MACHINE LEARNING MODEL SELECTION

Decision Tree & Random Forest

Introduction to Data Science

Universidad Carlos III de Madrid

Jamison Biddle (100492595)

&

María Victoria Vivas Gutiérrez (100452684)

Group 96

Data Preprocessing

For the data preprocessing, the first step was to run a summary of the dataset. One benefit was to make sure the test variable, in this case Survived, was a binary factor. It already was. Next, any variables that were not significant were removed. After project one, we determined that tickets were not significant and cabin floors were not consistently significant- however, whether a cabin was listed did seem to have an impact. Thus both were removed but "hasCabin" was added. An initial classification tree was done, though it was not ultimately included in results as it was without any hyperparameter selection. With all data preprocessed, machine learning strategies were tested to find an ideal model.

Machine Learning strategy

The main goal of this project has been the creation of a model which looks into the relationship of the "Survived" binary factor with all the other variables provided by the Data set. To accomplish this objective, different models have been trained by validating Decision Tree & Random Forest. Thus, the criteria for selecting the best model applied has been the accuracy. Accuracy could be defined as a measure of the model performance in general, i.e, how often the classifier is correct overall observations. Therefore, the accuracy computation has been reached by the next formula:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total Number of Predictions}}$$

Even though it is widely known the "Accuracy Paradox", it has been considered that in the Titanic Dataset this measure will perform correctly. That is, if one of the levels of the boolean factor "Survived" had been a quite smaller number of observations, accuracy will no longer perform correctly in evaluating the model. Nonetheless, as the number of survivors ("Survived == 1") is 256 in the given data and the observations of passengers who did die are 412, the difference has not been regarded as so relevant to consider the dataset as imbalanced. As a result, the use of accuracy as a criteria is trustful.

1. Classification Tree

1.1 First Classification Tree

First of all, an initial classification tree has been trained in the pursuit of the Survived factor prediction. Decision tree is a supervised machine learning method which operates suitably on labeled data sets contexts. The selection of the Classification tree is due to the nature of the titanic data set and of the Dependent Variable, which is a categorical one, i.e, a boolean factor which has two levels for those who survived and those who did not. The first model trained has included all the variables of the preprocessed data (remision to the first section) as predictors except the dependent one. The result can be graphically shown as it follows:

