Task 1

Decision Trees

2024-04-10

Table of contents

Delivery deadline: April 24 2024

Data description

File "HFCRD.csv" contains the *Heart Failure Clinical Records Dataset* with information on 299 patients with advanced heart failure in whom the following variables were analyzed:

- **age**: age of the patient (years)
- **anaemia**: decrease of red blood cells or - hemoglobin (boolean)
- **cpk**: level of the CPK enzyme in the blood (mcg/L)
- **diabetes**: if the patient has diabetes (boolean)
- **ef**: ejection fraction: percentage of blood leaving the heart at each contraction
- **hbp**: if the - patient has hypertension (boolean)
- **platelets**: platelets in the blood (kiloplatelets/mL)
- **sc**: level of serum creatinine in the blood (mg/dL)
- **ss**: level of serum sodium in the blood (mEq/L)
- **sex**: female/male (binary)
- **smoking**: if the patient smokes or not (boolean)
- **fup**: follow-up period (days)
- **death_event**: the patient deceased during the follow-up period (boolean

Questions

1. Do a short exploratory data analysis in order to know some characteristics of each variable

    - variability,
    - percentage missing values
    - …
    - Also, you can apply a multivariate techniques such as PCA, clustering, …

2. Separate the data into 2 sets: training set (2/3) and test set (1/3). Use this partition in the training phase (and validation phase if necessary) and the test phase of each of the sections that are presented below. Use the value `1234` as random seed to do the partition.

3. Fit the following statistical models:

    1. A classification tree to predict survival.
    2. A logistic classifiers with the same goals ad assessment.
    3. Build a third predictive model of your choice. Choose one that you consider most appropriate for the problem you are analyzing.

4. For each predictor do the required tuning, if needed, train the model and do a proper test-based evaluation.

5. Compare the predictors based on the adequate metrics.

Important remarks

- Answer the questions in a reasoned way, adding the necessary comments, not just only the code. Notice that some questions may seem ambiguous. They are indeed so that you have space for creativity while answering the questions.
- Provide your report in a reproducible manner
    - A readable report in pdf with explanations, results and discussions. Explain minimally what you do and use references to complement your explanations.
    - The Rmarkdown document or the Python norebook used to generate it.
    - Use relative paths instead of absolute paths to read / write files, to make it easier to run the code outside of your computer.

Delivery / Deadline

Upload a zip file to Atenea before the deadline ends. This file should have no sub-folders and contain only

- the R/Rmd/python/pynb file used as template for the report,
- the output reports in pdf or html files.

Tthe reports should have the same name with the following pattern:

`Group_XX-LastName1-LastName2-LastName3.extension`