POLITÉCNICO DE LEIRIA
ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO

Mestrado em Ciência de Dados
Big Data

PROJECT

1. Objectives

In this work, students should:

- Choose a dataset thought for a classification, regression, forecasting or clustering task;
- Follow the typical data analysis steps using Big Data tools:
  - Data loading (using at least two different data formats);
  - Data exploration (e.g., computing correlations, averages, standard deviations, counts, generating pertinent charts data visualizations);
  - Data preparation (e.g., data cleansing, field selection, adding new fields);
  - Data modeling: you should test at least three different learning algorithms;
  - Model evaluation: you should choose the appropriate metrics to compare the performance of the models;
  - Model deployment: develop a module to run the model on data ingested via streaming.
- Write a report describing all the steps, all the experiments and obtained results in detail. Alternatively, students may deliver well documented notebooks. The advantages/disadvantages of each file format should also be discussed regarding the computational performance.

The datasets should be chosen by the students, and then validated by the teachers. Each group should send an email to both teachers (carlos.grilo@ipleiria.pt; joao.f.ramos@ipleiria.pt) with the following elements:

- Members of the group;
- Link to the dataset and a brief description, which should include:
  - Number of instances/samples;
  - Brief description of the attributes, namely their types (datasets with both nominal and numerical attributes are preferred);
  - Problem that is supposed to be solved, namely, if it is a classification, regression or clustering problem;
  - Description of the objective class/dependent variable, if applicable.

Note that these elements should also be included in the final report/notebook(s).

The email with this information should be sent until November 14th. Teachers will answer as soon as possible to each group.

2. Links where to look for datasets (just some suggestions)

1. https://www.kaggle.com/datasets (Kaggle Datasets)
2. https://data.world/ (Data World Datasets)
3. https://archive.ics.uci.edu/ml/datasets.php (UCI Machine Learning Repository Datasets)
4. https://imerit.net/blog/the-60-best-free-datasets-for-machine-learning-all-pbm/ (Imerit Datasets for Machine Learning)
5. https://pub.towardsai.net/best-datasets-for-machine-learning-data-science-computer-vision-nlp-ai-c9541058cf4f (Towards AI Machine Learning Datasets)

3. Assessment

10% - Dataset/problem complexity
5% - Data loading
15% - Data exploration
15% - Data preparation
15% - Data modelling
10% - Models evaluation
5% - Model deployment (data streaming)
15% - Documentation (report or notebooks documentation, including justifications of decisions taken and conclusions)
10% - Extras

Examples of extras:

- Use more than three learning algorithms to generate the models;
- Dataset with multiples files/tables that may need to be integrated/joined/combined.

Deadlines, dates, rules, and instructions

1. Project delivery deadline: **January 9th, 2025, 23:59**;
2. Oral exam date: **January 11th, 2025, 10:00**;
3. The project should be developed in groups of 3 students. Groups with more than 3 students are not allowed;
4. The report should be written using the template provided in the course Moodle page. Alternatively, students may deliver well documented notebooks;
5. The project should be delivered as a zip file containing all the project elements, including the report, dataset(s) and the notebook(s). The file name should follow the format BigData_Project_#1_#2_#3.zip, where #1, #2 and #3 should be replaced by the student

numbers of the group elements. The report should be in the pdf or notebook format and its name should follow the same format (with pdf/ipynb extension).

6. Delivery method: through Moodle.