Second Assignment
Machine Learning Models
Titanic

Jaime Salafranca Pardo & Diego Stergar Vega

Wednesday, 24th of November 2021

Summary

---

1. Cleaning the data

2. Making the best tree by Hyper Parameter Selection

3. Making the best random forest by Hyper Parameter Selection

4. Conclusion of the questions of the first Assignment

4.1. For minors was gender an important characteristic for survival?

4.2. Were the children with parents more likely to survive?

4.3. Did money influence survival? If yes, was it the same on a large scale as in a smaller one?

4.4. Were the aged people more likely to spend more money on the trip?

4.5. Do the people that worked on the Titanic were more likely to survive?

4.6. The people embarked in different ports for reasons of money?

3. Conclusion

To create a predictive model to analyze the data and find the relation with survivors, we can use different methods. We could use subsampling, repeated random subsampling cross-validation, or even repeated cross-validations. In this assignment, we are going to use the same data set that shows the variables of age, gender, port of embarkation, Fare, number of Siblings and Spouses, and the number of Parents and Children on board. We can only use data that is represented in numbers, we have to clean the data.

1. Cleaning the data

Survived
TRUE
FALSE
FALSE
TRUE

First, we change the variable Survived that is initially filled with zeros, and ones by True or False. This will be very important because all our analysis will depend on the Survived variable. We can also change it by SURVIVED or DIED to make the tree more understandable.

After we replace the words “male” and “female in the variable sex by the numbers 0 and 1. This is optional because males and female can also work but for us, changing it to numbers was more helpful. We also replace the letters of each ports C for Cheersbourg by 1, “S” of Southampton by 2, and “Q” of Queenbourg by 3.

```
#We change the variable embarked :
titanic.train$Embarked= factor(titanic.train$Embarked,
                      levels=c("C","S", "Q"),
                      labels=c(1,2,3))
titanic.train$Embarked = as.numeric(titanic.train$Embarked)

#We do the same for the Sex:
titanic.train$Sex = factor(titanic.train$Sex,
                           levels = c("male", "female"),
                           labels= c(1,2))
```

Finally, we have two variables that we have to analyze before doing the research. The variable Ticket such as the variable Cabin doesn’t seem to be very useful. For us, it’s clear that we have to remove the variable Ticket. Even if for the Cabin we could do something like using the letters as groups of cabins, we don’t have too much data so we decide to erase also the variable.

```
titanic.train[,"Ticket"]=NULL
titanic.train[,"Cabin"]=NULL
```

We thought that standardizing the data would be very useful to make accurate predictive models. We did computing the means and standard deviation for each variable. Using this we saw that the accuracy results were very similar even if the standardized data was more difficult to understand so we used the non-standardized data.

After doing all these changes we can start creating this is what our set of data looks like.

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
TRUE	1	2	38.0	1	0	71.2833	1
FALSE	3	1	35.0	0	0	8.0500	2
FALSE	1	1	54.0	0	0	51.8625	2
TRUE	2	2	14.0	1	0	30.0708	1
TRUE	3	2	4.0	1	1	16.7000	2
FALSE	3	1	39.0	1	5	31.2750	2

2. Making the best tree by Hyper Parameter Selection

To obtain a tree accurate enough to predict in the function of all the other parameters the survival or the death of a passenger aboard the Titanic, we thought that the best option was using Hyper Parameter Selection. This way of searching consists in making a lot of little changes in the data while measuring the accuracy, specificity, and precision. All of that is done by k-cross-fold validation, a method that consists in dividing k times the data randomly in the test set and training set, (80 % percent of the data), to measure the characteristics mentioned before.

Hyperparameters selection

To use this technique we have to make a vector that determines the minimum split, another for the maximum depth, and finally one for the cp. Using the expand. grid function we can make a matrix with all the parameters to start our research for the best trees.

```
d_minsplit=seq(from=10, to= 100,by=10)
d_maxdepth = seq(from=1, to = 30, by = 1)
d_cp = 2^(-11:-1)
paramet= expand.grid(d_minsplit, d_maxdepth, d_cp)
```

To verify each model corresponding with each set of parameters (max_depth, min_split, and cp), we use the 10 different test sets and training sets.
For each one of these 10 repetitions we create a confusion matrix that could look like this one:

Thanks to this kind of matrix that shows the Classified prediction in comparison to the actual values we can compute the accuracy, precision, and specificity by the following calculus.

		Classifier prediction	
Actual value		TRUE	FALSE
	TRUE	190	66
	FALSE	40	372

The accuracy is the number of values that are True in both cases and False in both cases divided by the total number of values.

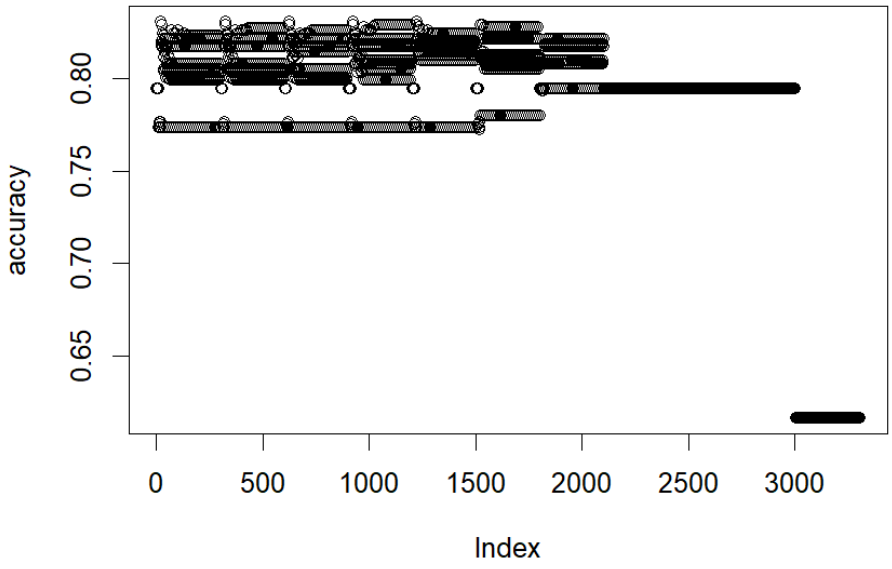
```
accuracy = sum(diag(conf_matrix))/sum(conf_matrix)
precision = conf_matrix[1,1]/sum(conf_matrix[,1])
specificity = conf_matrix[2,2]/sum(conf_matrix[,2])
```

The precision is the number of values that are true in both cases in comparison to the values that are actually true.
The specificity is the number of values that are predicted False divided by all the False values.

