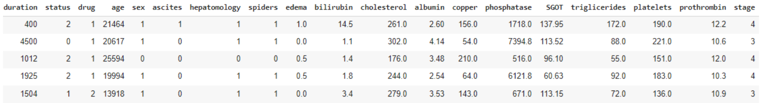
Shallow ANN

Aldo N. Cao Romero

June 2022

1 Introduction

It is presented the data set from the Fleming, T.R. and Harrington, D.P. (1991) Counting Processes and Survival Analysis on the primary billiary Cirrhosis. The Data frame can be seen in Figure (1). Where the shape of the data is 312 rows and 20 columns. Where the encoded variables are as follows: drug: 1=penicillamine, 2=placebo. sex: 0=male, 1=female. presence of ascites, hepatomegaly, spiders, edema: 0=no, 1= yes, (0.5 presence of edema without or resolved by diuretics). The aim of this project is, taking as dependant variable the Status, which is encoded as: 0=censored, 1= tansplantation or 2=death, and use it to predict, base on certain characteristics, if a person is candidate to receive a transplant or not



duration	status	drug	age	sex	ascites	hepatomology	spiders	edema	bilirubin	cholesterol	albumin	copper	phosphatase	SGOT	triglicerides	platelets	prothrombin	stage
400	2	1	21464	1	1	1	1	1.0	14.5	261.0	2.60	156.0	1718.0	137.95	172.0	190.0	12.2	4
4500	0	1	20617	1	0	1	1	0.0	1.1	302.0	4.14	54.0	7394.8	113.52	88.0	221.0	10.6	3
1012	2	1	25594	0	0	0	0	0.5	1.4	176.0	3.48	210.0	516.0	96.10	55.0	151.0	12.0	4
1925	2	1	19994	1	0	1	1	0.5	1.8	244.0	2.54	64.0	6121.8	60.63	92.0	183.0	10.3	4
1504	1	2	13918	1	0	1	1	0.0	3.4	279.0	3.53	143.0	671.0	113.15	72.0	136.0	10.9	3

Figure 1: Original data sit on Cirrhosis disease

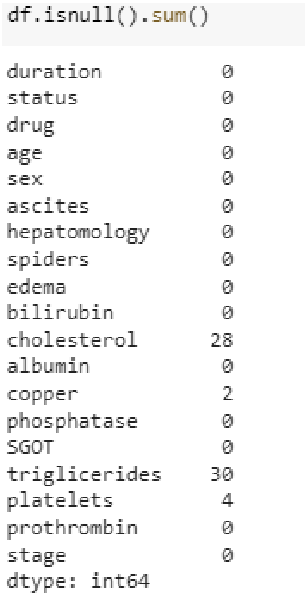
2 Cleaning Data and EDA

In overall the data cleansed enough. However, there were some issues we have to deal with. For example, we have several Null/NaN values (Table 2), and since the amount of the data is small we use some exploratory analysis to see how to deal with these values.

First, we take a look to the box plots of the Cholesterol and Triglycerides columns. It is observed in the cholesterol plot that, in spite of the outliers, the median and the mean are really close, so the null values were replaced them by the median. In contrast with the Triglycerides plot, we use the mean. The rest of the Null values where just dropped, giving a final data shape of 306 rows.

Now, to begin with the exploratory Analysis, let's take a look to the correlation matrix in a heat-map format, figure (4). We see that, in particular the status column has great correlation among several variables, since we have a small amount of data we focused in the variables status is more correlated with. Additionally, we show the distribution of the values that the variable status can have. Figure (5). In this pie chart we principally observe that most of the data is located in the 0 label, that is to say, this information is censored. However we are going to focus ourselves in the possibility of a transplant, i.e., label 1. In this case, we also observe that we have really small data, which can generate problems in the model.

