Second Assignment
Machine Learning Models
Titanic

1. Cleaning the data

2. Making the best tree by Hyperparameter Selection

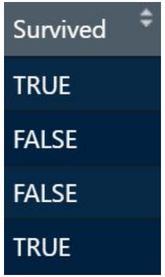
3. Making the best random forest by Hyperparameter Selection

4. Conclusion of the questions of the first Assignment

5. Conclusion

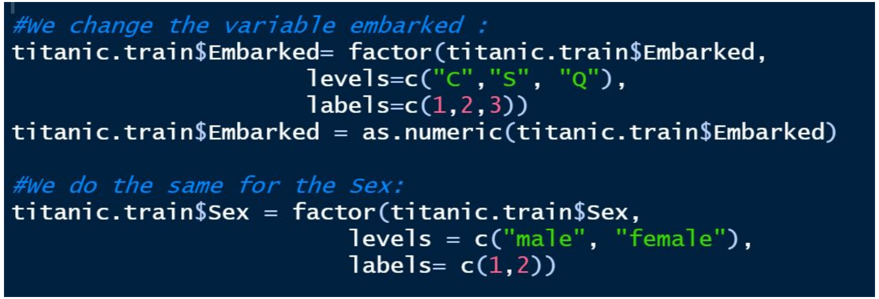
Jaime Salafranca Pardo & Diego Stergar Vega

1. Cleaning the data



Survived
TRUE
FALSE
FALSE
TRUE

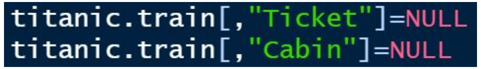
We replace the variable survived filled with (0, 1), with True False or “SURVIVED” and “DIED”



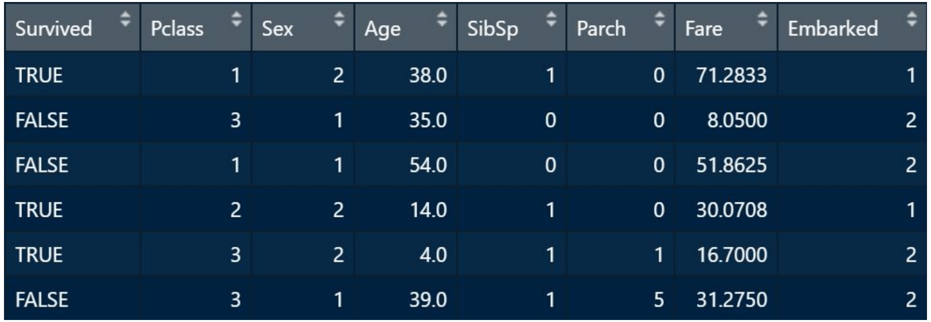
```
#We change the variable embarked :
titanic.train$Embarked= factor(titanic.train$Embarked,
                           levels=c("C","S", "Q"),
                           labels=c(1,2,3))
titanic.train$Embarked = as.numeric(titanic.train$Embarked)

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

We change the **non-numeric** variables in numeric



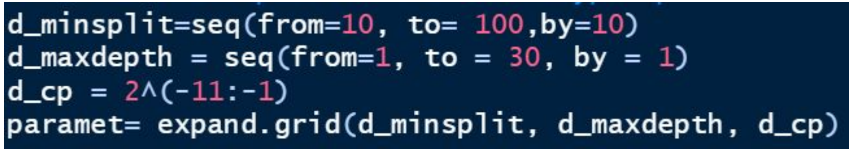
```
titanic.train[,"Ticket"]=NULL
titanic.train[,"Cabin"]=NULL
```



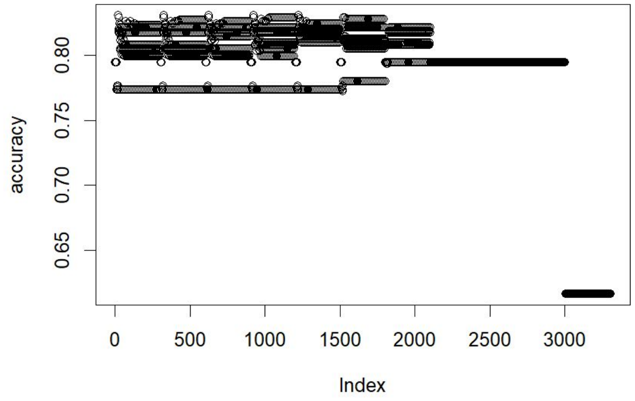
Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
TRUE	1	2	38.0	1	0	71.2833	1
FALSE	3	1	35.0	0	0	8.0500	2
FALSE	1	1	54.0	0	0	51.8625	2
TRUE	2	2	14.0	1	0	30.0708	1
TRUE	3	2	4.0	1	1	16.7000	2
FALSE	3	1	39.0	1	5	31.2750	2

- The change of the variable Sex is not necessary.
- The variable **Ticket** doesn’t seem to be relevant
- We decide not to use the variable **Cabin**, as in the First Assignment
- We don’t replace, by the mean, the values for **Fare = 0** .
- We **standardize** to see if the accuracy is the same, but **we don’t use it.**

2. Making the best tree by Hyperparameter Selection