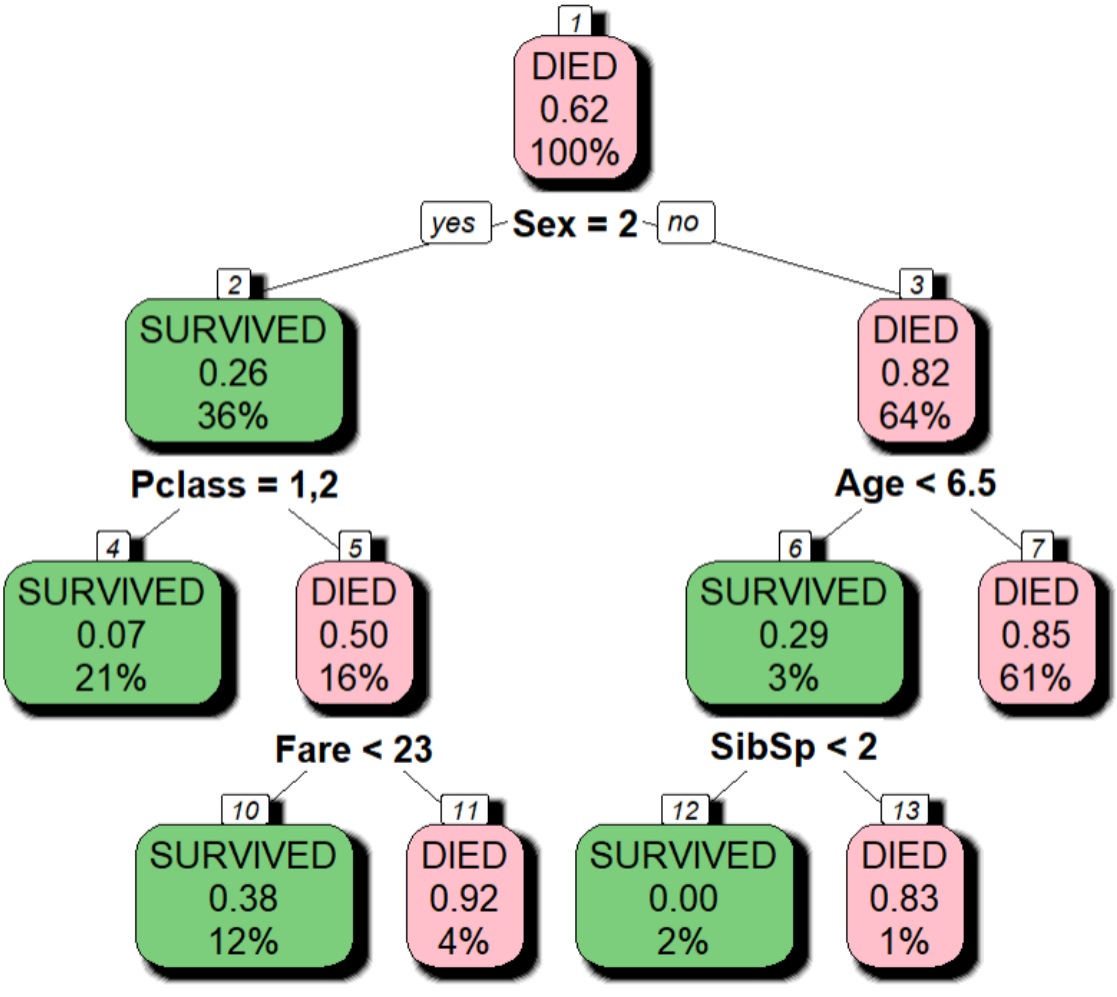
Once we have these three important pieces of information for every one of the ten folds we can compute the means for the model with the parameters used. We repeat the process for every single combination of parameters until we find the best mean of accuracies(accuracy, specificity, and precision).

For us, accuracy and precision were the most important, because in a practical case we have to be more accurate when we say that someone is alive than that someone is dead. After all, if we make a mistake the consequences are worse.

We can plot a graph for all the accuracy that will be the base to choose the best tree. Here we see that the best accuracies are next to 83%. We are going to create the tree for the hyperparameters of those accuracies.



We found these parameters for our best accuracy, even if for us it was not very high we created our tree following these parameters and this was the result. We will analyze it after.



