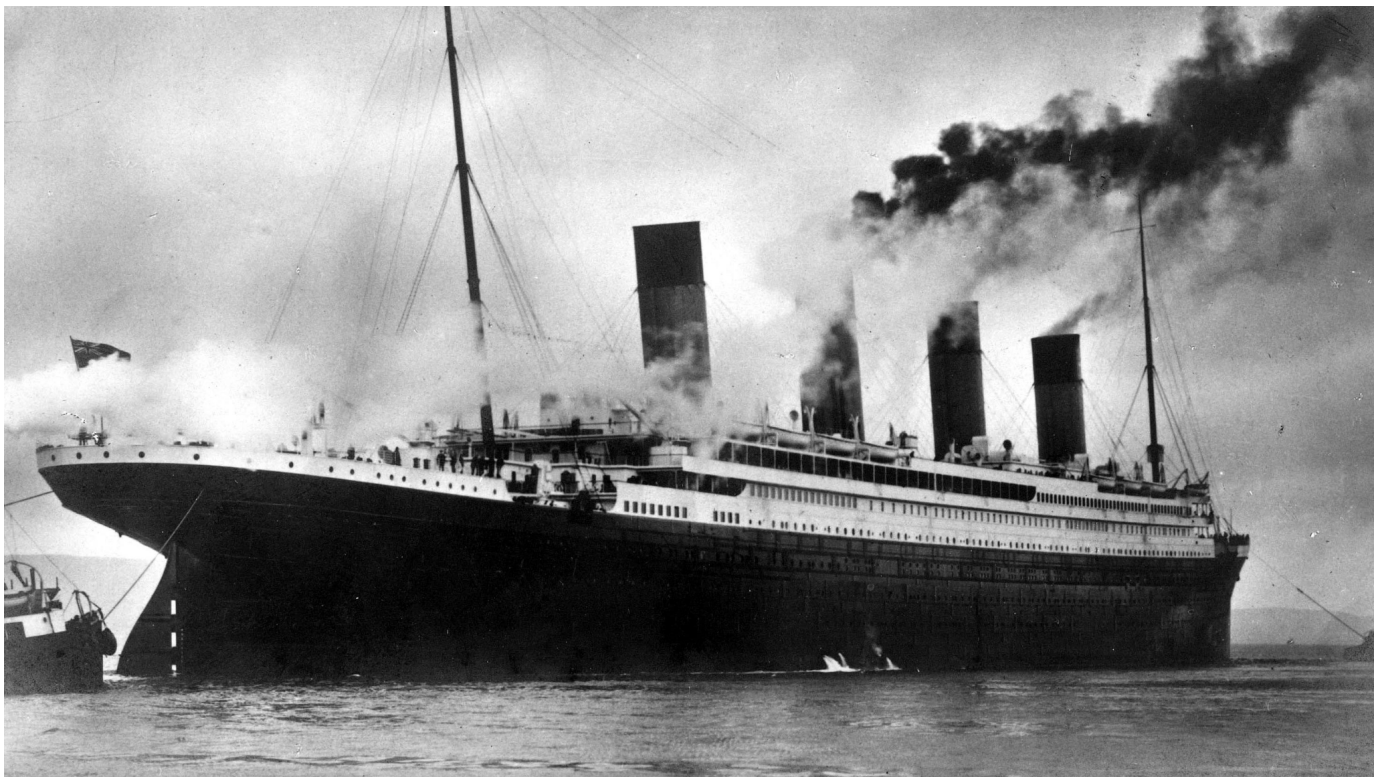
INTRODUCTION TO DATA SCIENCE. SECOND ASSIGNMENT. MACHINE LEARNING MODELS.



CONTENTS.

SECOND ASSIGNMENT: MACHINE LEARNING MODELS.

Introduction.

The assignment consists in using the machine learning model techniques taught in class to analyze additional relationships between variables that might influence the survival, and making a model to predict the survival as a function of the rest of variables. To do so, we are working with a Titanic dataset.

