Statistical Learning

Bachelor in Data Science and Engineering

Second Homework: Supervised Learning

Deadline: December 17, 2023 at 13:00
Upload to Aula Global: notebook (.Rmd and .html) and data

uc3m | Universidad Carlos III de Madrid

General Objectives

- Apply supervised-learning tools to an open data set
- Select two important variables as targets: one categorical (for classification) and the other one numerical (for regression). Or select just one numerical and transform it into a categorical (as in the airline delays' case study)
- The HW is divided in two parts: classification and advanced regression
- In the two parts, you need to develop all the supervised tools you can (and tuning hyper-parameters) to predict the targets. It is also important to explain the relations between the target and the predictors
- The bigger and more diverse the dataset, the better
- Get convenient insights and conclusions from the performance of the tools, considering ideas from risk learning

Evaluation

1. Data preprocessing and visualization tools: 2 points
2. Classification (emphasis on interpretation): 1 points
3. Classification (emphasis on prediction): 3 points
4. Advanced Regression (emphasis on interpretation): 1 points
5. Advanced Regression (emphasis on prediction): 3 points

Instructions

- The assignment can be done individually or in pairs, it is your choice.
- In case of pairs, the assignment will be uploaded by both members of the pair.
- In case of pairs, the evaluation will be the same for both members.
- **No work will be accepted after the deadline**.

Important: you can take inspiration from other notebooks but you need to cite always the source

Remember the deadline: December 17, 2023 at 13:00