a class variable, eg: NA's, serve no purpose for training the algorithm and can thus be removed.

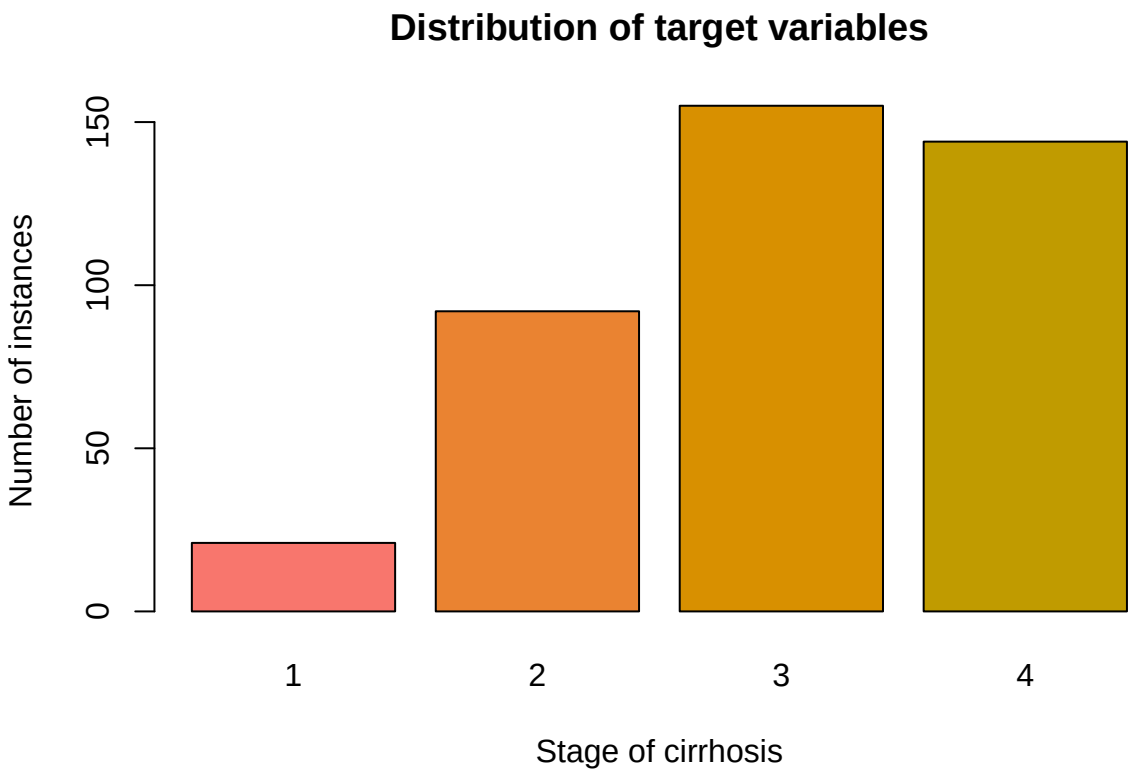
```
sum(is.na(data$Stage)) #checking the amount of na's in the stage collumn
```

```
## [1] 6
```

```
data <- data[!is.na(data$Stage),] #drop the rows with no class variable
```

We identified six rows with no class variable, exactly as the description of the dataset noted. We removed the six instances.