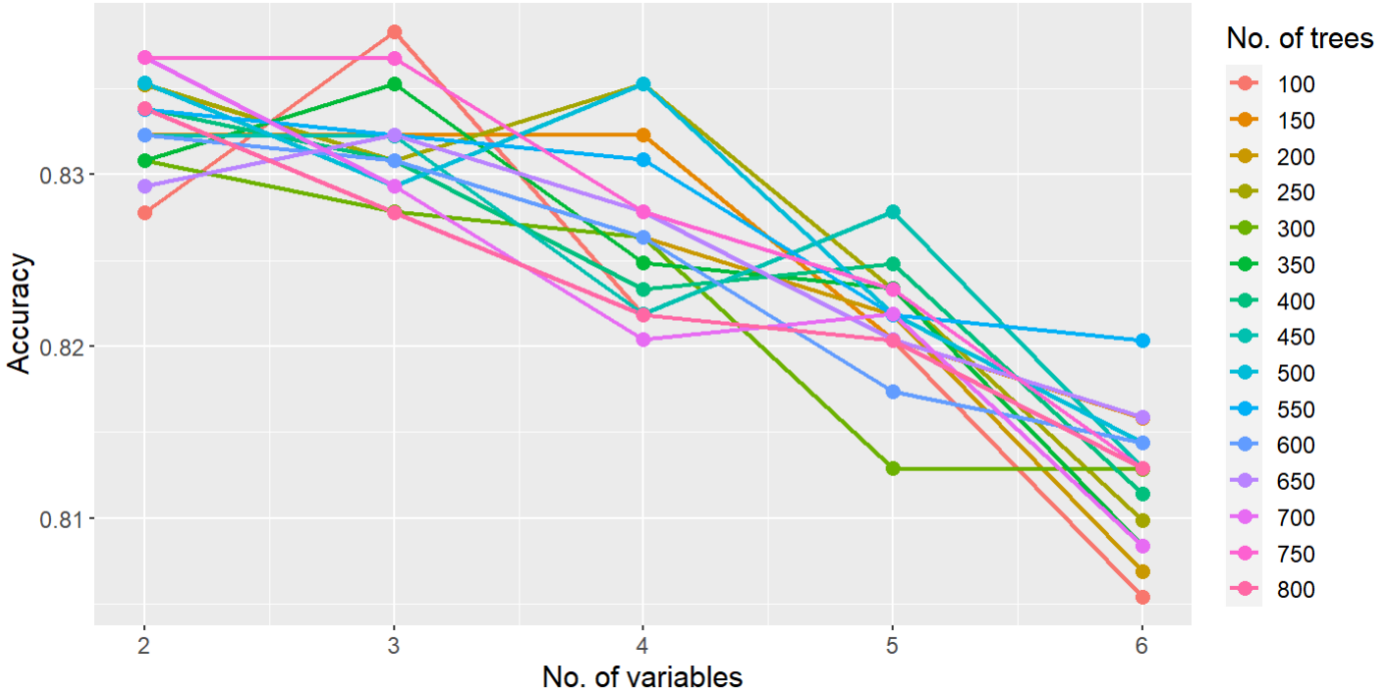
3. Making the best random forest by Hyper Parameter Selection

There are differences between normal trees and random forest that makes these latest interesting to work with. there are some benefits, like the recession of risk of overfitting, an augmentation of flexibility, and also it points very well the importance of each variable. In our work, these benefits seem very useful to limit the differences with the test set that we don't have and to explain our previous conclusion about the variables.

```
d_mtry=seq(from=2,to=6,by=1)
d_ntree=seq(from=100,to=800,by=50)
parameters = expand.grid(mtry=d_mtry,ntree=d_ntree)
```

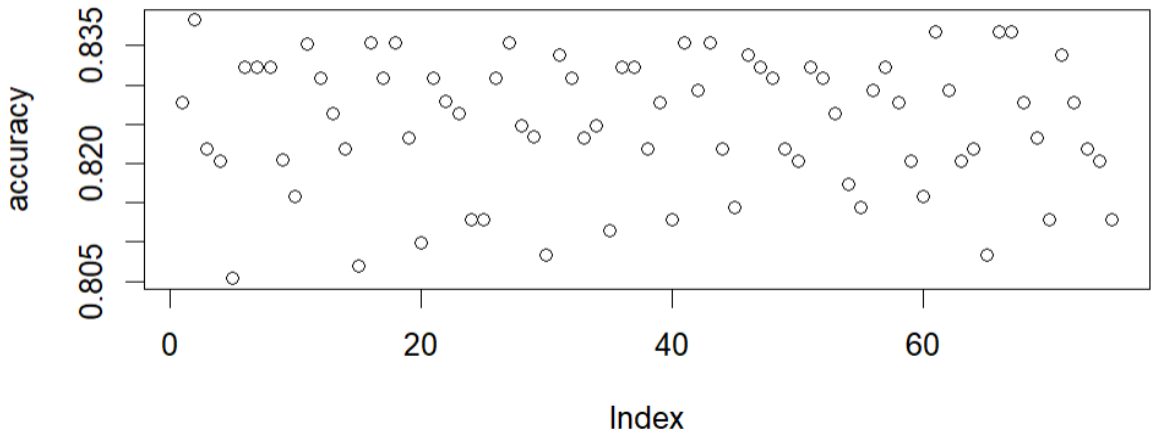
Here one more time we have decided that we were going to use the most time-consuming technique but also the one that was able to give us the best model, the Hyperparameter grid search. The number of variables that we have is 8 so the number parameter mtry will more or less be equal to 3.

