Practical Assignment
Sébastien Harispe

Machine Learning Assignment:

Model Evaluation and Comparison

This document defines the assignment on which you'll be evaluated.
Recall that this assignment has to be done with three of your colleagues (teams of 4 students);
it will contribute to 1/2 of your final mark.

Aim of the assignment

The aim of this assignment is to evaluate your ability to use Machine Learning for solving a given predictive problem of interest. We will not constrain you to solve a specific problem or to use a specific dataset. You can therefore construct or generate a dataset of your choice. Feel free to consider a dataset on a topic you like, whatever it is. Note however that you've only mainly be introduced to simple regression and classification problems and that you don't necessarily know how to represent and process complex inputs (video, sound, image). We ask you to:

- Generally define your task
- Analyse the corresponding dataset.
- Define the problem as a Machine Learning problem.
- Define the methodology and protocol you will use to tackle the problem.
- Implement various approaches that are suited to this problem.
- Discuss the results you obtain.
- Motivate the selection of your final solution.
- Deliver that solution in the form of a predictive model that can be used on new input data.

Several datasets can be downloaded at:

- UCI Machine Learning repository
- Kaggle
- opendata.aws

Do no hesitate to generate your own datasets. We will not evaluate the selected topic per se. Bonuses will be given for the dataset construction if you decide to spend time working on that aspect. Be careful however not to spend too much time building the dataset. Note that we are well aware of the fact that numerous comparative studies of several models are already available for numerous datasets - do not forget to cite used sources.

Detailed notes on Decision Trees and Random Forest: You will be asked to include Decision Trees and Random Forest models in your comparison. In addition, we ask you to provide an Appendix explaining these two models. Details will be provided for specific cases (modelling, training procedures...). Your explanations will be presented as detailed notes, up to 5 to 7 pages. Do not forget to mention references.

The evaluation procedure for the practical evaluation as well as the honor code are recalled below.

Evaluation procedure

This project will give you the opportunity to deeply cover a topic related to the domain - you'll have to write a dedicated report/notebook (up to 15 pages + Appendix, 5 to 7 pages).

The due date for this assignment is set to: **05/04/2023 23:59** (April 05)

Late penalties: 2 points/day.

Results must be zipped in an archive named “[ML1] NAME_1 – NAME_2 - ....zip” and posted on campus. They must only contain notebooks and additional data (link) and pdf files.

Honor code

This assignment is a practical evaluation of your skills: 'your' refers to you, you as an adult. You can talk with other teams about the assignment in accordance with expectations of academic integrity; you must nevertheless (i) offer your own solution, (ii) be able to explain and defend it. No extreme measures will be taken to prevent cheating. However, any suspicion of unpermitted behavior will be investigated. Violation of the academic integrity would imply a 0 mark and would be notified to the administration. No difference would be made between those giving unpermitted aid to others and those benefiting from such an aid.

In [ ]: