­Instituto Superior de Agronomia, ULisboa

Master’s in Green Data Science 2023-2024

Practical Machine Learning/Aprendizagem Automática Aplicada

Instructor: Manuel Campagnolo

**Final project guidelines**

Project proposal (due June 6)

Your project proposal should include the following information:

• What is the problem that you will be investigating? Why is it interesting?

• What are the challenges of this project?

• What dataset are you using? How do you plan to collect it? You can use your own data, if you have, or gather data from online data repositories.

• What method or algorithm are you proposing?

• How will you evaluate your results? What kind of analysis will you use to evaluate and/or compare your results (e.g. what performance metrics or statistical tests)?

Format of the project proposal: Your proposal should be a PDF document, giving the title of the project, the project category (tabular data, image classification, image segmentation, other—to specify), the full names and student IDs of team members (1-2 team members) , and a 300-500 word description of what you plan to do.

Submission (due June 28, 2024): Create a GitHub repository which contains your report, and a separate notebook or script with the code. In alternative, you can create a notebook that combines both. Submit the repository’s URL in Moodle

1. The report should contain a comprehensive account of your project. The report is expected to be thorough, yet concise, and organized by sections like “introduction”, “data”, “methods”, “results”, “discussion and conclusions”, and “references”. Broadly, it should cover the following:

* Good motivation for the project and an explanation of the problem statement (you can re-use the project proposal);
* A description of the data. It is possible that your data is not ready to be used, due to data errors, inconsistencies and formats. Perform and document the necessary data cleaning and transformation steps to prepare your data. Also identify the data types. Describe Data Cleaning, Feature Selection and Feature Engineering;
* Describe training, validation and test sets and how the data is organized.
* Which ML model of models are used to address the problem: describe hyper parameter and architecture choices that were explored
* Presentation of results: either tabular or graphical
* Analysis of results
* Any insights and discussions relevant to the project
* References
* Contributions: include a section that describes what each team member worked on and contributed to the project.

Format: A ~4-6 pages document, which extra pages for appendices and references if needed (but the main document should be self-contained)

1. Code: Python notebook or script with the code.

Grading (up to 10): The final report will be judged based on the following criteria.

* clarity of the report,
* relevance of the project to topics taught in class,
* the novelty of the problem,
* technical quality
* significance of the work.

Useful links:

* The Kaggle Machine Learning Project Template <https://www.kaggle.com/general/187601>
* Kaggle ongoing competitions:

<https://www.kaggle.com/competitions>

Examples of previous projects:

* Identification of Greenhouses with Satellite Images (Image segmentation)
* Deteção de doenças em folhas de milho através de imagens (Image identification)
* Condicionantes socioambientais para as piroregiões de Portugal continental (tabular data, clustering)
* Predicting covid-19 deaths in Portugal (tabular data, classification)
* App to help consumers to know more about the products they’re considering to buy at a grocery store (image classification + database)
* BirdCLEF Competition (Kaggle). Identifying Eastern African Bird Species by Sound: develop machine learning models capable of accurately identifying bird species in Eastern Africa based on their sound recordings (sound recording classification)
* Predicting crop production from country, year, yield, crops, rainfall, temperature and pesticides with data from FAO and the World Data Bank (tabular data, regression)
* Identify grapevine varieties from images (image classification)
* Identify from cellphone images the occurrence or not of trees in the foreground of the image (image classification)