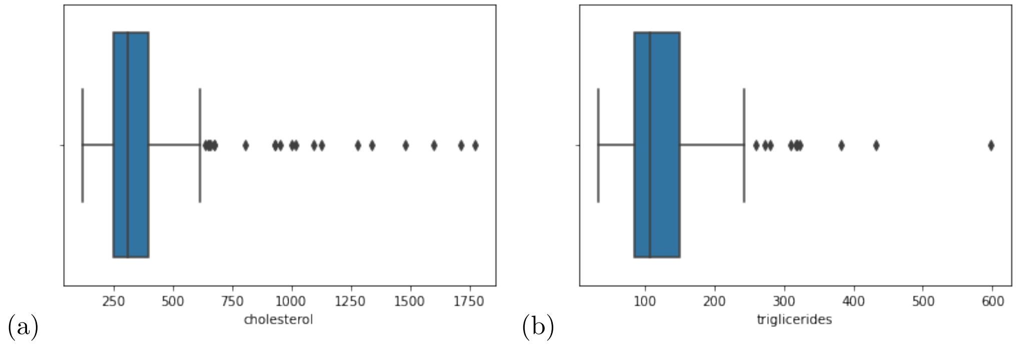
In addition, by observing the box plots of the age and copper variables against status (figure (7)), we infer that a great amount of copper has to do when a person dies or if it can receive an organ, even if the censored status is below the 1 and 2 status, it does have several outliers that can support the idea that most



```
df.isnull().sum()
```

```
duration         0
status           0
drug             0
age              0
sex              0
ascites          0
hepatomology     0
spiders          0
edema            0
bilirubin        0
cholesterol     28
albumin          0
copper           2
phosphatase      0
SGOT             0
triglicerides   30
platelets        4
prothrombin      0
stage            0
dtype: int64
```

Figure 2: Visualization of Null and NaN variables of the data set.



(a) (b)

Figure 3: Box plots of the Cholesterol and Triglycerides columns

of those patients can be death. It happens the same with the age. We see that the majority of the people who receive the organ is just below the threshold of the people who died, and the censored information has similar characteristics to the status 2, which makes us think, once again, that most of those patients that belong to the status 0 can be death.

Finally, we analyse the amount of patients that belong to any of the classes of the status value and the stage of the disease. These results can be seen in figure (8). One hand we observe that most of the who received a transplant were in the stage 3 and 4 of the disease, also, it was expected that most of the people who finally died were in the forth stage of the disease.

3 Creation of the model

The model was trained using the 85% of the total data and the Shallow Artificial Neural Net. The parameters where chosen by plotting each one of them against the accuracy and by taking the best one. The plots (9) show this reasoning. We see, that the pots against epoch and learning rate tend to converge at some point, on the other hand, with the number of neurons, we are not able to see a proper behavior. Anyway, we use the max values of each, this is, neurons=10, Learning Rate= 0.05 and epoch=67010.

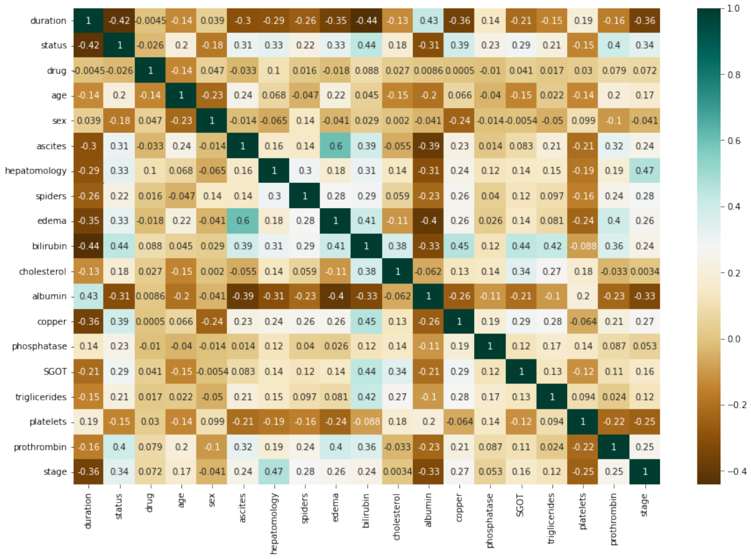
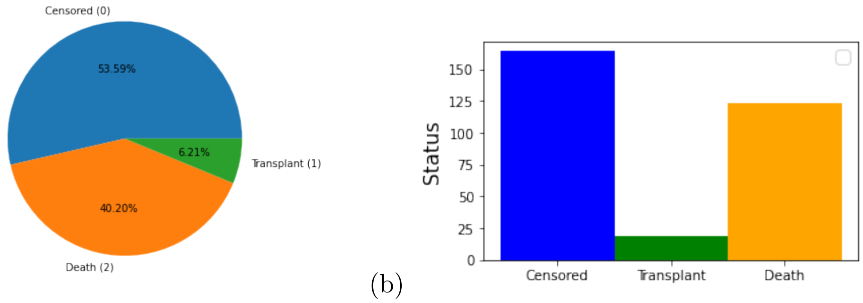
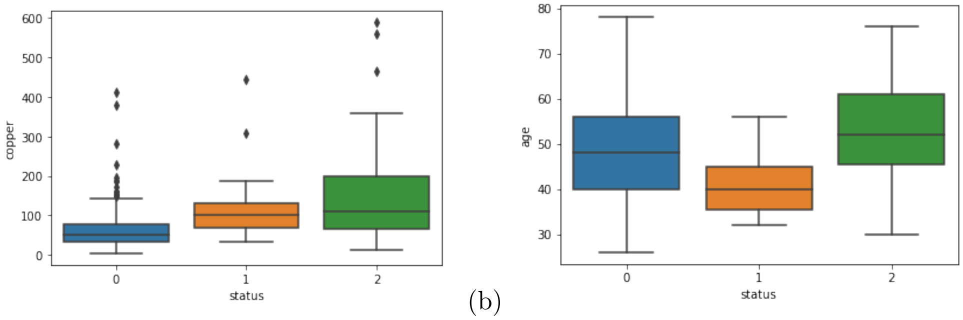