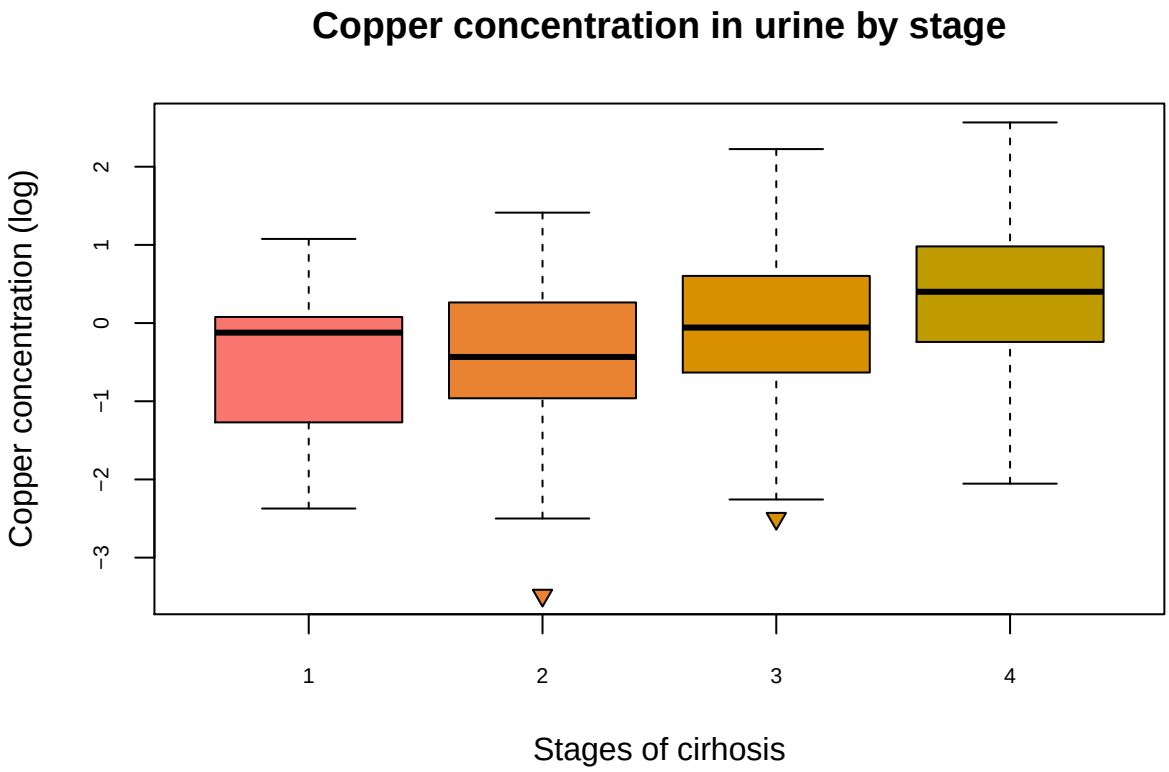
Distribution of class variables



Because early stage cirrhosis is very hard to diagnose there are significantly less patients labeled with stage one compared to the other stages. Ideally we would like for all the stages to have about the same amount of instances, the level to which the model will be affected by the skewed data remains to be seen.

