UNIVERSIDADE DE AVEIRO
DEPARTAMENTO DE ELECTRÓNICA TELECOMUNICAÇÕES E INFORMÀTICA

PROJECT 1 2024/2025 – Instructions

Deadline for Project 1 submission: 27/November 2024
Submit to elearning: report (pdf+ source files); presentation slides; implementation code

Each group of two students is supposed to work on one project topic.
You are strongly encouraged to propose a machine learning problem you would prefer to work, not listed below, that may reflect better your interests. Please, discuss your idea with the instructor.
Choose a project with some usefulness for the society !!!

I. PROJECT GOALS

The goal of this project is to apply suitable machine learning algorithms learned in class or self-learned to solve a specific data science problem (classification, regression, clustering). Represent the results in graphical/table formats and make analysis and conclusions.

II. PROJECT PROPOSALS

Sign language understanding

Hand gestures and sign language are the most commonly used methods by deaf and non-speaking people to communicate among themselves or with speech-able people. However, understanding sign language is not a universal skill. For this reason, building a system that recognizes hand gestures and sign language can be very useful to facilitate the communication gap between speech-able and speech-impaired people.

Project proposal 1: Identification of digits from sign language images

Data source: https://www.kaggle.com/datasets/ardamavi/sign-language-digits-dataset
(Google: Sign Language Digits Dataset)

Project proposal 2: American sign language understanding
Data source: https://www.kaggle.com/datasets/datamunge/sign-language-mnist
(Google: Sign Language MNIST)

Project proposal 3: Mammographic Mass Data Set
This data set can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. The aim is to discriminate benign from malignant cases assuming that all cases with BI-RADS assessments greater or equal a given value (varying from 1 to 5), are malignant and the other cases are benign.

Data source: http://archive.ics.uci.edu/ml/datasets/mammographic+mass

Project proposal 4: Heart Disease Data Set
This dataset contains 4 heart disease related datasets. For the present project you will use the Cleveland database and the referred subset of 14 features form a total of 76 attributes. The goal is to distinguish presence (values 1,2,3,4) from absence (value 0) of heart disease in the patient.

Data source: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Project proposal 5: Bank Marketing Data Set
The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe or not a term deposit.

Data source: https://www.kaggle.com/janiobachmann/bank-marketing-dataset

Project proposal 6: The German Traffic Sign Benchmark (GTSB)
GTSB is a multi-class, single-image classification challengeheld at IJCNN 2011. Automatic recognition of traffic signs is required in advanced driver assistance systems and constitutes a challenging real-world computer vision and pattern recognition problem. A comprehensive, lifelike dataset of more than 50,000 traffic sign images has been collected. It reflects the strong variations in visual appearance of signs due to distance, illumination, weather conditions, partial occlusions, and rotations. The dataset comprises 43 classes with unbalanced class frequencies.

Data source: https://www.kaggle.com/meowmeowmeowmeowmeow/gtsrb-german-traffic-sign

Project proposal 7. Kaggle Credit Card Fraud Detection
https://www.kaggle.com/mlg-ulb/creditcardfraud

Recommended Data Repositories:

- Kaggle Data Repository : https://www.kaggle.com/datasets
- UCI Machine Learning Repository : https://archive.ics.uci.edu/ml/index.php

III. PROJECT ASSESMENT (25 % of the final grade)

1. **Report.** The project is evaluated based on a submitted report (IEEE Latex format). The work done by each student has to be explicitly specified.
   All project's files (pdf and Latex files of the report, the presentation slides and the code implementing the algorithms) are **submitted to the elearning.ua.pt page of the course** in section *SUBMISSION – PROJECT 1* in a compressed format having the following name: P1_ML2024_XXXXX_YYYYY (where XXXXX and YYYYY are substituted by the student number of each student.

2. **Oral presentation** of the report in class (about 10-15 min.).

IV. Evaluation criteria (total score 20)

1. Report content (12):

- Data description and preprocessing (if necessary normalization, feature selection, transformation, etc.). Motivation for choosing the particular problem.
- Data visualization (histograms, box plots, other plots).
- Short description of the implemented ML models.
- Model training (data splitting – train, validate, test, k-fold Cross validation). Visualize graphically the cost function trajectory over iterations. Training with regularized or non-regularized cost function.
- Model hyper-parameter selection – systematic approach instead of randomly chosen values.
- For a classification problem, you need to present the confusion Matrix (accuracy, precision, recall, F1 score, etc.).
- Performance comparison between the models.
- Results in graphical or table formats.
- Conclusions.
- Problem complexity.

2. Report formatting (2) :

- IEEE Latex format, affiliation (Department, University, subject, course instructor), abstract, keywords, work load per student.
- Sufficiently detailed report.
- References, reference citation in the report.
- Clear figures (title, legends, axis labels) and tables referred in the text.

3. Oral presentation (3)

- Slide Organization, slide numbers, affiliation.
- Clear and convincing presentation by both students.

4. Novelty and contributions (3)

- Compare your solution with the works of other authors (published references) , try to propose a better solution, e.g. improve the performance of the ML model in solving the problem you work with.