UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

Advanced Engineering Data Analysis
MUEI, MUEA, and MASE
Prof: Daniel Fernández

Exercise 2

Topics: 1 problem about Classification and Regression Trees and Random Forest.
Weight in Exercises grading: 50%

Notes:
- This activity must be done in groups
- The report must have 10 sheets at most (5 pages). You can add appendices.
- Written in 11-12 font size.
- Submitted as a report (PDF format) to Atenea (under Exercise 2 folder).
- **Each member** of the group must submit the report.

The due date is at 23:59 on March 22, 2022

1. CART & Random Forest

For the data set "Ionosphere" from the "mlbench" package:

Note: Use the following R command to load the data set and the *mlbench* library.

```
library("mlbench")
data("Ionosphere")
```

The data set has 34 predictors (V1-V34) and an factor/group variable (Class).

(a) Select randomly a test and training data set. The training and the test data sets must include the 80% and the 20% of the whole data set, respectively.
(b) Fit a classification tree with the training data set selecting a complexity parameter of 0.01. Use the 34 predictors to fit the outcome variable "Class".
(c) Plot the classification tree. How many splits and leaves have this tree?
(d) Use the "one standard deviation rule" (1SD) to decide if the tree obtained in (b) is optimal or not. How many splits have the optimal tree? If the classification tree obtained in (b) is not the optimal tree according to 1SD, then prune it, and plot the pruned tree with an optimal number of splits.
(e) Use the test data to assess the prediction power of the optimal classification tree you obtained. What is the percentage of the total misclassified observations in the test data set? (i.e., the tree predicts "Class=bad" when the true value is "Class=good" (from the test data set) or the tree predicts "Class=good" when the true value is "Class=bad"?
(f) Apply a Random Forest and do changes (e.g., the *number of trees* the random forest can use and/or the *number of variables* used for each tree) to improve the prediction with respect to final classification tree obtained from the previous sections.