Figure 4: Correlation matrix in a heat-map format.



(a) (b)

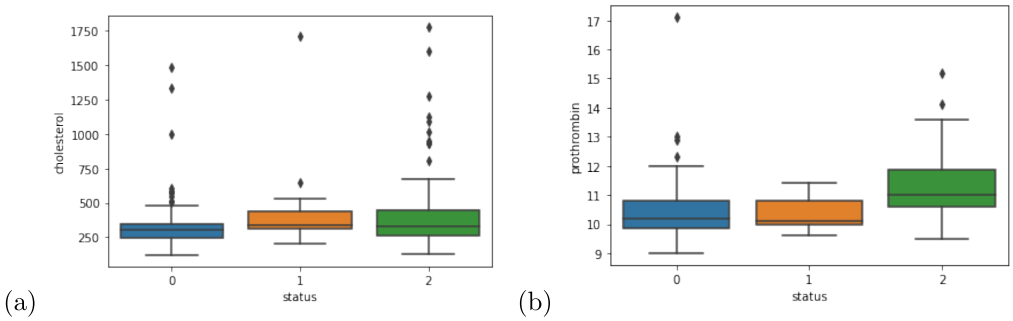
Figure 5: Pie chart and histogram of the status variable.



(a) (b)

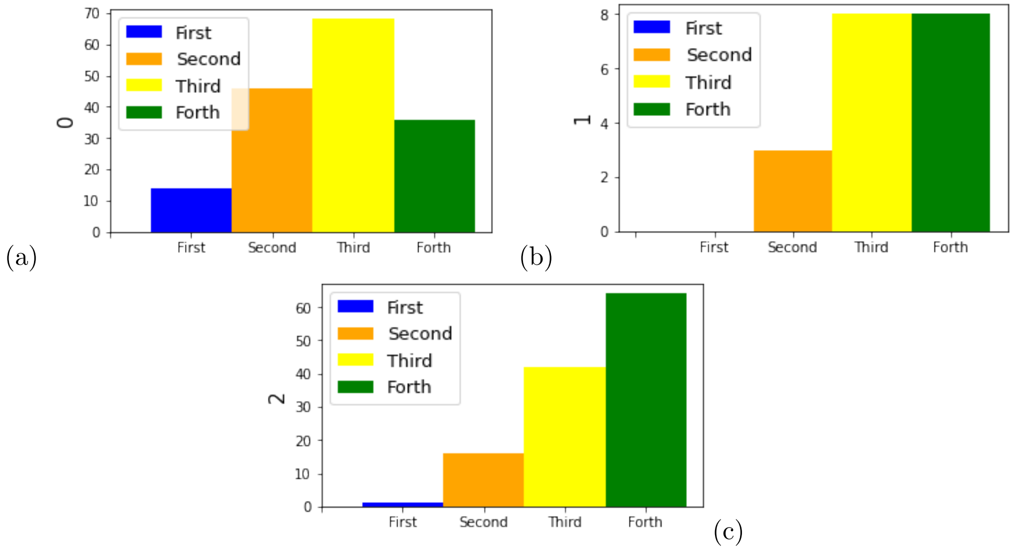
Figure 6: Box plots of the age and copper variables against status