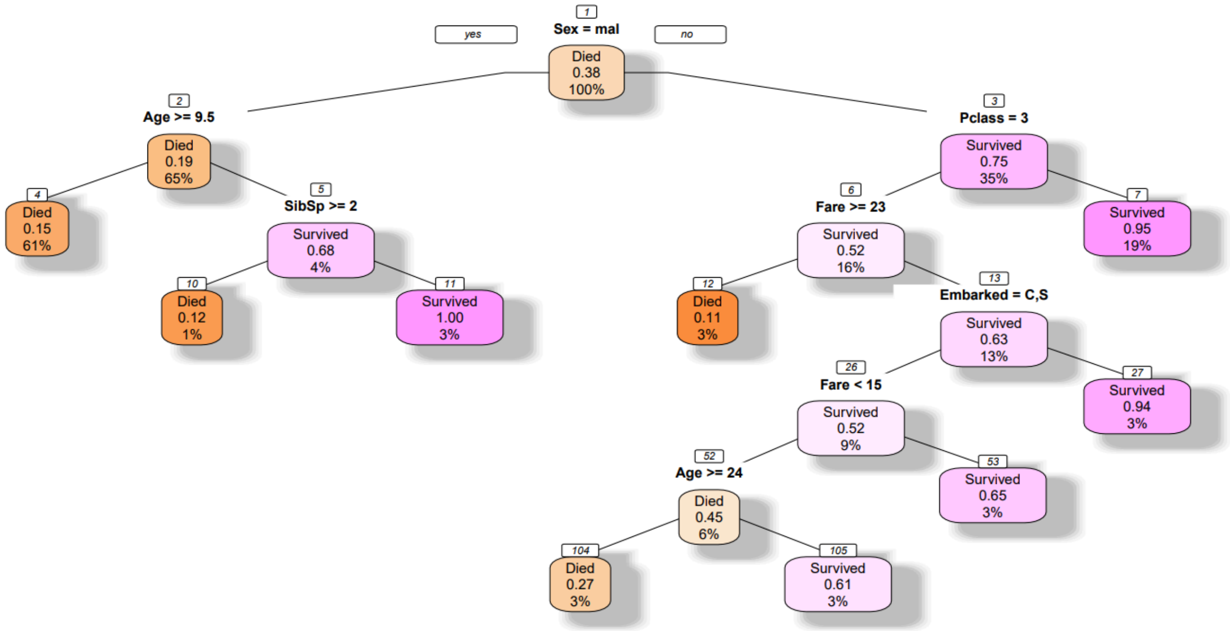


Figure 1: First Decision Tree estimated

The root node determines the overall Survival rate, which is 38%, meaning that the first classification will result in a 0 ("Died"). First split is made by the categories of Sex, suggesting that the 65% of passengers were men that had a survival probability of 19%. "Sex" is the most relevant variable according to the model. The second node inside the men category has been splitted by age. Belowing to the group of male sex and above 9.5 years old passengers reduced the survival rate of 61% of passengers to a 15%. The Exploratory Data Analysis has already shown the influence of "Sex" & "Age", so the prediction agrees with the descriptive statistics analyzed before. But being a man younger than 9.5 years old could increase the survival rate under the condition of having more than two spouses or siblings aboard. If they had, the Survival rate augmented to 100%, but only including the 3% of the data in this node. If they do not, the probability of Survival is only 12%, but again, this leaf node only covers the 1% of the data, i.e, this split could be removed when pruning the tree.

After the root node, it has been remarked that being a woman (35% of the data) implied a Survival Rate of 75%. The 3rd node divides the women according to their Ticket Class, which resulted in a 95% probability of survival for women that traveled by 2nd & 1rst class. For those women who traveled in the 3rd one, the situation was somewhat worse: the result is still survived but with a probability of 52% and depending on the "Fare" variable. Those women who paid less than $23.000 had a Survival Rate of 63%; a situation which is much better if they embarked on Queenstown because this node displayed a probability of Survival of 94%, even though only for the 3% of data. Female passengers who embarked on Cherbourg & Southampton are grouped by "Fare" and "age". However, this split seems to be not as significant as it should be to be conserved in the best decision tree model in terms of accuracy: it will probably disappear when pruning the tree.

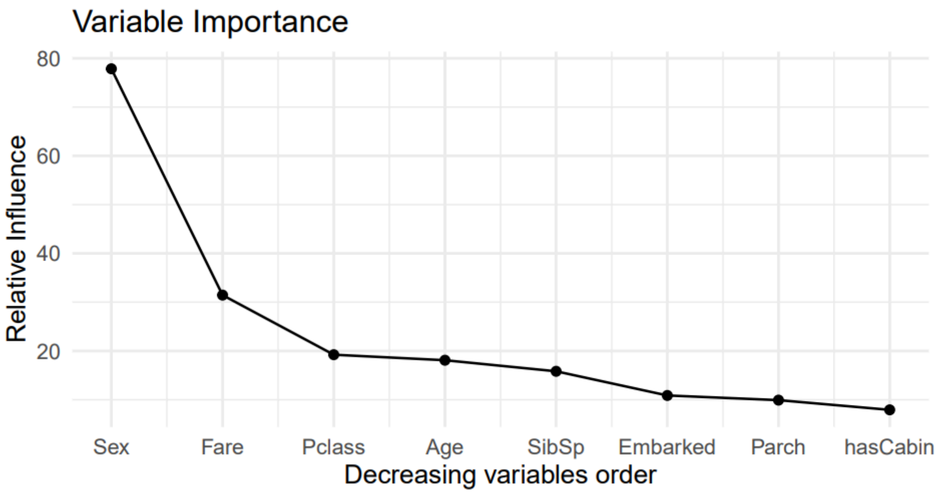


Figure 2: First Tree "Decreasing variable importance"

This graph is a visualization of the decreasing importance of each variable in the first model. Variable importance is established by computing the relative influence of each variable in two aspects: whether that variable was selected to split on during the estimation of the tree and how much the variance decreased as a result of its effect in the model. There is a huge difference between the relative importance of "Sex" related to other variables. The consequence is that these variables which are in the right on the X axis in the graph are not useful in the task of reducing the estimation error, so in the best decision tree model controlled by the hyperparameters which provides the highest accuracy, they will not be included.