Significant differences between attributes

Copper



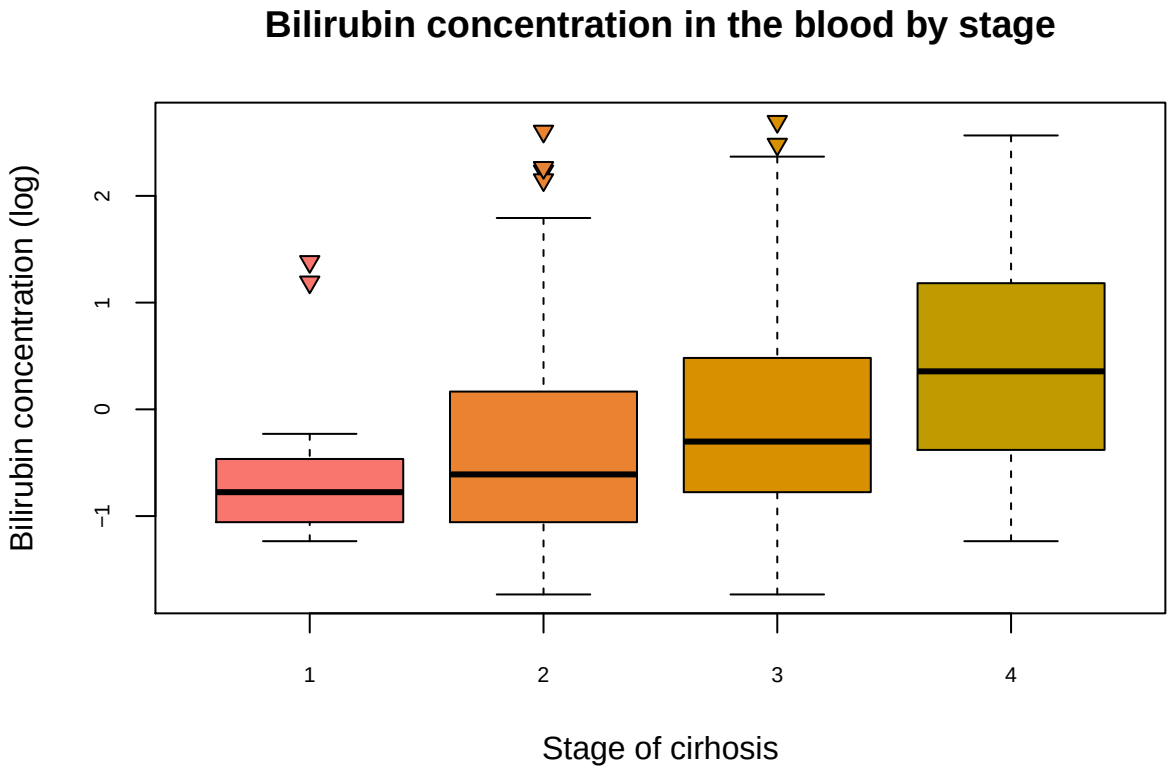

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## factor(Stage)   3  28.86   9.620    10.5 1.47e-06 ***
## Residuals     272 249.17   0.916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Copper ~ factor(Stage), data = na.omit(data))
##
## $`factor(Stage)`
##          diff         lwr       upr     p adj
## 2-1 0.1374042 -0.64608726 0.9208957 0.9689383
## 3-1 0.5745676 -0.17726679 1.3264020 0.1998138
## 4-1 0.9557620  0.19732392 1.7142000 0.0069029
## 3-2 0.4371634  0.03854371 0.8357830 0.0252546
## 4-2 0.8183577  0.40741860 1.2292969 0.0000030
## 4-3 0.3811944  0.03439905 0.7279897 0.0247919
```

Except for the first stage, copper concentrations tend to increase as the cirrhosis progresses. If we look at the adjusted p-values from our anova and the following Tukey multiple pairwise-comparison, we see

significant differences in all except the first two. The analysis factors in the differences of sample sizes, but because the early stage of cirrhosis is hard to diagnose this amount of data might be insufficient.

Bilirubin



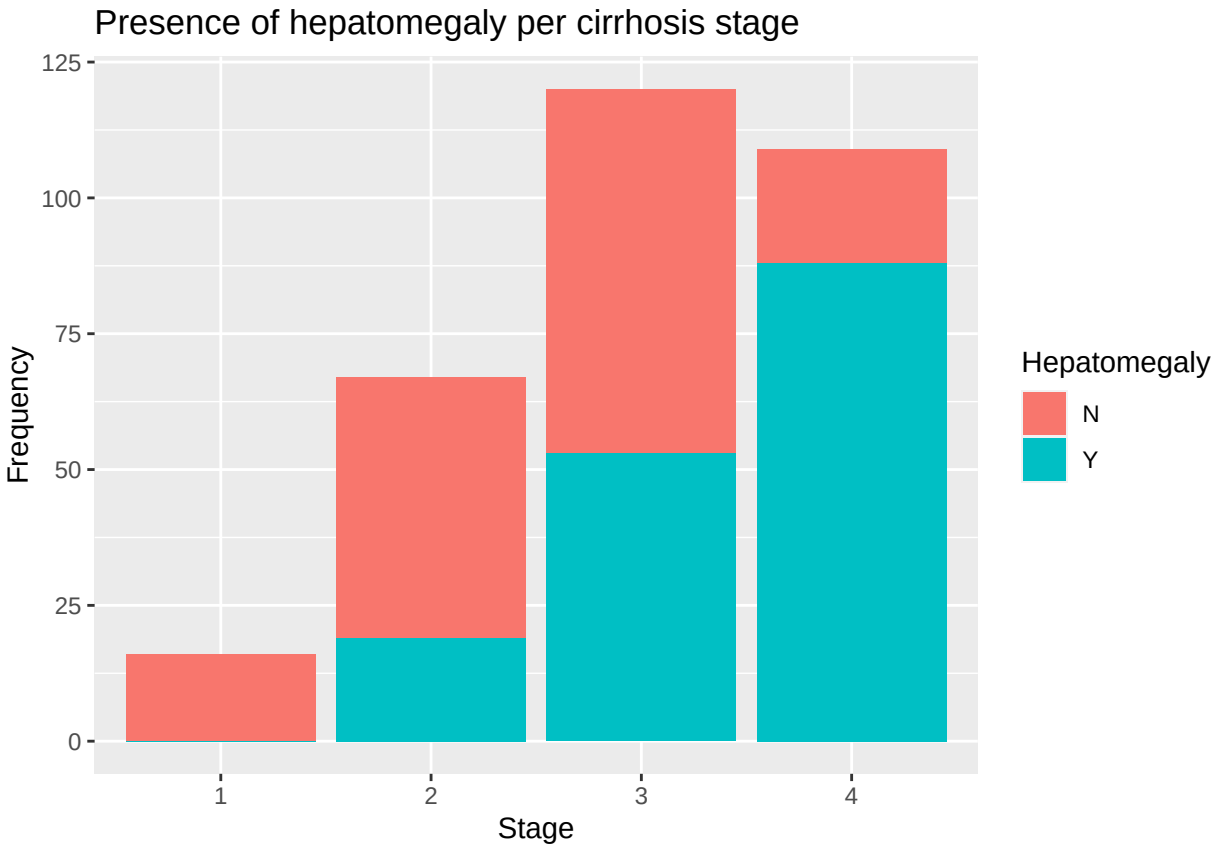

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## factor(Stage)   3  37.24  12.413   13.92 1.81e-08 ***
## Residuals     272 242.56   0.892
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Bilirubin ~ factor(Stage), data = na.omit(data))
##
## $`factor(Stage)`
##          diff         lwr       upr     p adj
## 2-1 0.2818209 -0.49120910 1.0548509 0.7820040
## 3-1 0.6373405 -0.10445511 1.3791360 0.1201415
## 4-1 1.1688809  0.42056981 1.9171920 0.0004066
## 3-2 0.3555196 -0.03777754 0.7488167 0.0923487
## 4-2 0.8870600  0.48160789 1.2925121 0.0000002
## 4-3 0.5315404  0.18937565 0.8737052 0.0004439
```

Bilirubin concentrations also increase with the progress of the hepatic disease. We again notice the first stage only showing significant differences with the fourth stage. But the overall pattern is more distinct and lets us to believe this is a meaningfull correlation.

Hepatomegaly

```
##
##  Pearson's Chi-squared test
##
## data:  ct
## X-squared = 71.211, df = 3, p-value = 2.349e-15
```



