ISEL - DEETC

FINAL PROJECT – A1

Subject: AMD – Aprendizagem e Mineração de Dados
Course: MEIM | MEIC

Goal: Classify using a dataset from the machine-learning repository at UCI (University of California, Irvine). Use the implemented classifier in a "bigger" dataset with some missing values (in features). Use Orange DM to explore additional classification methods.

Scenario A1: Your classifier is ready for new challenges! Now, the "FungiData" institute is asking you to analyze a dataset with information regarding de edibility of mushrooms (i.e., if they are good to eat or poisonous). The dataset is available in this folder and was downloaded from `https://archive.ics.uci.edu/ml/datasets/Mushroom`.

Project Items:

To overcome this challenge you need to develop the following items:

1. Analyze both the dataset and its description available in this folder: `dataset_long_name_ORIGINAL.csv` and `dataset_description.txt`.

2. Transform the dataset (`dataset_long_name_ORIGINAL.csv`) into the proper `.tab` format (for Orange processing). Achieve such transformation with whatever method you prefer and describe such method in project report. Build a file similar to the one provided: `dataset_long_name_PTS_INPUT_v01.tab`.

3. Apply your 1R algorithm and save into a file, named `oneR_OUTPUT.txt`, the output where the selected attribute (chosen by 1R) shows the following information:

   ```
   ( attr, valueAttr, valueTarget ) : (error, total)
   ```

   where `attr` is the name of the selected rule attribute, `valueAttr` is each value for `attr` and `valueTarget` is each value for the class; the `error` is the number of wrong classifications for the tuple `(attr, valueAttr, valueTarget)` tuple and `total` is the total number of instances for that same tuple.

   It is important that you provide this information with the proper format because it will be compared with a reference solution (during the final discussion of this projectA/A1).

4. Use your knowledge of Orange DM and build a workspace where you can visualize (and show me) the: a) total number of instances, b) number of instances of each class, c) a "tree" classification, d) a "random forest" classification. e) make "predictions" using those two classifiers. Notice that all thses "Orange goals" can be achieved via "visual compositions", e.g., using Orange operators such as "Data Table", "Select Columns", "Distributions", "Tree", "Tree Viewer", "Predictions", "Random Forest", "Pythagorean Forest", "Test and Score".

Important dates (deadlines) and deliverables:

- 13.DEC.2021 – Deliver all the project elements described below (in the next "Rules" paragraph).

Rules:

- Extend your projectA report with at most 5 pages and provide, for the discussion purpose, a printed version.
- Join project A and A1 and deliver, in electronic format (file named AMD_XX.zip, where XX is the working group number), the project report and all the information regarding the implemented system; the ".ppt" (powerpoint), ".bat", ".exe", ".py", ".tab" (data) and any other file needed to properly execute your solution.