INTRODUCTION TO DATA SCIENCE

Machine learning models:
the Titanic ship

A project made by: Marina Gómez Rey and María Ángeles Magro Garrote

SECOND ASSIGNMENT: MACHINE LEARNING MODELS

This report goes through the analysis done to the titanic data set. It will explain the preprocessing done to it, the validation process used with the techniques which predicted the survival and the selection of the best parameters values. Also, conclusions will be drafted and compared to the initial hypothesis.

1. Preprocessing

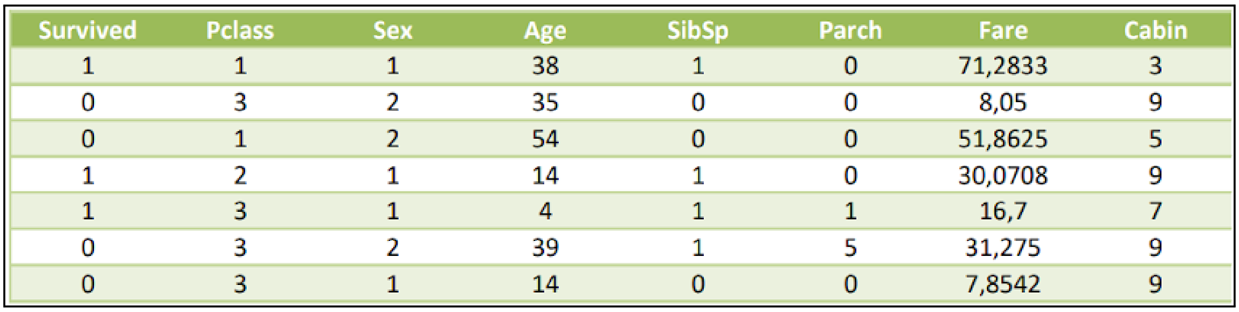
First and foremost, the numerical variables Age and Fare were studied. No empty values were found so there was no need to replace them by the mean of its category. Then, it was decided to normalise the categories in order to have more reliable results, as the ranges of the variables are different. However, as the result did not affect the decision trees and the random forest, it was decided that the normalising was not necessary as the results are more understandable without that normalization.

Afterwards, Cabin and Ticket were considered. Although the Ticket variable was deleted as it was considered that it did not provide any useful information, Cabin was left as in the previous assignment it was proved that it was an important factor when surviving. Due to this, the cabin names were divided by the letter it contained and afterwards, that letter was converted into a number. So, the cabin names which contained an "A" were converted into 1, the ones with "B" to 2, the ones with "C" to 3, …, until "G" with 7. It was taken care of the cabin names which had both "F" and "G" and they were saved with an "F". Then, the cabin names with "T" were converted into a 8 and the empty cabin names to 9. It was done this way as it was proved that the people without a cabin were likely to die, whereas the passengers with an "A" or a "B" (as they were of a better class) had a better survival rate, so it was better to put the empty cabins in the end.

Then, the categorical variables Sex and Embarked were transformed into numbers. Because of this, female was converted into 1 and male into 2, and embarked from "C", "Q" and "S" to 1, 2 and 3 respectively.

Finally, to make sure these data were treated as integers and not as strings, every necessary variable was transformed into an integer value.