```
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)
```

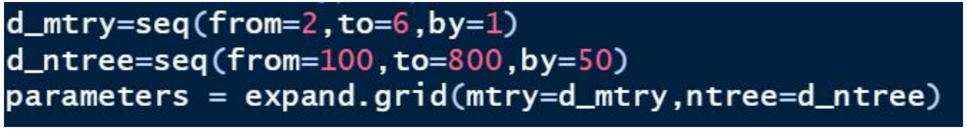


-Find the **accuracy**, of each parameter thanks to the **k-fold cross validation**.

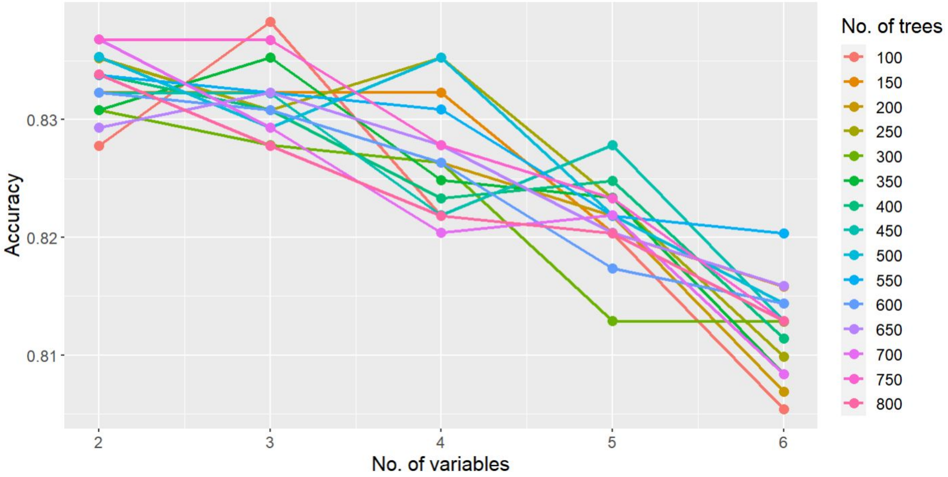
-Find the values of the **best hyper parameters**.

-Create the tree and plot it

3. Making the best random forest by Hyperparameter Selection



