Instituto Superior de Agronomia, ULisboa

Master's in Green Data Science 2024-2025

Practical Machine Learning/Aprendizagem Automática Aplicada

Instructor: Manuel Campagnolo

**Final Project Guidelines**

**Project Proposal (Due June ..., 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 ..., 2025)**

Create a GitHub repository that contains your report and a separate notebook or script with the code. Alternatively, you can create a notebook that combines both. Submit the repository URL in Moodle. All group members should submit the same link, regardless of who owns the repository.

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 processes.
    - **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, 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**: A ~4-6 page document, with additional pages for appendices and references if needed (the main document should be self-contained).

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

**Grading (Up to 10 Points, After Discussion)**

The final report will be judged based 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).
- **Clarity**: Clear and concise presentation of the report.
- **Relevance**: Relevance of the project to the topics taught in class.
- **Technical Quality**:
  - **organization:** modularity, clear pipeline
  - **soundness**: appropriate methods
  - **validation**
- **Results and Conclusions**: Meaningfulness of the results and conclusions.

By adhering to these guidelines, you will ensure that your project is well-organized and thoroughly evaluated, showcasing your understanding and application of the course material.

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 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)