INTRODUCTION TO MACHINE LEARNING

Project (**by groups of 2 students**)

Objectives: apply your skills in Machine Learning and practice Python programming

By October, 31st, send me an email with the members of the team.

Input : the dataset **waveform.data**
https://archive.ics.uci.edu/dataset/107/waveform+database+generator+version+1

- 5000 data
- 3 classes of waves
- 21 attributes all of which include noise
- Optimal Bayes classification rate = 86% accuracy

List of possible experiments to perform (non exhaustive)

- **Tune the best *k* of a *k*NN classifier** by cross-validation (plot the accuracies over the validation subset w.r.t. *k*) from 4000 randomly drawn **training** examples (you will keep apart 1000 waves for the **test** set).

- **Reduce the complexity** by running the Data Reduction algorithms studied in class on the training data. Compare the accuracy (with a 1NN) on the 1000 test waves before and after reduction of the training set.

- Using the original dataset, compare (in terms of time) the two methods studied in class for **speeding-up the calculation** of the 1NN with a brute force 1NN algorithm.

- **Generate artificially imbalancy** in the training data and analyze the impact on the accuracy on the 1000 test waves. Tune k w.r.t. the F-measure and compare the performance with the accuracy.

Output: you are required to provide a report limited to 2 pages using the two-column Latex style from the international conference ICML (title, authors, abstract, core, conclusion)
See: https://media.icml.cc/Conferences/ICML2022/Styles/icml2022.zip
Deadline: upload on claroline connect your 2 pages document by **January, 8th** along with your ipynb jupyter notebook file.