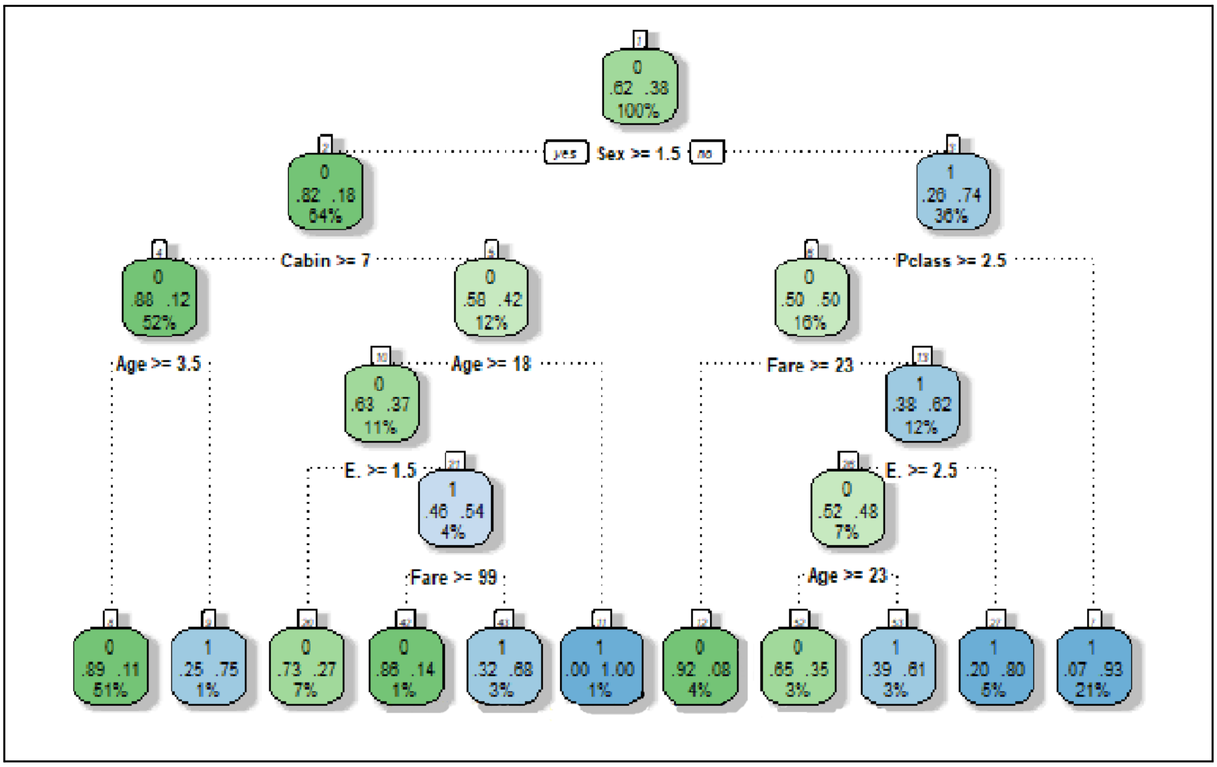
As a result, the head of the data set is:



Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin
1	1	1	38	1	0	71,2833	3
0	3	2	35	0	0	8,05	9
0	1	2	54	0	0	51,8625	5
1	2	1	14	1	0	30,0708	9
1	3	1	4	1	1	16,7	7
0	3	2	39	1	5	31,275	9
0	3	1	14	0	0	7,8542	9

2. Analysis

In order to extract hidden relationships among the variables a decision tree was done with the titanic tree data:

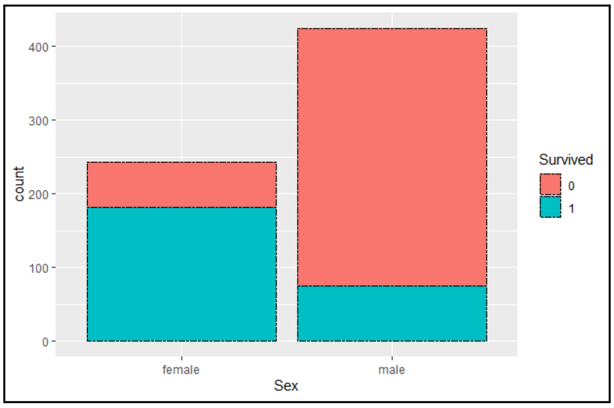


* where E corresponds to Embarked (the name was abbreviated to avoid overlapping in the graphic)

* This tree was created using the Rattle library, that has the function FancyRpartPlot