One time more inside the loop of the parameters we use the k-fold-cross-validation process to train the data and find the best mean accuracy as we have already explained for the trees.
Here is the graphical representation of the accuracy of each random forest. In this case, we are only going to work on the accuracy, because we search for the best model for accuracy.
We find the best parameters and we create the random forest.



In this graph, we represent the accuracy in the function of the number of variables tested in each forest in the function of the number of trees. We can see that when the number of variables increases the accuracy decreases. This has sense because the more variables are taken into account, the more useless the relationship between them will be, because it will be an extremely overfitted model.

We can see that the maximum accuracy will have a parameter number of variables next to 2 and the number of trees could be between 100 and 750. We are going to find the parameters for the best accuracy and we are also going to see the graph that presents all the accuracies.



In this graph, we show the accuracy in the function of the index that represents each model of random forest. We see that the number of random forests tried is less than 80 and for every one of these trials, the accuracy is between 79% and 84%. Among all these accuracies we have to find the only index that creates the best accuracy. In this case, the best parameters are the number of variables (mtry = 2) and the number of trees (tree = 650) for an accuracy = 0.839. For these precise parameters, we build our best classifier. And we plot it to see the errors.

With this best classifier, we built a function in where we enter a test set, in which we test the model with this best classifier. It builds accuracy precision and specificity thanks to a confusion matrix. It returns the prediction and the confusion matrix.


```r
mymodel = function(test_set){
  #We clean the data inside the function
  test_set$Embarked= factor(test_set$Embarked,
                            levels=c("C","S", "Q"),
                            labels=c(1,2,3))
  test_set$Sex = factor(test_set$Sex,
                        levels = c("male", "female"),
                        labels= c(1,2))
  test_set$Survived = factor(test_set$Survived,
                             levels = c(1,0),labels=c(TRUE, FALSE))
  test_set[,"Cabin"]=NULL
  test_set[,"Ticket"]=NULL

  pred = predict(bestclassifier, test_set, type = "class")
  conf_matrix = table(test_set$Survived,pred,
                      dnn=c("Actual value","Classifier prediction"))
  conf_matrix_prop = prop.table(conf_matrix)
  accuracy = sum(diag(conf_matrix))/sum(conf_matrix)
  precision = conf_matrix[1,1]/sum(conf_matrix[,1])
  specificity = conf_matrix[2,2]/sum(conf_matrix[,2])
  return(list(prediction = pred, conf_matrix = conf_matrix, accuracy))
}
```

All this process of the random forest helps us to create a model a lot more accurate than the tree to predict the result of a set of data in terms of Survival. In the model we clean the data the same way we did, and after we make the predictions. We built a confusion matrix and give back all this and the accuracy of the model.

4. Conclusion of the questions of the first Assignment

With the models built in the second and third parts, we can draw some conclusions about the questions in the first assignment. First of all, we can say that the variables mentioned at the top of the tree are the more relevant when we talk about the variable Survived. The first variable is Sex where we divide at the first stage between male and female. Men are represented with Sex ≠ 2 and women with Sex = 2.