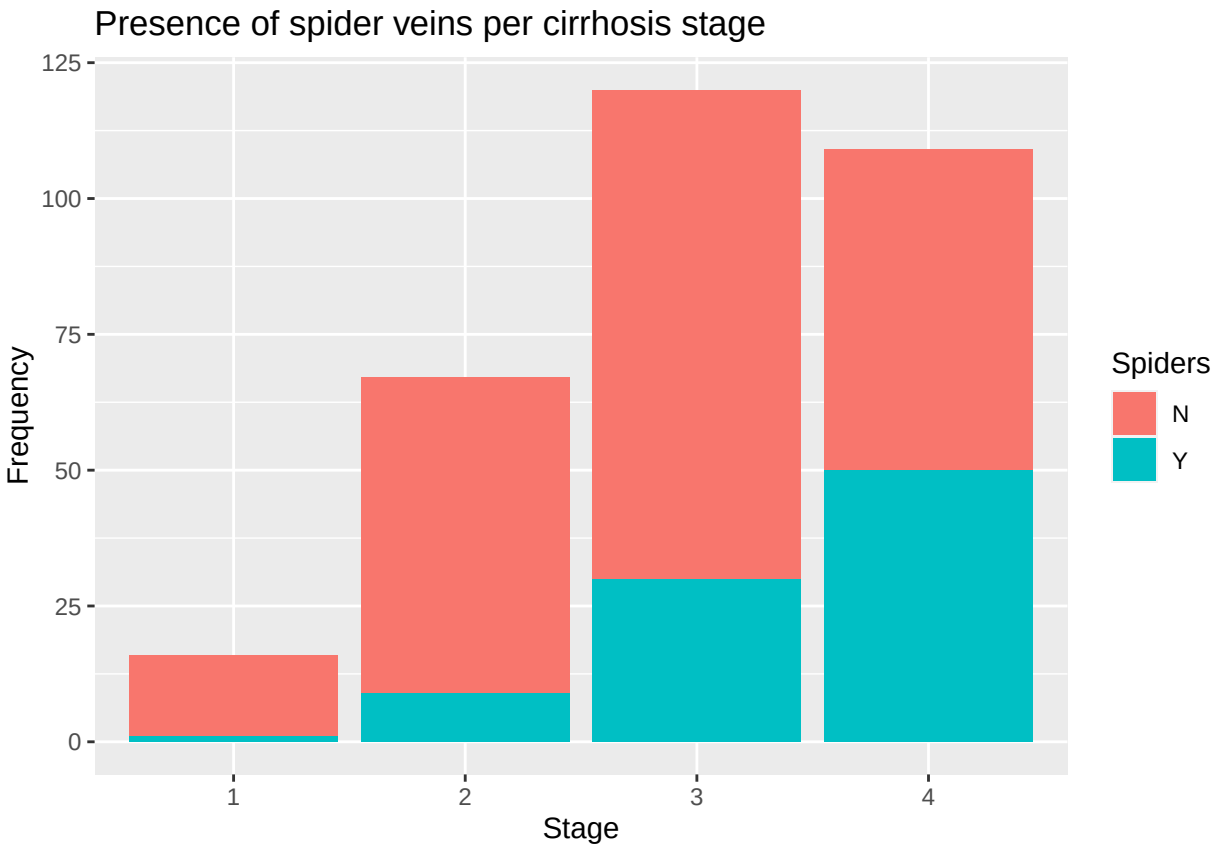
The relationship between hepatomegaly and the stage of cirrhosis is very clear, hepatomegaly becomes more prevalent as you go up the stages. It is makes sense that the liver enlarges while it is scarring, therefore almost all the patients in the last stage of cirrhosis experience swelling of the liver. The chi-squared test confirms with a value of 2.349e-15, far below the bar of significance at 0.05.

Spider veins

```
## Warning in chisq.test(ct): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  ct
## X-squared = 27.993, df = 3, p-value = 3.644e-06
```



Although not as apparent as hepatomegaly, the frequency in which spider veins occur also increases as cirrhosis progresses. Here nearly half of the patients in stage four cirrhosis experience spider veins, where only a handful experiences them in the first stage. Again a very low p-value from the chi-squared test at 3.644e-06.

Clustering analysis

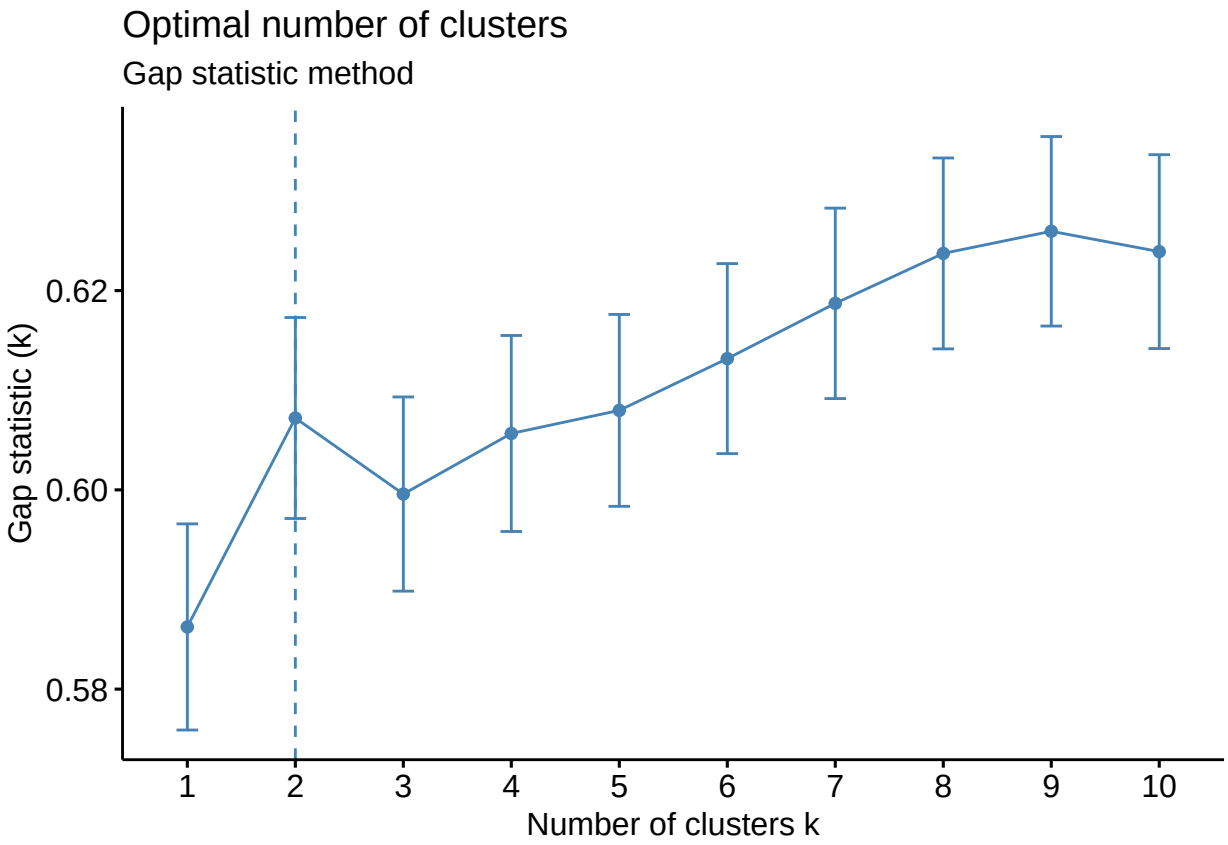
We will preform a k-means clustering on all the numerical data to see if it cat already be categorized in a meaningful way. Before we will we can do the actual clustering we need to determine the amount of clusters we want to divide the data into.