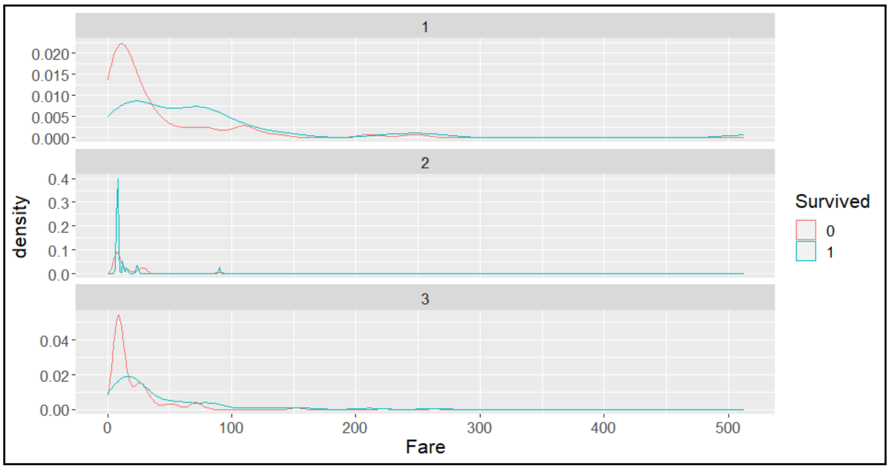
The following conclusions were reached from it:

1) **Sex** was the most important factor when it came to the survival rate. Also, as concluded in the previous assignment, female passengers had a higher chance of survival as proved in the decision tree. It can be seen as the value for women is 1, whereas the value for man is 2. In the first ramification, the value for not surviving is being more or equal than 1,5 in sex (that corresponds to male passengers).

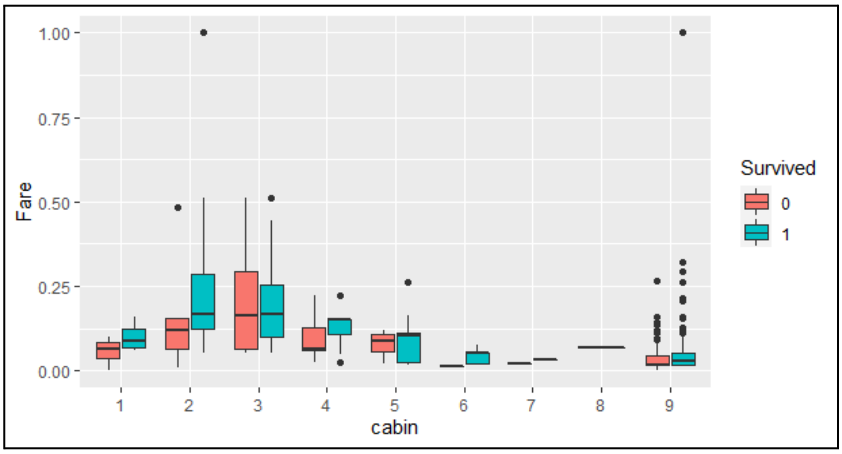


This graphic shows the massive difference between the survival rate of men and women, being this last group more likely to survive (as proven in the tree)

2) **Cabin and Pclass**, which were a representation of the wealth status of the passenger (as it was demonstrated in the first assignment, a higher Fare corresponded to a better Pclass and cabins with letters "A", "B" and "C", which are here 1, 2 and 3, so they are a representation of the Fare variable), also played an important role in the survival rate, taking place in the second ramification.



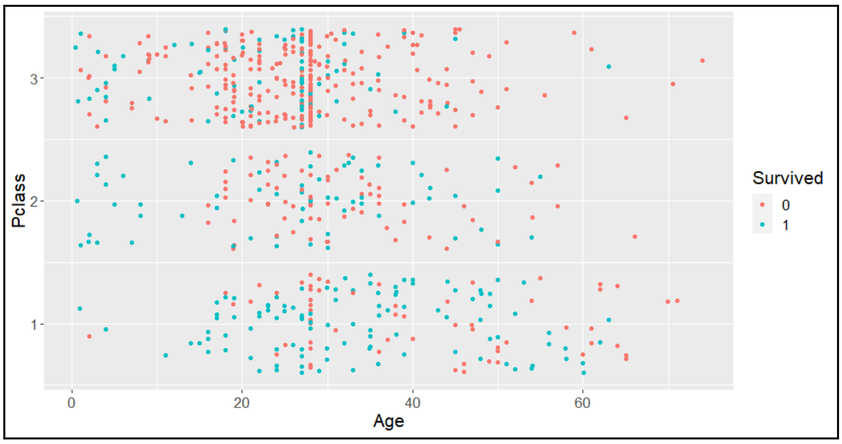
This first graphic represents how a higher Pclass equals both to a higher fare and a higher survival rate



This second graphic shows how the cabin of the first positions have a higher survival rate (and also a higher fare)