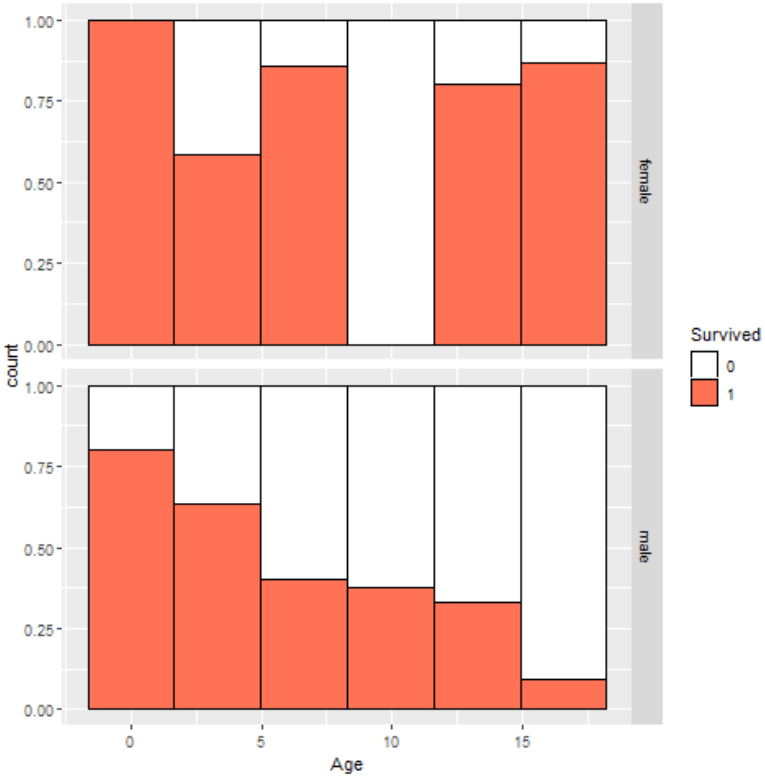
We can also say that the most common thing is to die in the Titanic because 62 % of people didn't survive in the whole data set.

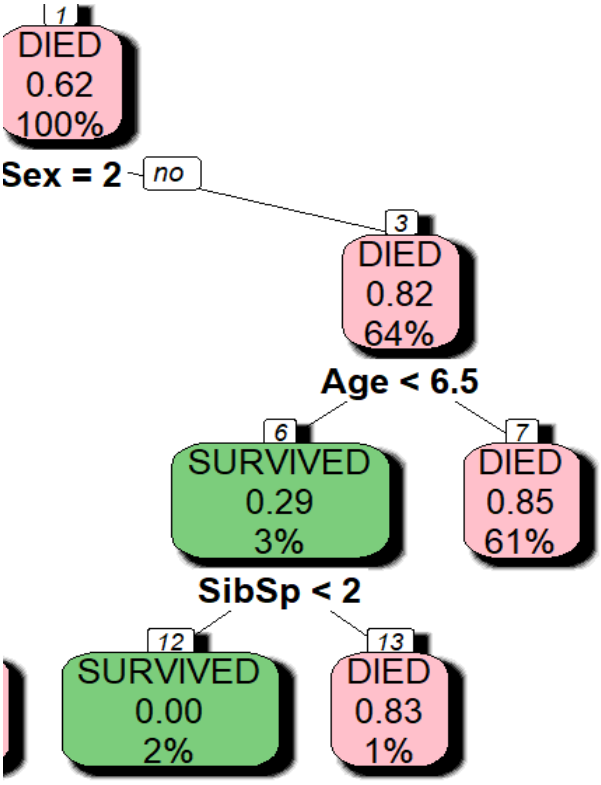
4.1. For minors was gender an important characteristic for survival?

Our conclusion to this question in the first Assignment was that gender was not the most important thing. Between minors, the gender seemed important but it was the combination with the age that seemed more interesting.

The more the little men were growing the more their chances of survival were extremely reduced. In fact, after 5 years old the chances of survival seem to be lower than 50 %. We justified this because the society of that time was expecting more future men with a little age than future women.

We could ask ourselves if this hypothesis was also justified by the tree that we made.