It has been computed a decision matrix of the first tree to obtain the Accuracy parameter. The Titanic data set was split using 80% of data for training the model and 20% for testing it in a random way. The results of using the testing data set in the model trained by the training one could be represented in the following figure:

Actual value/ Classifier Prediction	Died (prediction)	Survived (prediction)
Died	**.5640**	**.0526**
Survived	**.0752**	**.3083**

Figure 3: First confusion matrix

Accuracy is computed as it follows;

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{0.5640 + 0.3083}{0.5640 + 0.0526 + 0.0752 + 0.3083} = 0.8722$$

1.2. Hyper Parameter Selection: “Pruning the Tree”

Decision trees provide a nonlinear & nonparametric tool for prediction with a simple interpretation that makes them advantageous to be trained. Even though the application of the decision tree is justified, this Machine Learning Method has the limitation of facility to overfit the data. For ensuring the model to work in predicting new observations, hyperparameter selection is suitable to choose a better model. The objective is not to increase the complexity of the model but ensuring the maximum level of accuracy.

It has been fixed the “minsplit”, “maxdepth” & “cp” that maximize accuracy in a K-fold cross validation algorithm. Minisplit (18) fixes the number of observations for every node to be splitted and maxdepth (5) sets the maximum depth of the nodes. The complexity parameter (0.001) implies that every split that does not decrease the error on the estimation of “Survival” in this measure will not be performed. Therefore, the best model selection includes a lower number of predictors.

The Best Decision Tree model applying these hyperparameters that maximize the accuracy is graphically shown as follows:

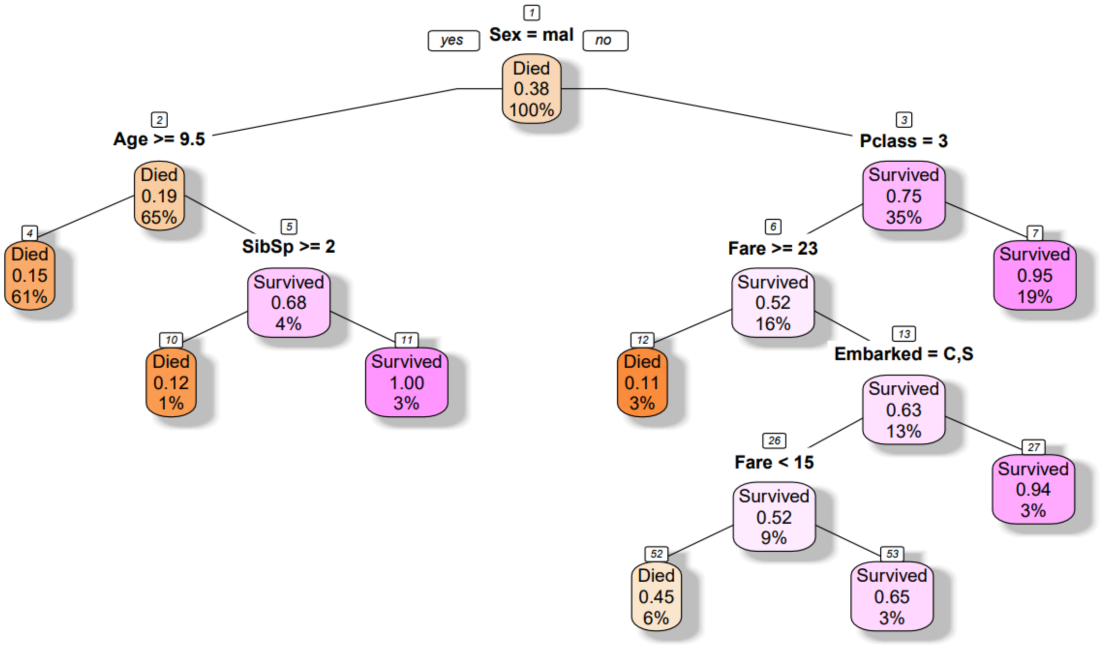


Figure 4: Best Classification Tree

The Accuracy-maximicer Decision Tree Model did remove the last split by “Age” that was included in the first model without the hyperparameters. The “Variable Importance” has changed by including the hyperparameters too, but the difference seems to be meaningless.

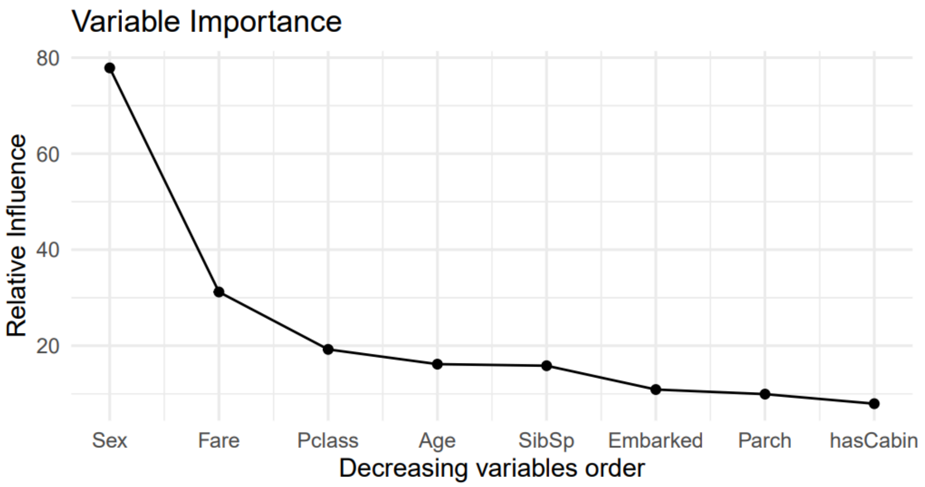


Figure 5: Best Tree "Decreasing Variable Importance"

As it was mentioned before, decision trees could overfit the training dataset. For solving this problem, different validation methods have been applied to the pruned tree. Validation is the process of ensuring that relationships of the variables included in the predicting model of Survival one describe in the right way the behavior of the data. As it have shown in the previous section referring to the first Tree, a Hold-Out Validation has been checked in the best model, as it can be represented in the following Confusion Matrix:

Actual value/ Classifier Prediction	**Died (prediction)**	**Survived (prediction)**
Died	**.5714**	**.0451**
Survived	**.0977**	**.2857**

Figure 6: Best Tree Confusion Matrix

Accuracy is computed as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{0.5714 + 0.2857}{0.5714 + 0.0977 + 0.2857 + 0.0451} = 0.8571$$

1.3 Validation (Repeat, K-Fold)

The selection criterias obtained through a simple Hold-Out Validation are not enough to decide the utility of the model. This result may depend on the randomly selected data. First of all, a Simple Validation for the tree has been repeated a hundred times. That is, it has randomly selected different observations for every 80% training data and 20% testing data for every repetition. The algorithm in a list with the parameters and its values that have been unlisted into a dataframe with two vectors with a length of 300 rows. The average accuracy obtained was .81 approximately.