It was got and accuracy of 72% of the training data and 67% for the test data. The results can be seen in figure (11). We observe that the zeros is mainly confused with the other two values, in particular with the class one, due to the lack of data that we have in that case. On the other hand with the class 2, where we obtain some confusions due to the amount of zeros we have in the test data. That is to say, in the test data, as well as in the training, there are more zeros. Also, we show a data frame with the probabilities of each class given by the model. Figure (10). With this we are able to observe why exactly



(a) (b)

Figure 7: Box plots of the cholesterol and prothrombin variables against status



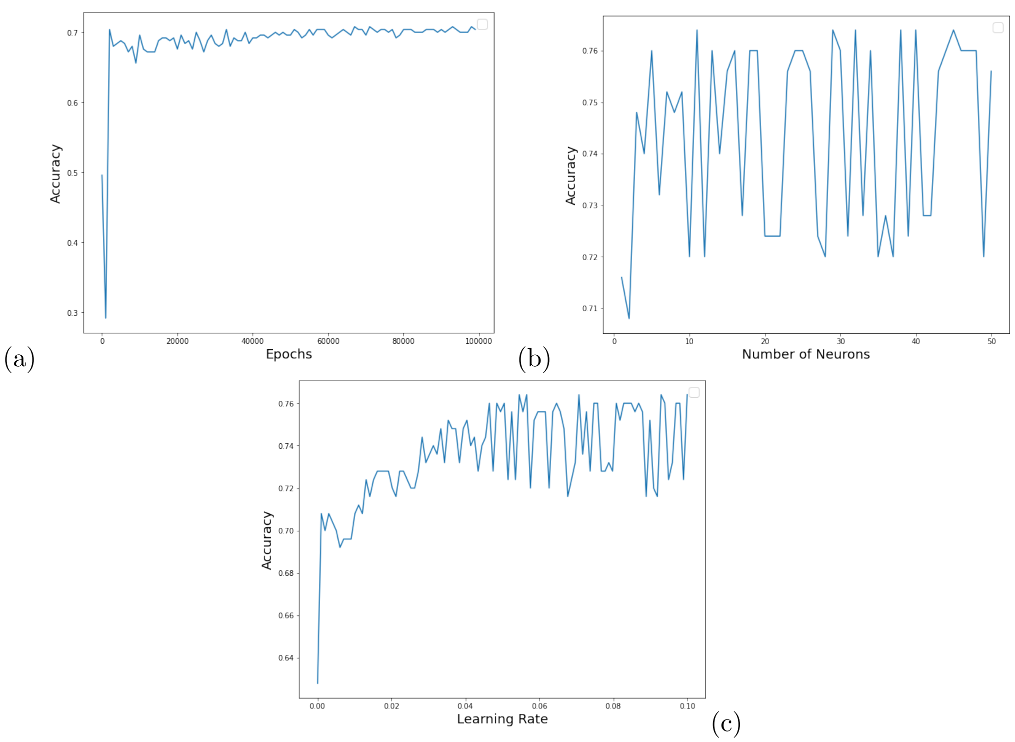
(a) (b) (c)

Figure 8: Box plots of the cholesterol and prothrombin variables against status

Finally, we studied in which manner the variables affect our prediction. In order to expose these results, we can take a look to the figures (12).

4 Conclusions

The exploratory data analysis showed the relation that exists between some usual variables such as age and some medical ones, like copper or cholesterol. Also we see the behavior of the stages in the disease for each status. Additionally, we analyze the amount of values that the status variable has, and we found that this amount is related to the lack of accuracy in the model. Having said this, we train the model applying the Shallows ANN algorithm. Unfortunately we didn't have the best results, principally for the lack of data in the cases, transplant (1), we wanted to predict.