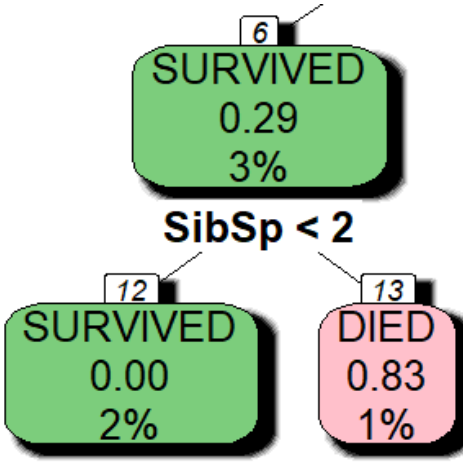


This part of the tree shows the data of men that here are represented with a one so if Sex ≠ 2, it is men. We see that the most common thing about men was dying but if we look after that we see that among kids with less than 6,5 years, it was most likely to survive.

This is very interesting because it shows that after 6,5 years more or less the growing men were as likely to die than the adults so it shows that our first hypothesis is confirmed. Maybe the cause isn't the one that we subject but we can deny that age and gender were important factors in the children's survival.

4.2. Were the children with parents more likely to survive?

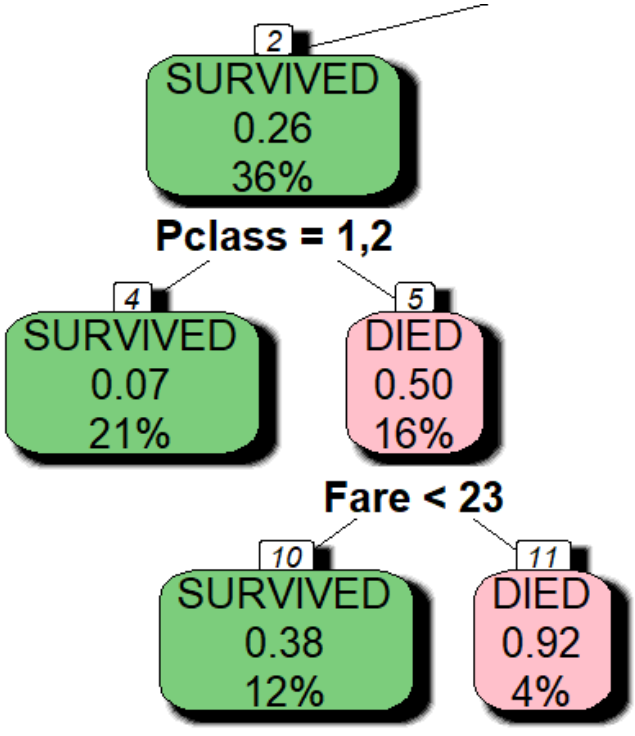
Our conclusion in the first assignment to this question was that between minors the size of the family was in some kind relevant. The thing that surprised us was that it was not the biggest families that tend to survive but the ones that were little families without being alone. In our tree, we don't have a lot of information about family sizes, except for one little side.



We find this brach just after the one shown on minors just before. So this is just applicable to males under the age of 6,5. We see that here the size of the family will matter in some kind, having in to account that normally a group of people studied here doesn't have a wife so the number of Siblings Spouse is the only le number of siblings. Here we see that the people that have more chances to survive is the one that has fewer siblings. Even if this prediction follows the one we made in the first assignment we have to be realistic and say that, this is only relevant for 3% of the data so our hypothesis is not strongly confirmed.

4.3. Did money influence survival? If yes, was it the same on a large scale as in a smaller one?

Our conclusion to this question in the precedent Assignment was that money was very relevant among pooer people. In our tree, the only place where money plays an important role is in the women's branch.



Among the women, we make a first classification where we separate women in first and second class with women in third class. This classification is already economic and we see that women in the 2 best classes represented 21% of the crew and were 99,93% likely to survive. On the other hand, the women traveling in third class had the same probabilities of dying and staying alive.

But what we found more interesting was that among the women traveling in the class the Fare was relevant but not in the way we thought. The women that had a Fare of less than 23 were more likely to survive than the ones with a superior fare. This means that the reacher women were more likely to die.