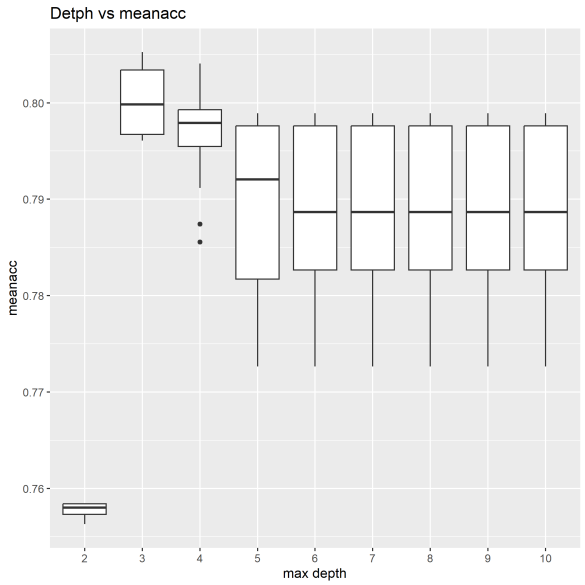
Data exploration and preprocessing.

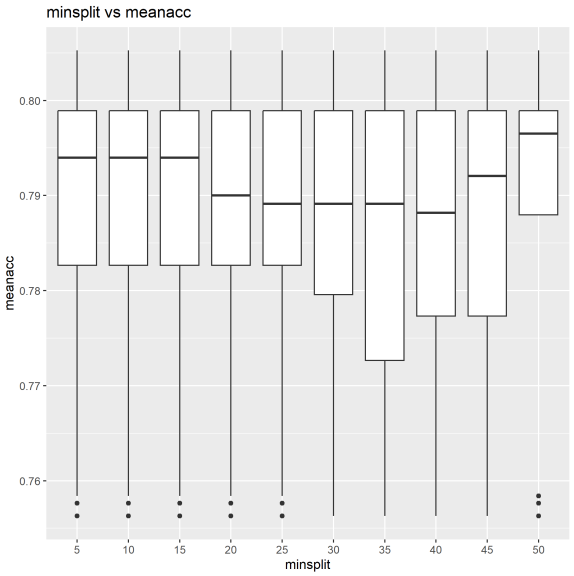
First, we studied the data with the function str() to see, among other characteristics such as the number of observations (668) and the of variables (10), and its type.

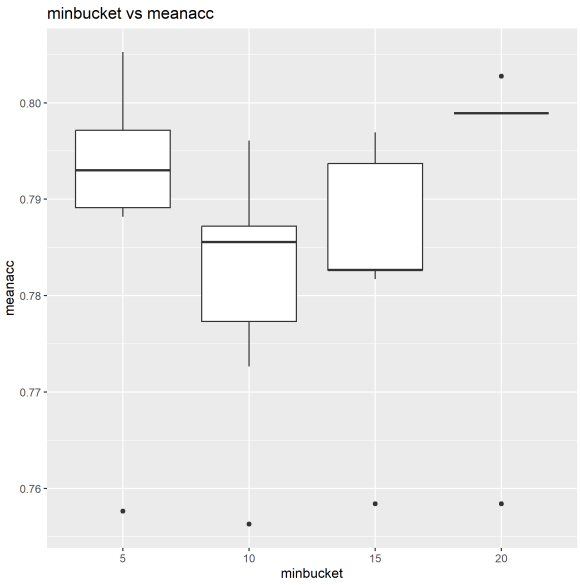
For the preprocessing of the information, we found that for numerical variables such as Age, SibSp, Parch and Fare, there were no missing values, and therefore no need to replace any. Then, we encoded the variable Cabin as a factor and removed the variable Ticket which didn't seem useful, as well as changing the labels of the variable Survived so it shows "Survived" or "Not survived" in the graphical representations instead of 0 and 1.

Decision Tree.