Determining amount of clusters

```
## Warning: did not converge in 10 iterations
```

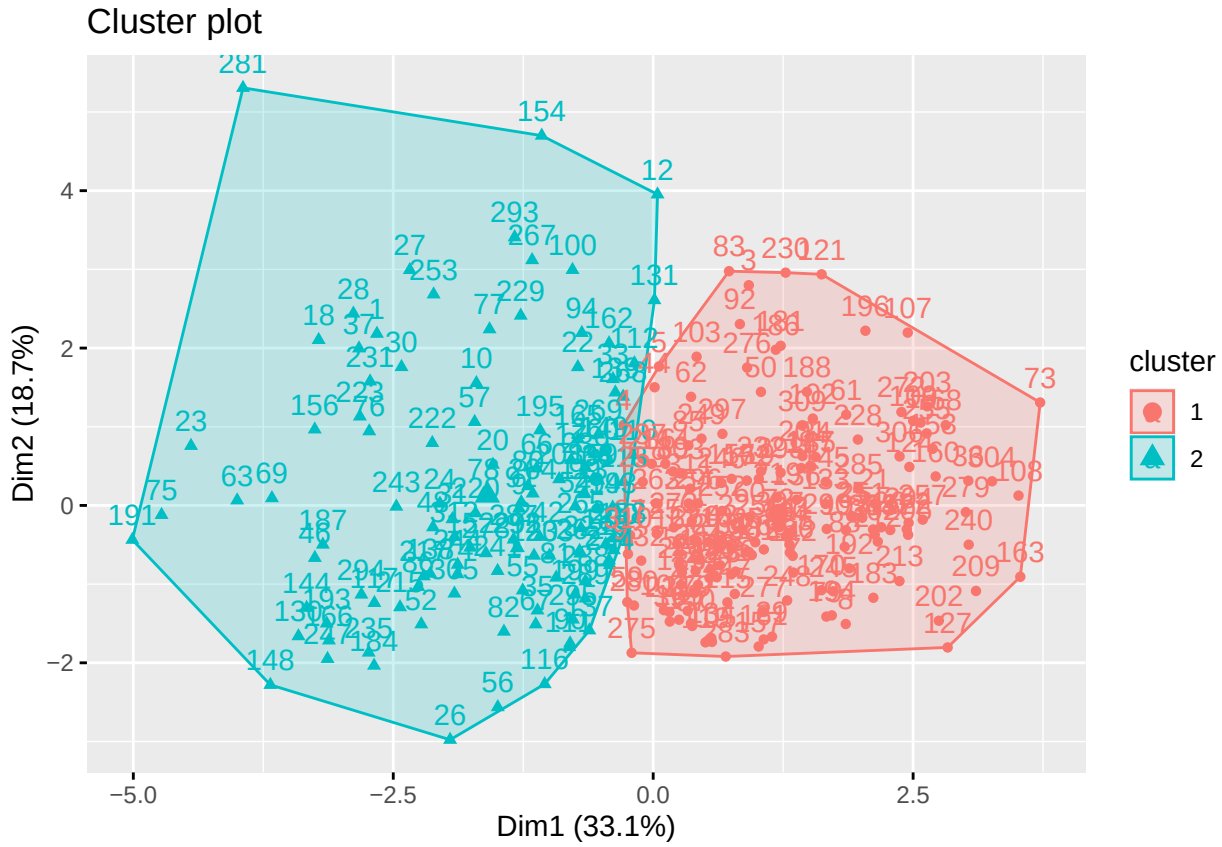
```
## Warning: did not converge in 10 iterations
```

