«Algorithms for massive datasets» «Statistical methods for ML»

Joint project for 2021-22 (validity: until 31/05/2023)

The project is based on the analysis of the «Airline Delay and Cancellation Data, 2009 - 2018» dataset published on Kaggle. The analysis can also consider a part of the dataset (i.e., a subset of the available attributes or a subset of the items to be processed—e.g., only those related to a time interval, or both).

The task is to implement from scratch (without using libraries such as Scikit-learn) a classifier based on logistic regression and to train it to predict whether or not a flight will be canceled, only on the basis of information available at flight departure. As an alternative, the prediction can concern flight diversion, or either canceled or diverted flights.

Important: the techniques used in order to infer the classifier should be time and space efficient, and scale up to larger datasets.

The project can be carried out individually, or in groups of two students. Code should be written in Python 3 (different choices must be preliminarily agreed with both instructors).

The project report, preferably written in LaTeX, will be evaluated according to the following criteria.

- Correctness of the general methodological approach.
- Reproducibility of the experiments.
- Correctness of the approach used for choosing the hyperparameters.
- Scalability of the proposed solution.
- Clarity of exposition.

The report should contain the following information.

1. Parts of the dataset which have been considered.
2. Data organization.
3. Applied pre-processing techniques.
4. Considered algorithms and their implementations.
5. How the proposed solution scales up with data size.

6. Description of the experiments.
7. Comment on the experimental results.

The report must also contain the following declaration: *"I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study."*

If the proposed solution is based on the ones published in Kaggle, this must be clearly stated, and the report should explain the differences and compare the experimental results.

The project should be made available through a public github repository, containing code and report. The dataset should not be added to the repository, but downloaded during code execution, for instance via the kaggle API (https://github.com/Kaggle/kaggle-api). Code should be implemented using a jupyter notebook executable on Google colab, possibly adding a badge/link directly from the repository to the colab version of the notebook.

Once the project has been finalized, students should send an email to Prof. Cesa-Bianchi (nicolo DOT cesa-bianchi AT unimi DOT it), Prof. Malchiodi (malchiodi AT di DOT unimi DOT it) and Dr. Clerici (giulia DOT clerici AT studenti.unimi.it), specifying:

- their names and student IDs,
- the program they are enrolled in,
- a github link to the project.

This project is valid for the academic year 2021/22, according to the following deadlines: 12/06/2022 (June session), 07/07/2022 (July session), 15/09/2022 (September session); Deadlines for the forthcoming sessions will be communicated later on.

After the project is evaluated, the teachers of the two courses will separately schedule an appointment for the oral discussions.