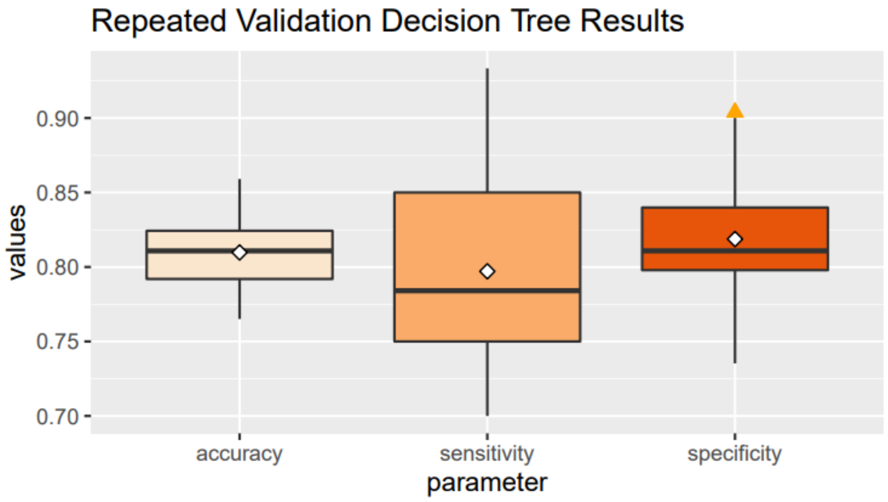


Figure 7:Boxplot "Repeated Validation" Decision Tree

Secondly, a K-Fold validation algorithm was trained. This time the data set has been splitted into ten folds with random observations for each one. In every repetition of training the tree, a split is excluded from the training data to test the model. Average accuracy value (.83 approx.) is achieved by K-Fold Cross Validation , as it is represented in the 8th figure.