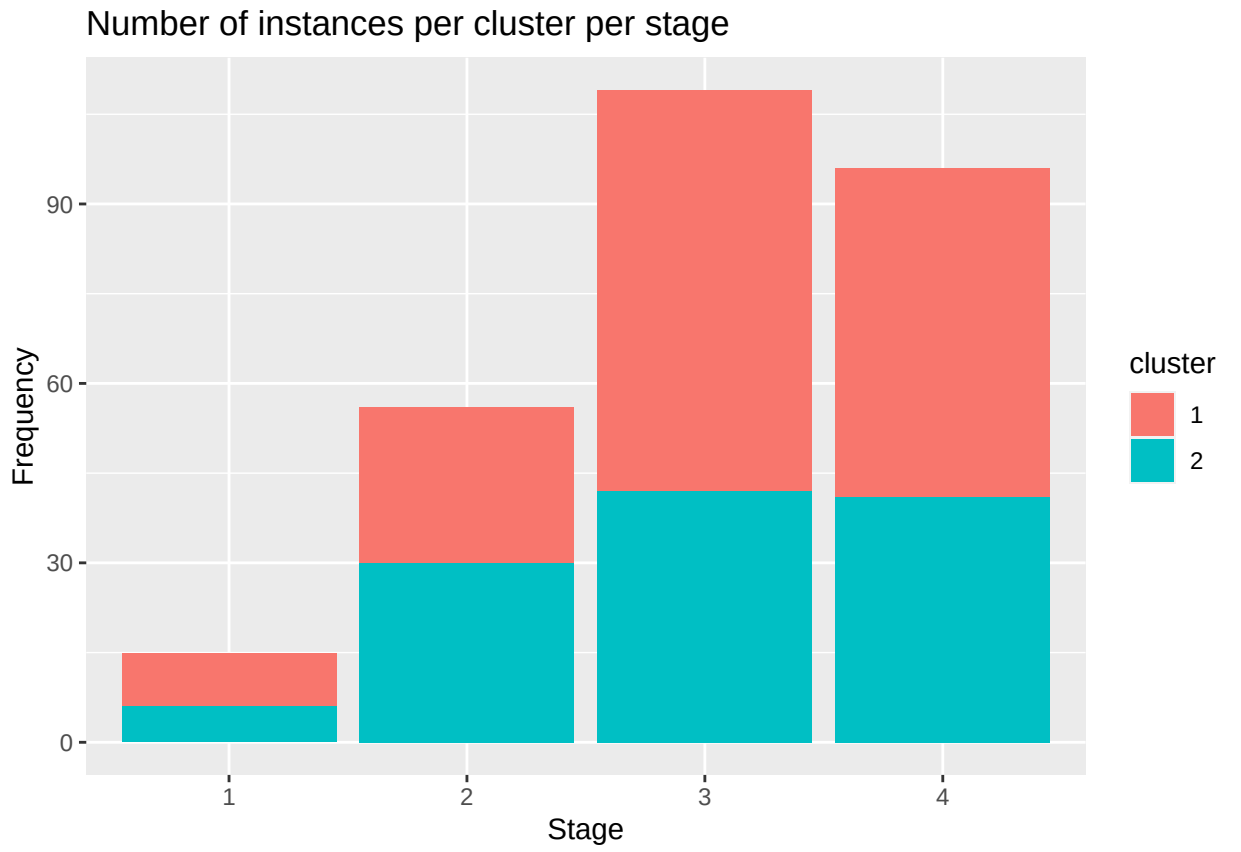
```
## Warning: did not converge in 10 iterations
```



To determine the amount of clusters we used the gap statistic method, the highest value suggest the amount we should choose. We see a peak at two, this means we should choose two clusters