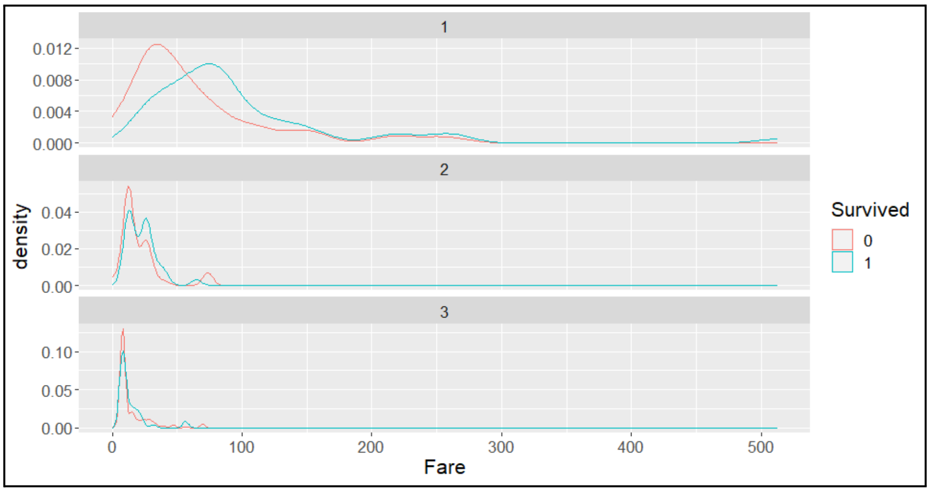
The conclusions of these graphics can be seen in the tree as the ramification of the cabin takes the valor 2,5 to make the difference, meaning that if the Pclass was not higher than 2,5 (most of first and second class) they are highly likely to survive, whereas if they are, further analysis must be done. In the case of the right ramification, the next branch is the fare. Similar to this, the one about the cabin makes a high difference. If the cabin (left part) is 7, 8 or 9 (poorest ones), only little children survive, whereas if it has other numbers, a further analysis can be done (with more possibilities to survive).

3) **Age** is another crucial factor that takes place in the third ramification. As previously said, in the poor cabins only little children survive (<= 3,5). Also, in the wealthier cabins, children under 18 years also survived, whereas older people require a further analysis. Last but not least, in the right part of the tree, there is a final ramification that shows how people (in that category) that were 23 or less years old were likely to survive, whereas the others not.



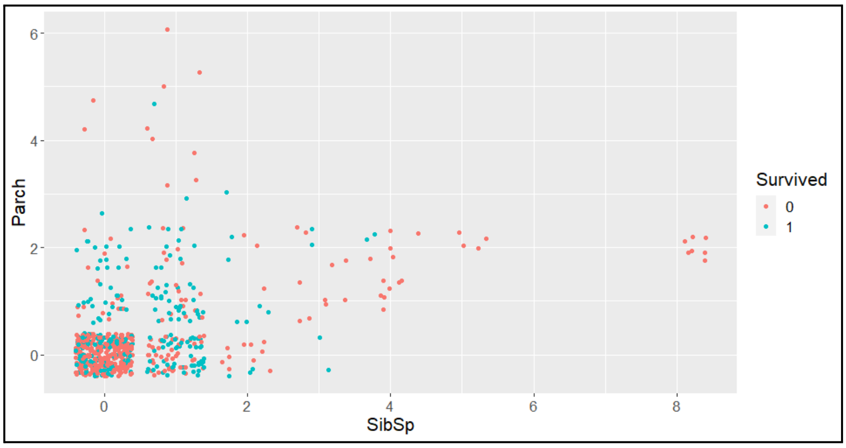
This graphic shows the importance of age, as younger passengers have much more possibilities to survive. The older passengers that survived were mainly of first class as proved in the last assignment

4) **Embarked (port)** is the last variable that divides the tree. The values are: 1 for Cherbourg, 2 for Queenstown and 3 for Southampton. In the left part of the tree, the value of embarked is >= 1,5, which means that passengers embarking in Southampton (and some of Queenstown) were not likely to survive. In the right part (men) the value ascends to 2,5, meaning that passengers from Southampton were not likely to survive, whereas the other ones were (if they were young).



This graphic is a good representation of these conclusions, with a high rate of death in Southampton. On the other hand, Cherbourg had more survival rate because of the wealth of the passengers there

5) **SibSp and Parch** have been proved to not be important factors when it comes to the survival rate, as it has not appeared in any of the branches of the decision tree.



This graphic shows the correlation between SibSp, Parch and Survived. Although the extreme values are red, there is not a real representation of any relationship between these variables as everything is mixed

3. Creating a model

Before creating the final model of the project, it was necessary to choose between a validation process and a technique. For that reason, both a decision tree and a random forest model were created in order to compare the accuracy of each model and choose the best one.

3.1. Choosing a technique

With this goal in mind, a k-fold cross validation was computed in both to choose the best hyperparameters that make the models the most perfect possible.

Decision tree

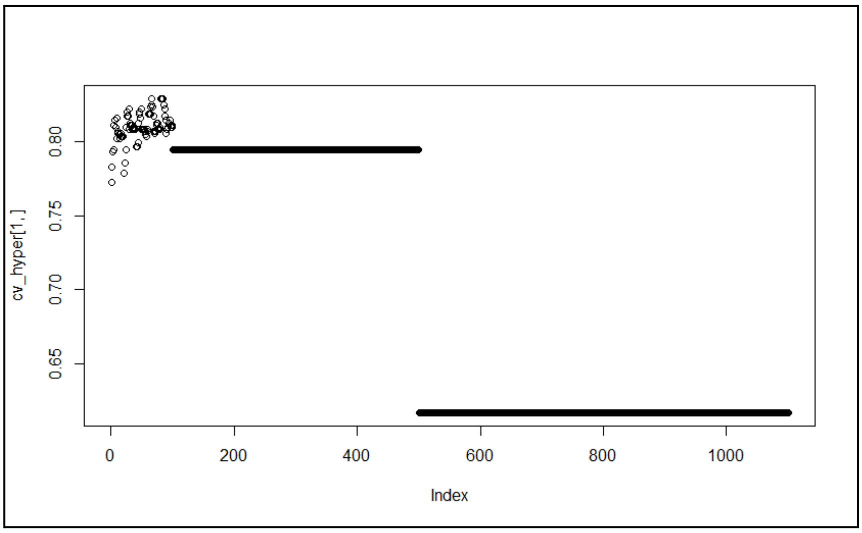
To compute the best accuracy of the decision tree, first, the preprocessing is applied to the data. After that, to make the k-fold cross validation, a seed is decided to obtain the same results always (123 was chosen) and the number of folds is also specified (10 was chosen). After that, the hyperparameters are included, with the recommended values (after the validation the best combination will be obtained). The hyperparameters considered were:

- The **minsplit** (minimum number of observations), whose values considered were from 2 to 40 by 2 each time.
- The **maxdepth** (maximum depth of the final node of the tree), whose values considered were from 1 to 5 by 1.
- The **cp** (complexity parameter), whose values considered were from 0 to 1 by 0,1.

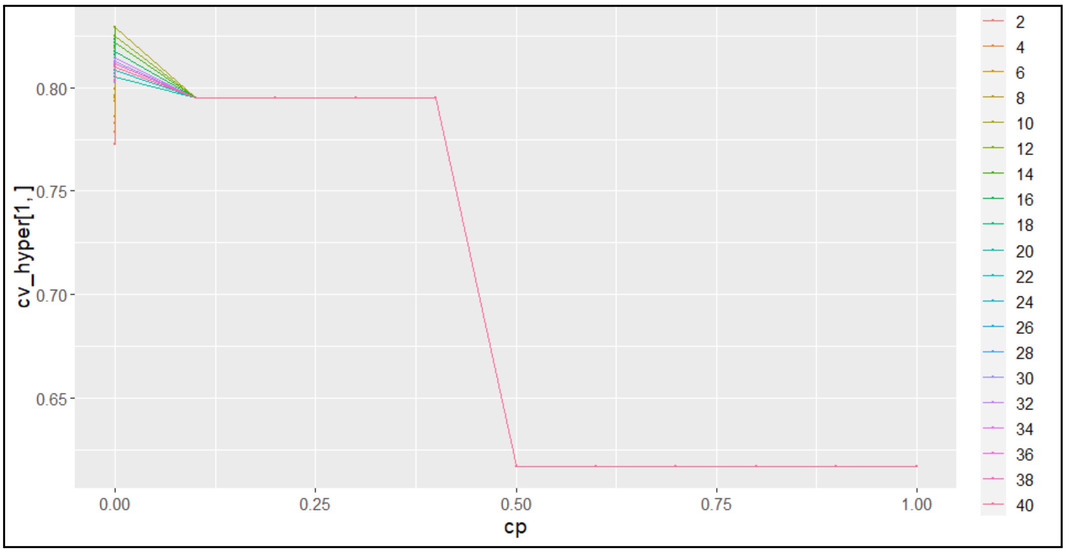
Then, all the possible combinations of the values are generated with the expand.grid function.