To start, we perform a grid search over different values for the hyperparameters to find the optimal ones for the Decision Tree, and then we perform a K-fold cross validation to see how well it works. Finally we calculate the mean accuracy, sensitivity and specificity for each set, and we store it in a variable called meanacc, to find the best parameters for the tree.



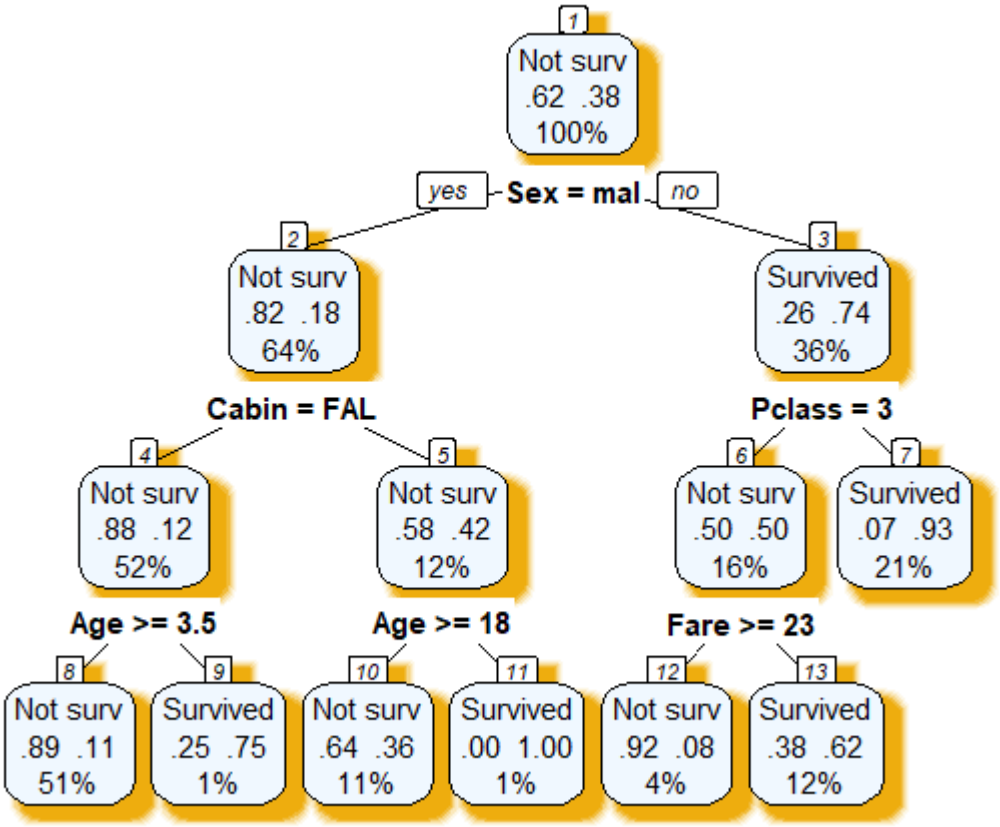




In this graphics plotted with ggplot we can see the relations between these 3 hyperparameters and the precision of the classification tree, being the higher the variable “meanacc” the better.

After having the best parameters, we build the “best tree” with this parameters, and we plot it with the rpr() function from the rpart.plot library.

Plot of the decision tree with the most optimal hyperparameters:



From analyzing the tree, we can see from node 1 that only 38% of all the passengers survived. Then, we can notice that the variables that are more related with the survival are the following:

- Sex: Really significant when coming to survival rate, as it is the root node. In the previous assignment we were wrong in Assignment 1 where we concluded that more than ⅔ of the survivors were men, since we analyzed the graph wrongly, thinking more men survived. Here we can see that men have 18% probability to survive while women had 26% to do so.

- Cabin: The tree suggests that 88% of males who did not have cabin (node 4) did not survive,, which makes sense since people without cabin had lower chance of survival, which also coincides with the outcomes of Assignment 1. In node 5 we can see that men who did have cabin had a 42% survival rate, which is higher than people who did not have a cabin, although they only represent 12% of the people in the titanic.