Clustering and visualization

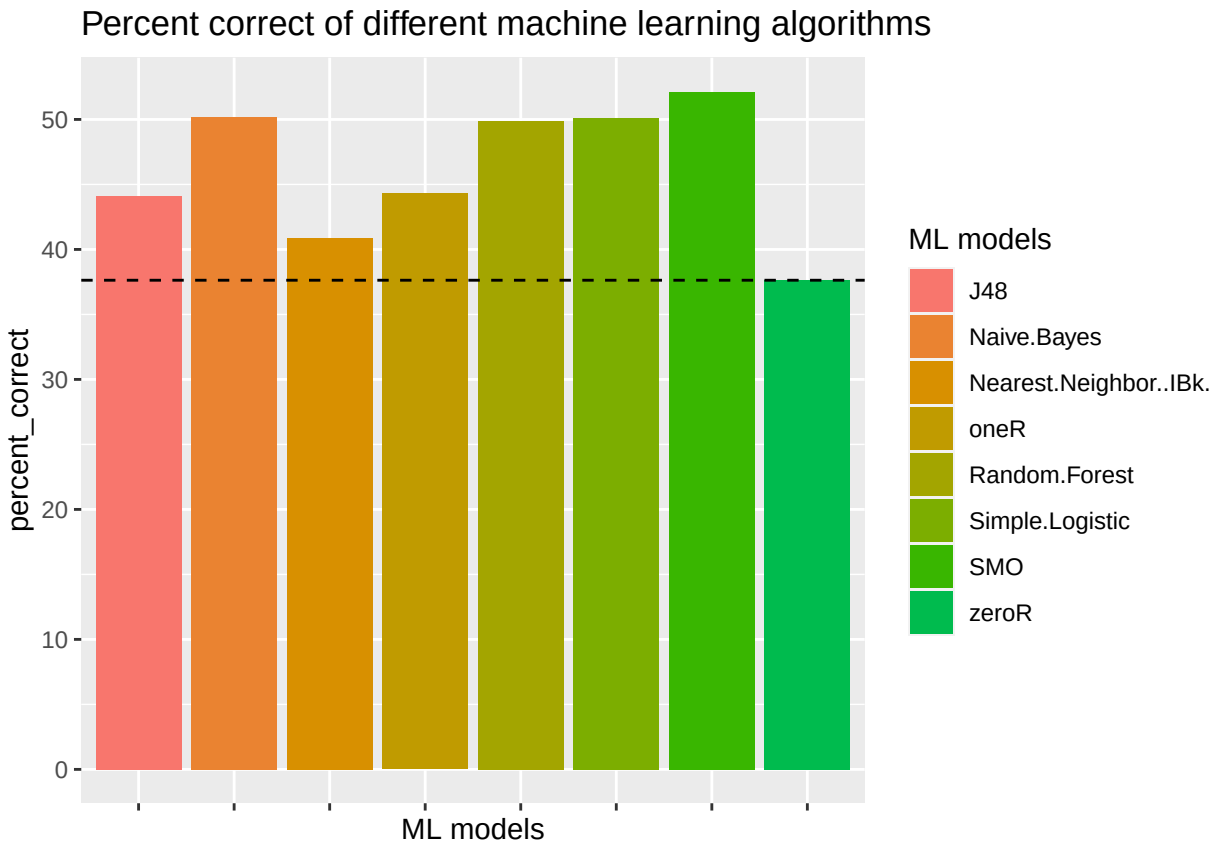


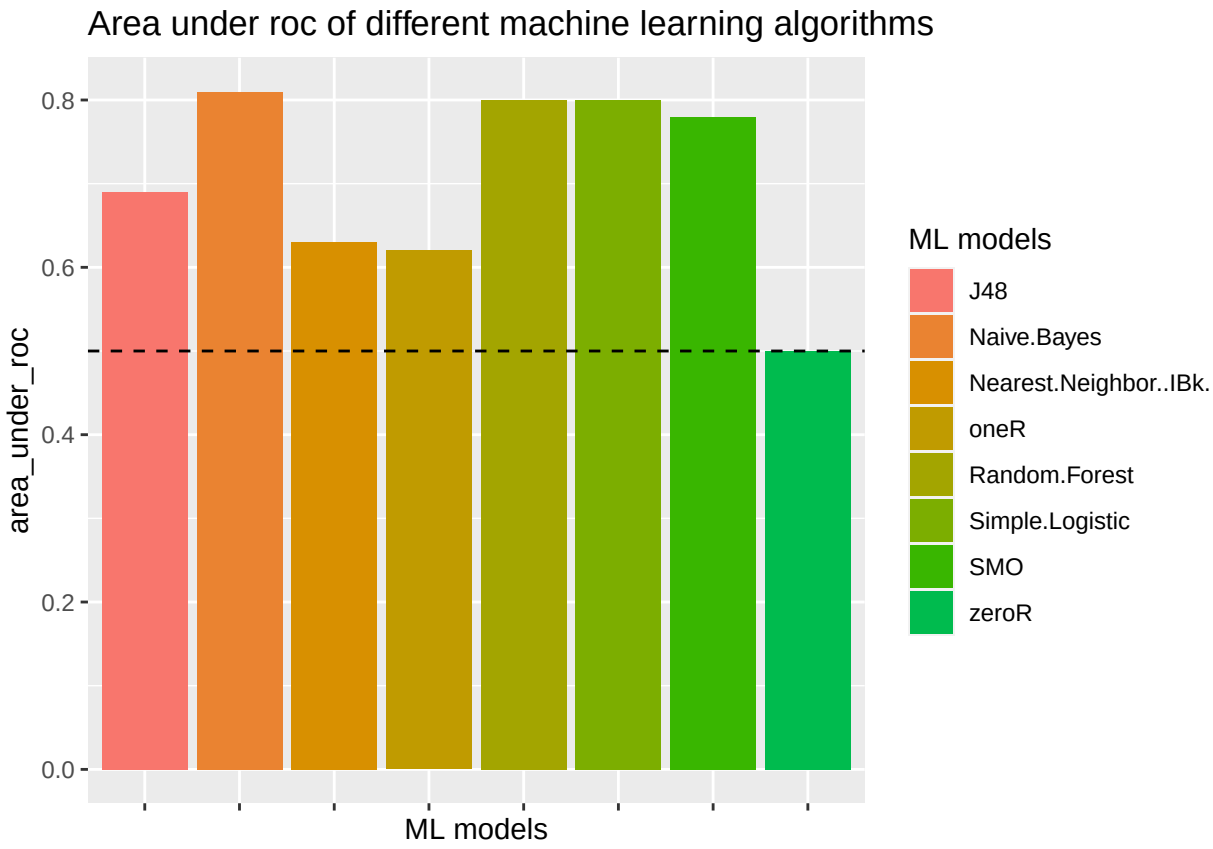


In the first plot we see the clusters and the respective instances plotted. It is hard to draw conclusions from these kinds of plots, but we do see minimal overlap and therefore distinct groups. Looking at the clusters plotted against the different stages in a barplot, we see no obvious correlation between the two. So we can conclude the k-means clustering of the nemeric variables alone cannot give an indication of the stage of cirrhosis