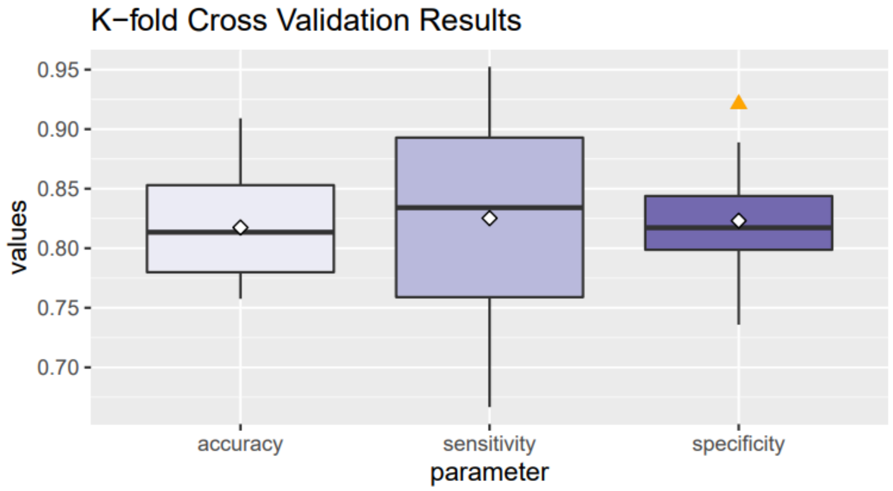


Figure 8: Boxplot "K-fold Cross Validation" Decision Tree

2. Random Forest

2.1 Hyper Parameter Selection

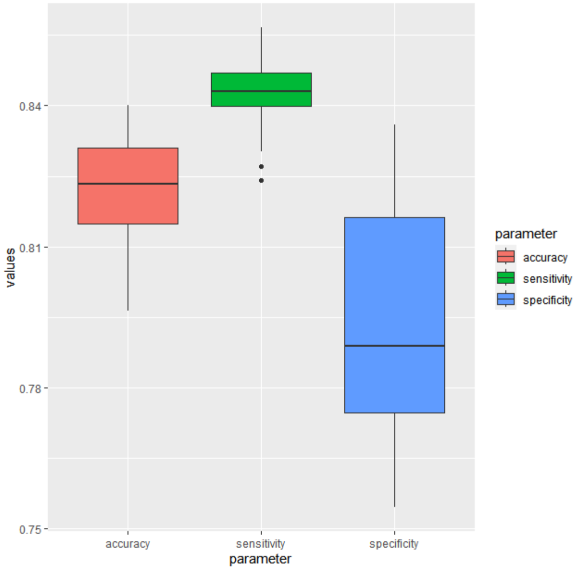


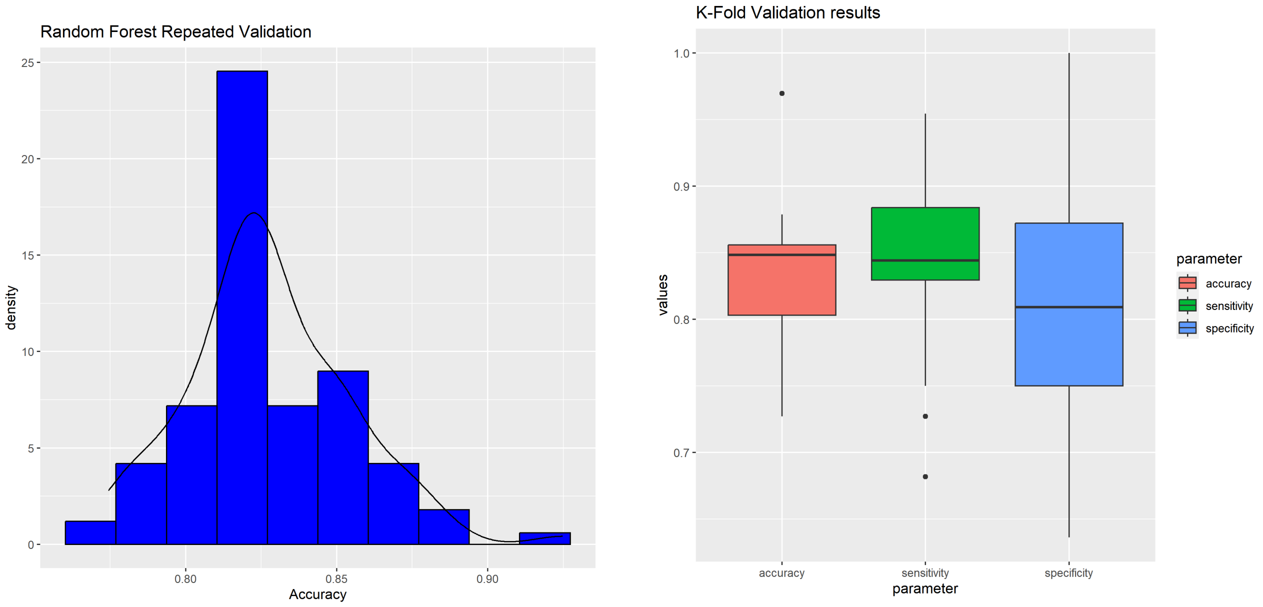
Figura 9: BoxPlot “Hyper Parameter Selection”

To select the best random forest to use, hyper parameter selection was used. There are 8 variables, so m was tried for 2-7. For the number of trees, 100-600 were tried, by 50 tree intervals.

This model has a fairly wide range of potential values, especially for the specificity. Accuracy is maximized at about 84%, with sensitivity’s max being closer to 86% and specificity being lower and more widespread than both, maximized at almost 84%. The best overall models came from m = 3 and ntree = 450, so this is the model that was used as the best random forest tree.

2.2 Validation (Repeat, K-Fold)

First, repeat validation was done on the selected model. Simple validation was not done as repeat validation covers many rounds of simple validation. The validation was carried out 100 times, and the accuracy, specificity, and precision were stored for each execution. The average accuracy was 0.82759, the average precision was 0.84166, and the average specificity was 0.806926. These numbers roughly line up with the medians on the graph of the models from hyperparameter selection.
For the K-Fold Cross validation results, values yet again line up roughly with the repeat validation results, although they were slightly better. The histogram shows all three have large ranges, with Accuracy being about .84, as the highest peak is at 8.5 but there is a lower significant peak at .75, so ultimately the accuracy is about .84. The sensitivity is a strong .82, and the specificity is less concentrated and peaking at around .8.