```
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)
```

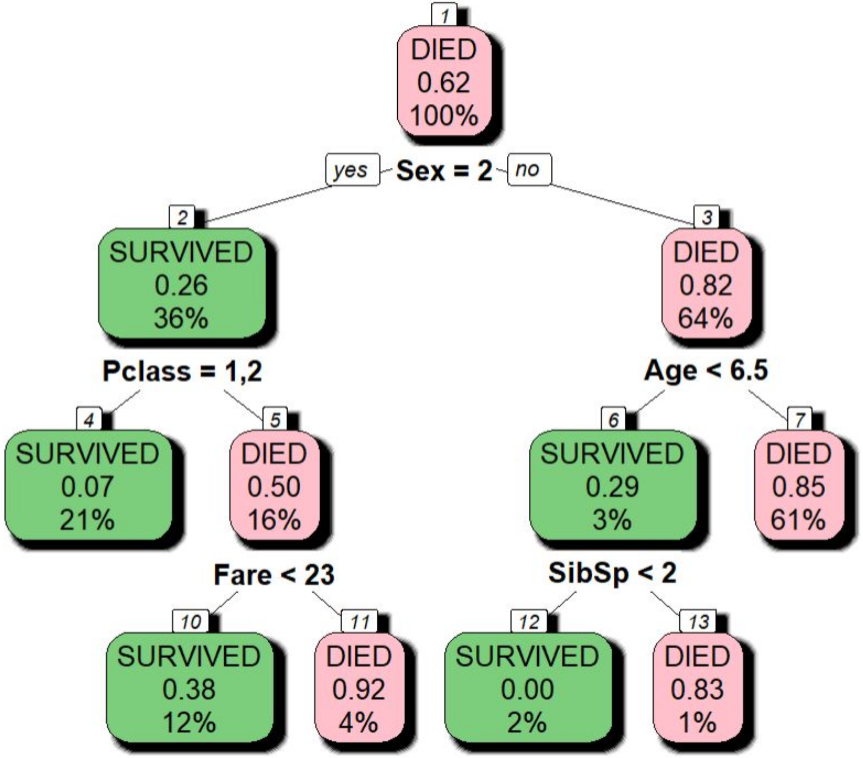


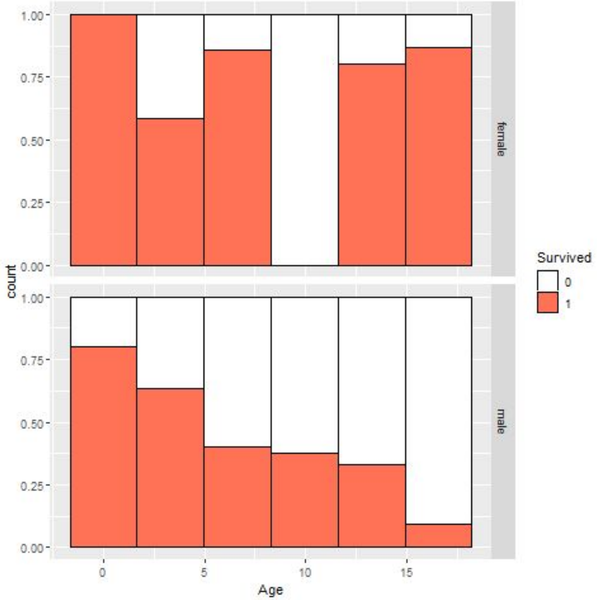
-Create the random forest and find the one with the best accuracy

-Find the hyper parameters for the best random forest

-Create the forest and the **model**.

4. Conclusion of the questions of the first Assignment





-Relationship between **survival**, **gender** and **minors**

-Relationship between **survival** and the **Family size**.
-Relationship between **money** and **Survival**.
-The **stereotype of a Survival** and the **stereotype of a non Survival.**

5. Conclusion

PROS OF MACHINE LEARNING MODELS

- Helpful to **predict** values
- Shows relations between variables that are difficult to find
- Create a **hierarchy** between the variables by the order of influence

CONS OF MACHINE LEARNING MODELS

- Sometimes difficult to visualise the meaning of some relation between variables.
- Some calculus are very **time consuming**, specialty for random forest.

OUR CONCLUSION

-Mix **Exploratory Data Analysis** and **Machine Learning Models** is the best way to find clear and relevant information.