ML model comparison

	zeroR	oneR	Naive.Bayes	Simple.Logistic	SMO	Nearest.Neighbor..IBk	J48	Random.Forest
percent_correct	37.63	44.31 v	50.17 v	50.14 v	52.14 v	40.90	44.13 v	49.91 v
area_under_roc	0.50	0.62 v	0.81 v	0.80 v	0.78 v	0.63 v	0.69 v	0.80 v
TP	0.00	7.30 v	8.86 v	9.72 v	9.51 v	7.74 v	8.63 v	9.47 v
FP	0.00	7.35 v	4.47 v	6.00 v	5.00 v	7.47 v	7.22 v	5.58 v
TN	26.80	19.45 *	22.33 *	20.80 *	21.80 *	19.33 *	19.58 *	21.22 *
FN	14.40	7.10 *	5.54 *	4.68 *	4.89 *	6.66 *	5.77 *	4.93 *


Percent correct of different machine learning algorithms
percent_correct
50
40
30
20
10
0
ML models
ML models
J48
Naive.Bayes
Nearest.Neighbor..IBk.
oneR
Random.Forest
Simple.Logistic
SMO
zeroR



Naive bayes

Comparison of different Naïve bayes algorithms.

	naive_bayes	naive_bayes_simple	bayesnet
percent_correct	50.17	50.25	48.02
area_under_roc	0.81	0.81	0.79

no significant difference

	naive_bayes	bagging	stacking	boosting
percent_correct	50.17	50.78	37.63	50.17
area_under_roc	0.81	0.81	0.50	0.76

SMO

	SMO	bagging	stacking	boosting
percent_correct	52.14	50.34	37.63	51.51
area_under_roc	0.78	0.80	0.50	0.73

Random forest

	random_forest	bagging	stacking	boosting
percent_correct	49.91	49.78	37.63	49.56
area_under_roc	0.80	0.80	0.50	0.79

Simple logistic

	simple_logistic	bagging	stacking	boosting
percent_correct	50.14	49.29	37.63	50.16
area_under_roc	0.80	0.81	0.50	0.77

Choice

We choose to build upon the naïve bayes algorithm. It shows the highest area under the roc and has the second highest percent correct of all the algorithms. Furthermore, none of the meta learners seem to create a significant increase in accuracy.

But the percent correct is really not what is it expected to be. The discrepancy between the area under the roc and the percent correct is explained by the fact that the weka gui by default only show the roc of the last class label. Therefore we will transform the dataset where there will be two class labels, yes and no. Referring to whether cirrhosis is present or not. We do this because determining the exact stage of pbc will be unfeasible with this dataset.

We read the data again, log-transform it and delete the ID and state attributes. We do not have to normalize the data again since we wont be comparing it anymore. Then we alter the class attribute.

Results ML

=== Summary ===

Correctly Classified Instances 313 75.9709 %

Incorrectly Classified Instances 99 24.0291 %

Kappa statistic 0.469

Mean absolute error 0.2726

Root mean squared error 0.4368

Relative absolute error 59.9203 %

Root relative squared error 91.5961 %

Total Number of Instances 412

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,646	0,179	0,660	0,646	0,653	0,469	0,803	0,689	yes
0,821	0,354	0,812	0,821	0,816	0,469	0,803	0,880	no

Weighted Avg. 0,760 0,293 0,759 0,760 0,759 0,469 0,803 0,813

=== Confusion Matrix ===

a b <– classified as

93 51 | a = yes

48 220 | b = no

Above we see results of the naïve bayes algorithm on the altered dataset. It shows a very big improvement over the previous dataset, with an percent correct of almost 76%.

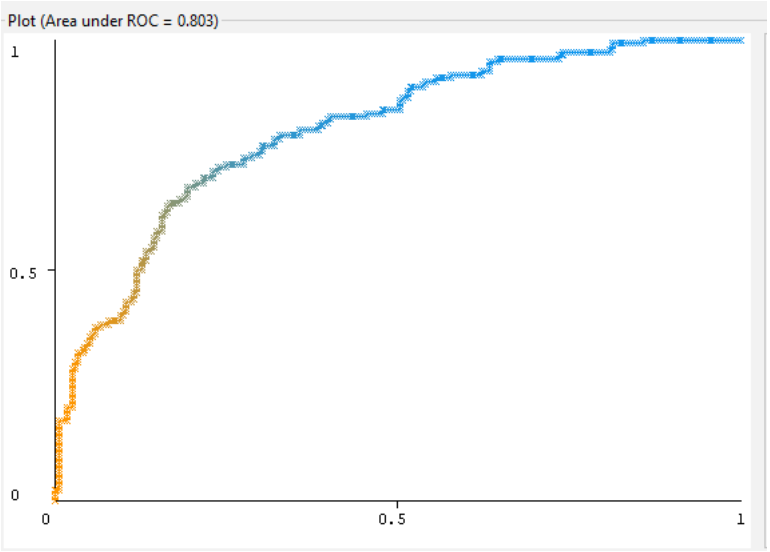


Figure 1: ROC curve of the naïve bayes algorithm