- Pclass: The class of the passengers. The tree divides this into having 3rd class or no.
    - People with 3rd class had a 50% of surviving
    - People who had higher classes had a 7% of surviving, which is surprisingly, really low.
- Fare: The fare of the passengers. From these ramifications we see that women with less money had more chances of not surviving, which totally makes sense.
- Age: Age is another really important factor. 75% of the childs under 3.5 years did not survive, while male adults over 18 who had cabins all died, although they only represent a 1% of the data.

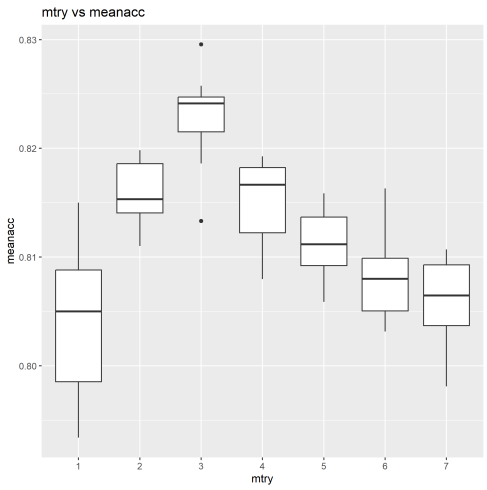
Using the summary function summary(best_tree) we obtain:

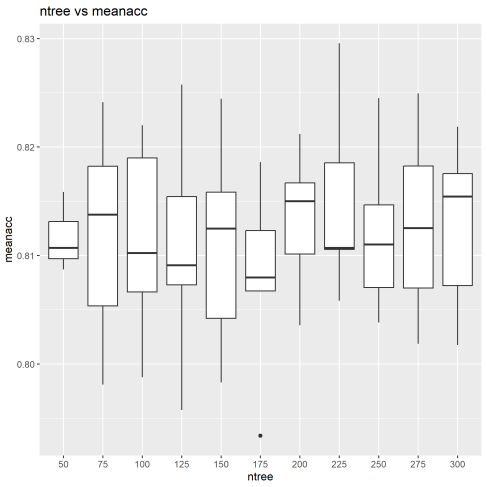
```
Variable importance

     Sex      Fare    Pclass     Cabin       Age     Parch     SibSp Embarked
      43        16        14         8         8         6         3         2
```

Random forest.

First of all, we created a table with every combination of parameters (ntree between 50 and 300 by 25, mtry between 1 and 7 by 1) from which we will train the models and select the best one based on "meanacc". In order to validate the results, we used K-fold cross validation and, for sake of precision and to reduce randomness, we used the same folds from decision trees validation.

After training 77 models, here are some graphs based on the parameter selected and the meanacc obtained.





From these graphs we can see that, in general, ntree = 225 and mtry = 3, give us the best "meanacc", which concurs with the function which.max when applied to the "meanacc" of the different models. Also, it can be seen that the ntree variable doesn't have as much impact on "meanacc" as mtry has.

Finally, the best random forest within this set of parameters is one with ntree = 225 and mtry = 3, with a meanacc of 0.8396 (acc = 0.8383, sensitivity = 0.8882, specificity = 0.7621)

Determination of the best technique.

In order to compare both techniques, we compared the best "meanacc" of models trained with the best parameters of each technique and with the full dataset of titanic. In this first check, random forest presented better "meanacc".

However, just to be sure, we also checked both techniques right after finishing the hyperparameter selection, in which random forest also had better "meanacc".

Final model.

Due to the reasons mentioned above, we decided that the best model we could present was a random forest with parameters ntree = 225 and mtry = 3.

This model, when tested with the full set of titanic, gives us a "meanacc" of 0.9537, acc of 0.9491, sensitivity of 0.9335 and specificity of 0.9784. However, these statistics are rigged as it was tested on the same data it was trained in.