Having done this, a function that computes the values of the accuracy, precision and specificity of each combination is generated. That function uses the lapply function (as it uses a list) and will select the training and test set according to the current split (to only use the training set as the test set must not be used until the final model), it generates the different possible trees and the confusion matrices of each of them and return them.

Finally, the values of the accuracies can be represented visually using the plot function:



After that, the which.max function allows to search for the highest values in the return of the function. In this case, the highest accuracy is **0.8293672**, with the hyperparameters of the 67th combination: minsplit = 14, maxdepth = 4, cp = 0, as shown in the following graphic (it shows that the best value for cp is 0 and the best for the minsplit is 14, as the green line has the highest accuracy).



<u>Random forest</u>

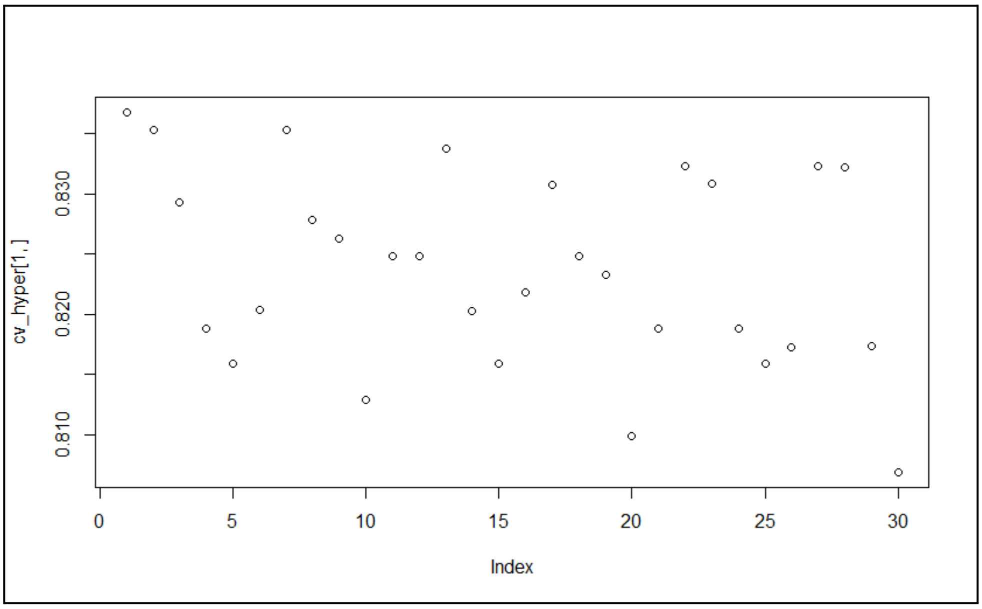
Subsequent to getting an accuracy of 0.829 approximately with the decision tree, the random forest classification was tested. As in the previous one, a k-fold cross validation technique was applied, also using 10 folds (the same seed was also applied). In this case the following hyperparameters were used:

- The **ntree** (number of trees generated) which went from 100 to 600 by 100.
- The **mtry** (number of variables randomly sampled as candidates at each split) which went from 2 to 6 by 1.

