Group assignment of
Machine Learning I

Objective: To develop a Machine Learning (ML) system for solving any prediction/classification problem.

I. General Instructions:

1. This assignment will be developed in groups of 3-4 people.
2. For the implementation of the ML system, you should include the pieces of code implemented in the Notebooks of the first part of the course. This should be included as additional modules of Julia, i.e., Julia script file(s).
3. Any prediction problem is suitable. Bear in mind to consider any publicly available database that can be used with your implemented ML system.
   3.1. The prediction problem must be solved by at least 4 different approaches.
4. The solution of your selected prediction problem should be delivered as latest on 12th December at 23.59 via Moodle (virtual campus).
   4.1. The files to be included are:
   - Selected database.
   - Source code and report of the entire process.
   - Auxiliar code developed in the previous practices.

   4.2. Please keep the following **folder structure** for your deliverable compressed file:
   - In the root folder, the report in PDF with a significant name. In this report, you should explain the selected problem, the objective, the approaches and methodology followed to those approaches, and a discussion of the results. Please, see the criteria evaluation in the next section for detailed instructions.
   - Also in the root folder, the Julia script (main.jl) which has the main part of the code, and which may call other modules with the approaches.
   - A folder with a name like *datasets*, where the data used in the different approaches will be located. Within this folder new folders can be created, and this structure can change as decided by the requirements of the work team.
   - A folder with a name like *utils*, where the code from the previous practices will be placed.
5. An oral presentation of your work will be on December 19th (estimated duration about 15 min per group).

II. Specific instructions and evaluation criteria

In this section, the evaluation criteria are presented in more detail in the form of a list of points for each of the respective elements, i.e., report, code and oral presentation. The scores for each of these elements are also specified.

1. Report (50% of the project mark)
   - Introduction (10% of the project mark). This should include:
     i. Explanation on problem to be solved.
     ii. Description of the dataset.
     iii. Justification of the metric or metrics to be used.
     iv. Explanation of the code structure and how it has been organised.

v. Bibliographic analysis. At least 3 scientific publications related to the problem should be briefly described. These works must be correctly referenced following a specific style for publication in books, journals, conference proceedings or similar publications. Web pages are not considered as scientific publications.

- Development: (30% of the project mark) The students are tasked with investigating various approaches for processing the dataset. The highest achievable score for each attempted approach will be 25% of the total value of this section. Each approach should encompass the following elements:
    i. Description of the database used in this approach. Although in the "Description of the problem" section the data have already been described, it is possible that in each approach may have slightly variations, so it is convenient to have a description including how many patterns have been used, how many entries, classes, etc. Furthermore, the description of the dataset characteristics included in this approach should be supported by graphs and/or images.
    ii. Data preprocessing. Usually referred to data normalization. It is necessary to justify the reason for the type of normalization used, as well as the normalization parameters (minimum, maximum, mean, etc.), or why normalization is not performed.
    iii. Other data related to the experiments to be carried out, such as methodology, number of folds, selection of variables, reduction of the dimensionality, etc.
    iv. Any other material such as graphs showing the data may be of interest for this part.
    v. Results of the experimental part. The clarity of the explanations is valued, as well as the number of experiments. It is necessary in all approaches to perform experiments with the 4 techniques covered in this material (Artificial Neural Networks (ANNs), support Vector Machines (SVM), decision trees and kNN) and test each one with different hyperparameters.
        1. For ANNs, test at least 8 different architectures, between one and 2 hidden layers.
        2. For SVM, test with different kernels and values of C. At least 8 SVM hyperparameter configurations.
        3. For decision trees, test at least 6 different depth values.
        4. For kNN, test at least 6 different k values.
    vi. Additionally, at least one ensemble technique should be applied in each approach, that is, majority voting, weighted majority voting or model stacking. That ensemble must combine at least three of the individual models of the approach.
    vii. All experiments should be reported with the selected metrics and a discussion and comparison of the models obtained.
    viii. Other material which could support the claims, such as explanatory plots or confusion matrices.
- Final discussion (10% of the project mark) The overall process should be evaluated, and the results of the different approaches compared among them.

The spotlight should be put on the conclusions drawn from the development which should be supported by the results obtained.
- For the format of your report, please consider the following author instructions given at: ACM Primary Article Template

2. Code (30% of the project mark)
   The code must, at a minimum, meet the following requirements:
   - Set the random seed to ensure repeatability of the results.
   - Load the data.
   - Extract the features of that approximation.
   - Make a split by holding out a portion of the dataset to test the selected model.
   - Call the *modelCrossValidation* function to perform cross validation with different models and parameter settings.
   - Once a configuration has been selected, proceed to train a new model with that configuration using the entire training dataset. Afterward, create a confusion matrix based on the test dataset to evaluate the model. It is important to note that ANNs will necessitate a hold-out split within the training data since a validation dataset is required for training and fine-tuning.

   Additionally, the following points will be considered in the evaluation as positive elements:
   - The use of separated modules for the functions is preferred and evaluated.
   - At least simple comments for each module and function describing their behaviour.
   - The in-code explanation of the most important part by comments.
   - The use of easy following names for the variables and functions. Following some convention will be a plus.
3. Oral Presentation (20% of the project mark)
   - The clarity of the explanation as well as the presentation timing (15 minutes) will be considered as a high point.
   - The explanation should contain at least contain a brief explanation of the problem, the pipeline used to perform the experiments, the results and the comparison of the approaches.
   - The answers to any question performed by the audience (5 minutes).

Major Penalties may be applied to the final grade if serious errors of concept are found in the source code, such as follows:

- Code that does not run (20% of the maximum mark of the project).
- Results do not match those shown in the report (10% of the maximum mark of the project).
- Using test samples to train the models (10% of the maximum mark of the project).