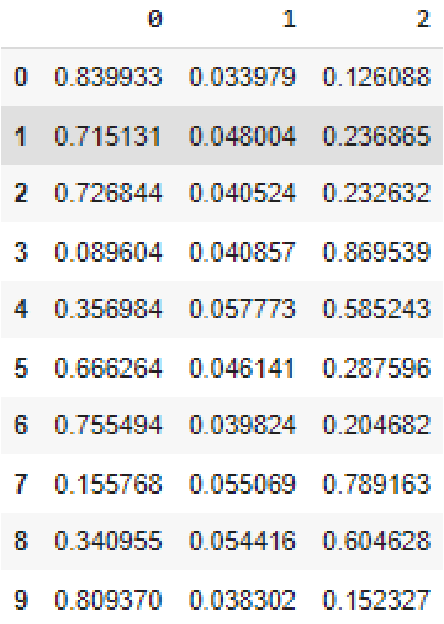


(a)

(b)

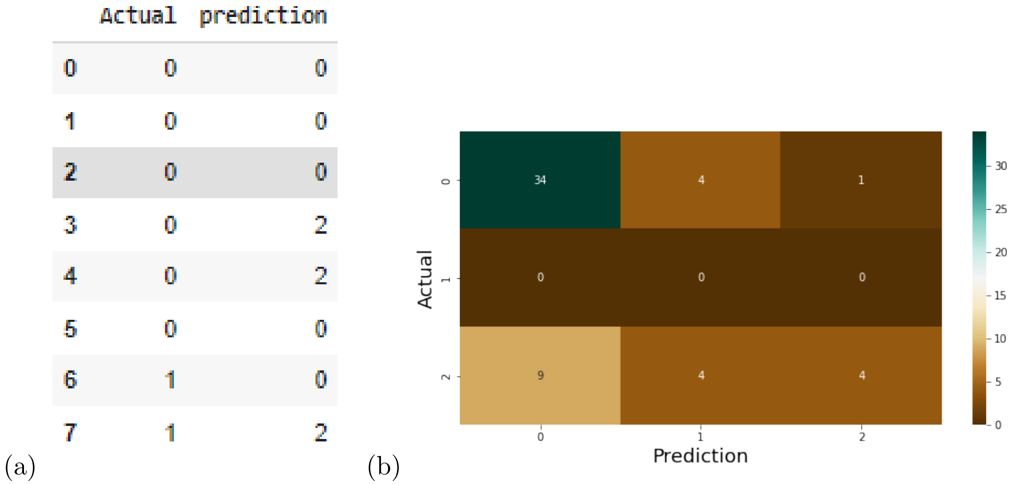
(c)

Figure 9: Plots of accuracy against epochs, learning rate and amount of neurons made in order to find the best fit.



	0	1	2
0	0.839933	0.033979	0.126088
1	0.715131	0.048004	0.236865
2	0.726844	0.040524	0.232632
3	0.089604	0.040857	0.869539
4	0.356984	0.057773	0.585243
5	0.666264	0.046141	0.287596
6	0.755494	0.039824	0.204682
7	0.155768	0.055069	0.789163
8	0.340955	0.054416	0.604628
9	0.809370	0.038302	0.152327

Figure 10: Data frame of the probabilities for each class given by the model.



	Actual	prediction
0	0	0
1	0	0
2	0	0
3	0	2
4	0	2
5	0	0
6	1	0
7	1	2

(a) (b)

Figure 11: Data frame and confusion matrix of the prediction with the Test Data

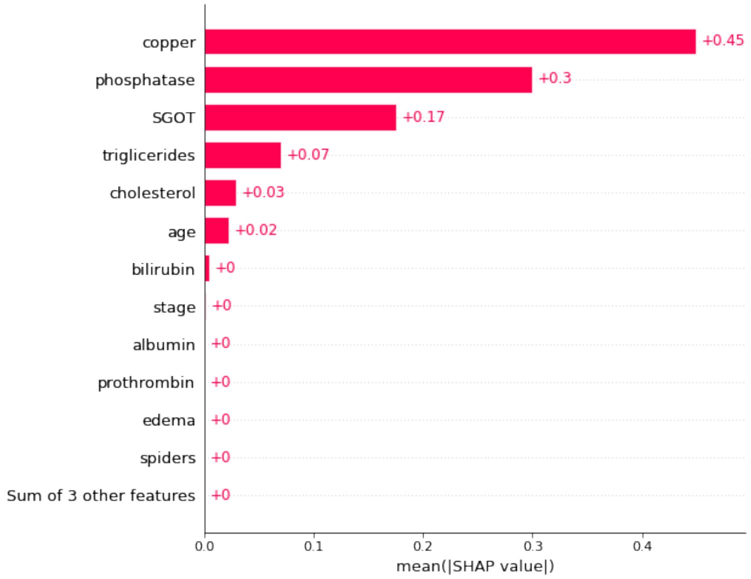


Figure 12: Bar plot that indicates the impact each variable has in the model.