Figures 10 & 11: Density Histogram & Boxplot for the Best Random Forest Validation

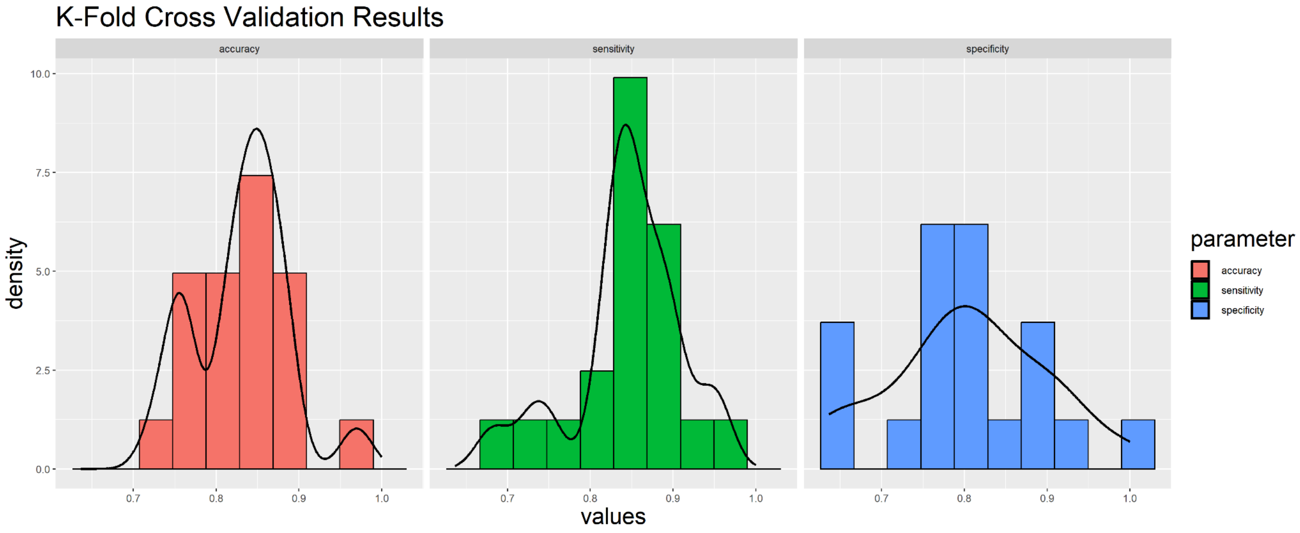


Figure 12: Parameters results of "K-fold cross validation" of the best Random Forest

Conclusion: Best Model Selection

At the beginning of the explanation of Machine Learning strategy it was stated that Accuracy would be the criteria-guideline selection. The reason is that, as it has been commented, the main objective is to maximize the correct predictions of overall observations on the testing Titanic Datasets. Following this criteria selection, the best model considered is the random forest model that was controlled by m = 3, i.e, considering three variables as candidates of every grown tree split, and by 450 trees.

In the previous Exploratory Data Analysis it was suspected that “Sex” would be the most important variable to predict the Survival Rate of Titanic and that “Age” would be relevant to explain the differences in the probability of survival among males. Furthermore, it was mentioned that “Sibsp” influenced the Titanic outcome in different ways depending on “Sex” & “Age” so “Embarked” did it too according to “Sex” and “Fare”. As a result, in the preprocessing data they have been included: it was considered important to use the trained models for determining until what node they started to be relevant for splitting. However, the cabin floor was revealed as not really relevant, so it was decided to transform “Cabin” into a boolean factor for inspecting a potential relationship.

Even though the decision tree did permit the interpretation of these hidden relationships that appeared to some extent in the EDA, the best Random Forest Model shows a higher accuracy than the only tree. Therefore, in the trade-off between the interpretability & the accuracy of the model, it was decided that given the task of predicting future observations, the model that provides the highest accuracy should be chosen. Thus, after validating both the pruned classification tree and the random forest controlled by the optimum, in terms of accuracy, selected hyperparameters; the forest performed better for achieving the objective. That is the reason why it has to be preferred, because the aim of this project is to classify new observations in an appropriate way and to accurately estimate the Titanic Survival rate.