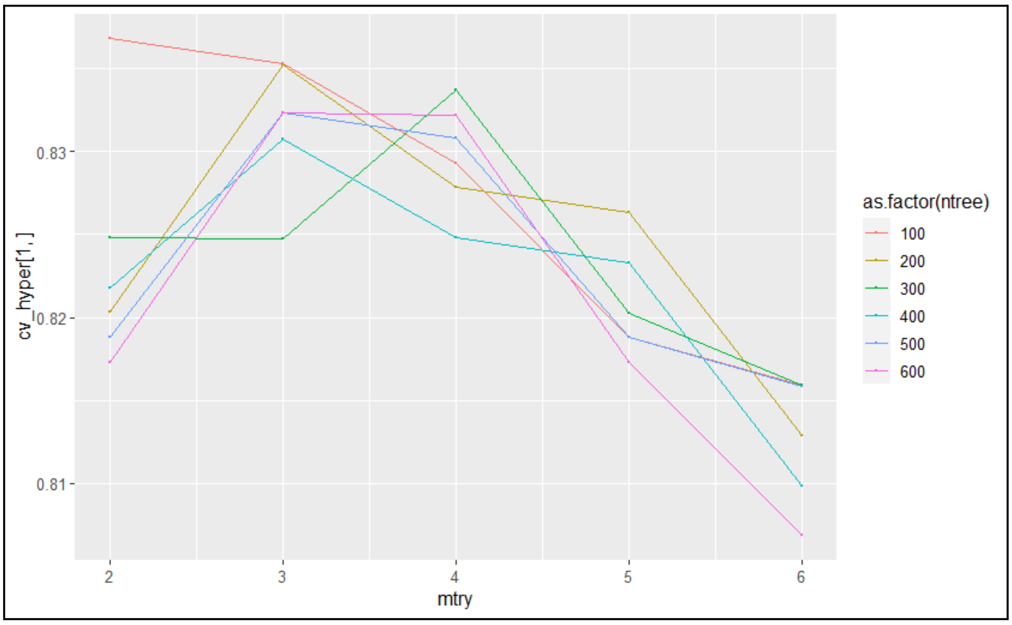
Afterwards, the expand.grid command was used to make all the possible combinations of both parameters (they were saved in the parameters variable).

To compute the accuracy of each combination, the cv_hyper function is created. It uses the function apply (that uses an array) using the parameters variable. Inside this function, another one is created (it is similar to the one created in the decision tree): it uses the lapply function (uses a list) and uses the test set and training set that corresponds to each iteration. After that, it generates the random forest that corresponds to the current hyperparameters and generates the confusion matrix (with the accuracy, precision and specificity of each model) and returns it. Finally, the cv_hyper function uses the values that cv gave and returns the final accuracies using the mean function (that means that it computes the mean of every accuracy of each tree in the random forest to obtain the accuracy of the random forest) and returns it.

With that process done, a graphic representation of the obtained accuracies can be done with the plot function:



Once this is done, the position of the best accuracy is saved (with the which.max function), and the parameters of that position (which are the optimum parameters) are obtained (mtry = 2 and ntree = 100). It can be plotted using the ggplot library:

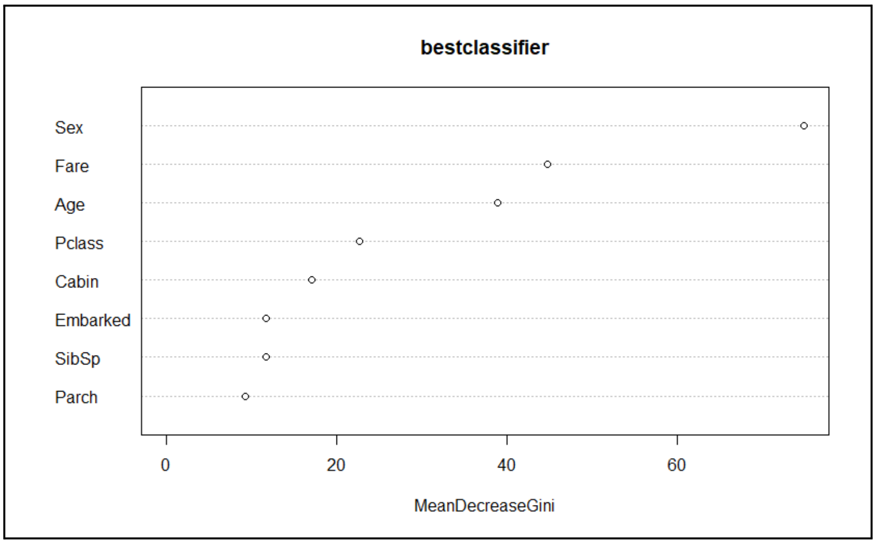


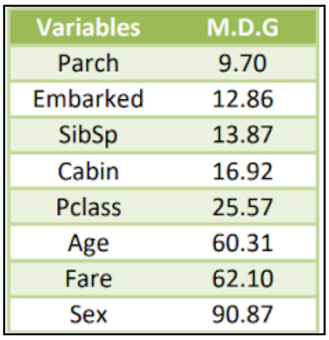
The highest accuracy obtained using the random forest was of **0.8367853,** higher than the one of the decision tree. Due to this, the technique selected for creating the model is a random tree.

3.2. The model

To compute the final model, once the random forest was decided, a function was generated. In it, the data preprocessing was added with the random forest with the best hyperparameters (applied to the test set). Then, it computes the confusion matrix and the accuracy, precision and specificity of the model.

With the model, the importance function can be used to obtain the order of importance of each variable in terms of being decisive when it comes to the survival rate. The following results were obtained:





Variables	M.D.G
Parch	9.70
Embarked	12.86
SibSp	13.87
Cabin	16.92
Pclass	25.57
Age	60.31
Fare	62.10
Sex	90.87

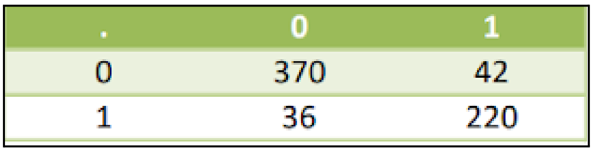
*The higher the mean is, the more importance the variable has when it comes to surviving.

The importance of the variables is as expected, as it arrives to the same conclusions that were extracted both from the previous assignment and the decision tree:

- **Sex** is the most crucial factor
- **Fare** is the next most important variable, that although it did not seem that principal in the decision tree (it appeared very low in the graphic), it really is as the variables Pclass and Cabin are representations of it.
- **Age** is the third most important variable
- The values of the **Pclass** and **cabin,** as representations of the fare, are the next most important but the values decrease significantly.
- **SibSp**, **Embarked** and **Parch** are the last ones and have very slight values. This means that these characteristics were not taken into account when choosing who to save and that it was another characteristic of the passenger what saved them.

3.3. Performance of the model

As seen in the values of the next confusion matrix, the performance of the model with the titanic train data set has improved to **0.8832335 % of accuracy**. A main reason for this could be that it has been used the same data set from checking the model than the one used from making it.



.	0	1
0	370	42
1	36	220

* confusion matrix obtained from the model created using the titanic train data
* It gives a precision of 0.91133 and a specificity of 0.8396947

4. Conclusions

On the one hand, regarding the data of the Titanic, the main conclusions that were obtained were about the priorities when it came to deciding who to save in the moment of the accident.

Both the decision tree and the importance graphic have proved that the **gender was crucial**, with a tendency to save women. Furthermore, **wealth was also a key** characteristic of the surviving passengers, as people in better cabins and with better Pclass (representation of the fare) had visibly a higher percentage of survival. On top of that, it was also demonstrated that **being a minor was also determinant:** the younger the better.

In spite of the previous statements, a formulated hypothesis from the first project was refuted: the embarkment place was decisive. As the importance plot has revealed, **there is not as much correlation as previously expected between having embarked in a certain port and being one of the survivors**. Apart from that, SibSp and Parch do not contribute any relevant information.

Having taken everything into consideration, the last three mentioned variables have nothing to do with the individual characteristics, while sex, age and wealth were really defining.

On the other hand, regarding the applied techniques, **the random forest provided better accuracy.** This is a logical resolution, taking into account the definition of random forest: the outcome of several decision trees (instead of just being one). An important issue that had to be addressed while programming was facing differences in the graphic when rerunning the program although having a seed. This was solved by making the preprocessing both in the decision tree and the random forest (erasing the environment with the remove function) as the variables were reseted. As a matter of fact, an **important concept learnt** is the **choice of the hyperparameters** (using a process such as k-fold cross validation) in order to extract the best possible results and locate them to obtain the optimum model (the values of the accuracy vary a lot between different hyperparameters).