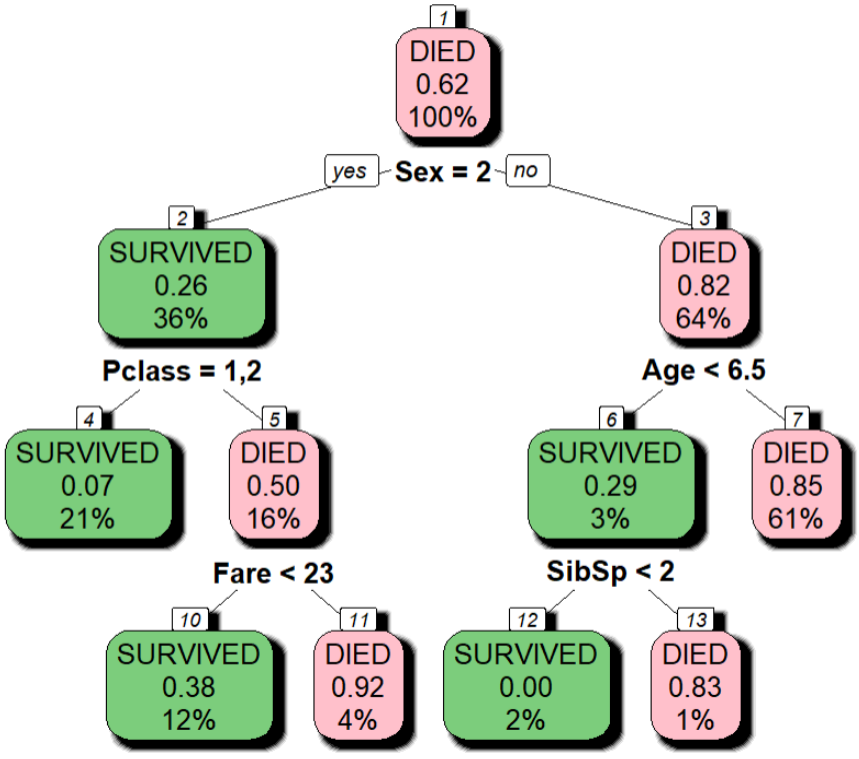
In a first look, we could think that the conclusion on this question from the first and second assignment are opposed one from the other. But we think that it is not the case.

We see that women who paid for a higher class were more likely to survive, which is the same conclusion as the first one. But we can deny that among the third class the money doesn't seem to have any importance with regard to the chances of surviving.

4.5. Do the people that worked on the Titanic were more likely to survive?

In our first assignment, we supposed that people who had a fare=0 were crew members on the board. In this work, we could have changed the value of that fare to the mean but we chose not to. We took this decision to be coherent with our first assignment. In our tree, the Fare condition only gives information in the specific case of women in the third class. Our sample of workers was very very low, so we can not use this condition of Fare to determine a conclusion over the survival of these people.

4.6 Other conclusions of our tree



We can underline other conclusions of our tree. For example, we can say that the most important variable to determine survival were first of all the Gender, and after the Age and the Class. The number of members in the family doesn't have a very big importance. And the Fare was not very important because it is also reflected in the Class. If we had to determine the stereotype of someone very likely to survive we could imagine a man aged 6.5 years and more or a woman in third class with a fare superior to 23. On the other hand, the stereotypes of someone with a lot of chances to survive are a woman in first or second class and a little boy that is less than 6.5 years old and has less than 2 siblings.

We can also say that in our first work we said that in the function of the variable Parent and Children was in some kind relevant. In this case, the variable doesn't seem important. This doesn't mean that the variable isn't important enough but not so important as we thought.

5. Conclusion

In conclusion, we can say that there are different machine Learning that has different benefits and inconveniences. Some are more accurate but more time-consuming while others are faster to make but can be less accurate. In order, to find the best model we have to test the accuracy for a lot of them and finally decide what's the best model. If we compare this to the Exploratory Data Analysis, we can say that it is easier to find the important variables and the ones that have less influence. With these techniques, it is also easier and faster to find the relationship between the variables even if sometimes The Exploratory Data Analysis allows us to understand the problem and the Data. We can finally say that work well balanced accurate and understandable should have Machine Learning Model mixed with Exploratory Data Analysis.