Instituto Superior de Agronomia, ULisboa

Master's in Green Data Science 2024-2025

Practical Machine Learning/Aprendizagem Automática Aplicada

Instructor: Manuel Campagnolo (ml@isa.ulisboa.pt)

TA: Dominic Welsh (djwelsh@edu.ulisboa.pt)

Final Project Guidelines

Project Proposal (Due June 6, 2025)

Your project proposal should include the following information:

- **Problem Statement**: What problem will you be investigating? Why is it interesting?
- **Challenges**: What are the challenges of this project?
- **Dataset**: What dataset are you using? How do you plan to collect it? You can use your own data or gather data from online data repositories.
- **Method or Algorithm**: What method or algorithm are you proposing?
- **Evaluation**: How will you evaluate your results? What kind of analysis will you use to evaluate and/or compare your results (e.g., performance metrics or statistical tests)?

Format: Your proposal should be a PDF document or a markdown (MD) file in your Github repository. All group members should submit the same repository link, regardless of who owns the repository. The proposal should include the following:

- Project title
- Project category (e.g., tabular data, image classification, image segmentation, other—please specify)
- Full names and student IDs of team members (ideally two members)
- A 300-500 word description of your project plan

Submission (Due June 30, 2025) – three steps

Step 1 of 3 (for the group)

Create a single GitHub repository with a readme file. The repository will contain the code and your report (see below).

Step 2 of 3 (individual)

Create an individual **short video** (that's no more than 5 minutes in length) in which you present your project to the world, as with slides, screenshots, voiceover, and/or live action. Your video should somehow include your project's title, your name, and any other details that you'd like to convey to viewers

Step 3 of 3 (individual)

Submit in Moodle the URL of the group's GitHub repository and your individual video.

Report, code and data

1. **Report**: Your report should provide a comprehensive account of your project. It should be thorough yet concise, organized into the following sections:
    - **Introduction**: Motivation and explanation of the problem statement (you can reuse content from the project proposal).
    - **Data**: Description of the data, including any necessary cleaning and transformation steps. Identify data types and document data cleaning, feature selection, and feature engineering.
    - **Data Organization**: Description of training, validation, and test sets.
    - **Methods**: Description of the ML model(s) used, including hyperparameter and architecture choices.
    - **Results**: Presentation of results in tabular and/or graphical form.
    - **Analysis**: Analysis of results, including insights and discussions relevant to the project.
    - **Deployment** (optional): possibly as an app
    - **References**: List of references used.
    - **Contributions**: A section detailing each team member's contributions to the project.

    Format: If in the pdf format, a ~4-6 page document, with additional pages for appendices and references if needed (the main document should be self-contained). If you prefer this can be included in the same notebook as the code, but in that case be sure to include sections and subsections, so a table of contents and links within the notebook are available.

2. **Code**: A Python notebook or script with the code available on your GitHub repository.
3. **Data**: Include the dataset in the repository if it can be made available on GitHub, otherwise make the link available in the repository.

Grading (Up to 10 Points, After Discussion)

The final report evaluation will rely on the following criteria:

- **Novelty and Significance**: Importance and originality of the problem (e.g., a Kaggle problem may be significant but might lack novelty). High: address a problem significantly different from the ones discussed in class; Low: do a straightforward classification of a standard tabular data set with no particular difficulties, i.e. which could be addressed just was done in class for the Iris data set for instance.
- **Clarity**: Clear and concise presentation of the report.
- **Relevance**: Relevance of the project to the topics taught in class. High: the report refers the relevant concepts discussed in class for the project at hand; Low: incomplete approach that misses relevant key aspects discussed in class.
- **Technical Quality**:
    - **organization:** modularity, clear pipeline. High: clear pipeline using the appropriate classes from scikit learn or PyTorch; Low: no clear pipeline, making it difficult to understand if the code is correct.
    - **soundness**: use appropriate methods to address the problem.
    - **validation:** follow correct procedures and address possible correlations among data observations

- **Results and Conclusions**: Meaningfulness of the results and conclusions. High: engaging discussion of results; Low: show results with no overall discussion

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 data, 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)
- App for bone fracture identification from X-ray images (image classification)
- Atmospheric Physics Climate Model, based on Kaggle competition "LEAP - Atmospheric Physics using AI" (tabular data, regression)
- PestTracker2: Identificação de praga de mosca-da-fruta (Ceratitis capitata) usando YOLOv8 (object detection on video)
- Estimation of soil salinity in rice production areas within mangroves from PlanetScope imagery with CNNs and RFs (image, regression)
- Identify from cellphone images the occurrence or not of trees in the foreground of the image (image classification)
- Creating a early fire detection model from ICNF fire occurrences (tabular data, classification)
- Genre classification of music tracks using the GTZAN dataset with using feature extraction and CNNs (sound data, classification)
- App for potato pest classification (image, classification)
- Air quality analysis from PM2.5 and PM10 concentration data (tabular data, classification)
- Identificação de pragas e doenças em tomateiros com recomendação de aplicações (image, tabular data, classification)