Project Handout: "Vinho Verde"
INF13262M – Applied Artificial Intelligence

António Anjos
`aanjos@uevora.pt`

Wednesday 22nd November, 2023
(*Draft Version*)

Overview

Welcome to the "Vinho Verde" analysis project! On Moodle you have access to two datasets: `red_wine.csv` and `white_wine.csv`. These datasets contain information about Portuguese red and white "Vinho Verde" wines. The goal of this project is to explore these datasets and to build machine learning models to predict wine quality and alcohol content, as well as to distinguish between red and white wines. The project is divided into three tasks: two classification tasks and one regression task. You are expected to compare two different algorithms in each task.

Tasks

Task 1: Regression (Alcohol)

- Predict alcohol content (red and white datasets merged).
- Use two different regression algorithms.

Task 2: Classification (Red vs. White)

- Distinguish between red and white wines.
- Use two different classification algorithms (they may, or may not, be the same as in Task 1).

Task 3: You decide (Quality)

- Predict wine quality (red and white datasets merged).
- Use two different algorithms.

Other techniques

- **Dimensionality Reduction:** Visual techniques were introduced in the lectures; however, feel free to explore other methods.
- **Noise/Outlier Processing:** Briefly mentioned techniques can be further explored.
- As a suggestion, only when you have a working model, start "playing" with these techniques to improve your results.

Deliverables

1. **Report (8 points)**
    - Maximum 8 pages, including the cover. You can start your report on the title page, if needed.

- Guidelines:
  - Identify the author, the project title, and the module name (i.e., Applied Artificial Intelligence).
  - Cite references, justify decisions, use appropriate visuals.
  - Don't include tables or figures that your text does not refer to.
  - Include all the instructions so that the lecturer can use your trained model file.

2. **Source Code and Trained Models (4 points)**
   - Submission: All code, and a trained model files (saved using `joblib`).
   - Considerations: Clear code, proper documentation.

3. **Oral Presentation (8 points)**
   - Duration: Max 20 minutes (+ 10 minutes for questions).
   - Content: Emphasize key aspects, decisions, results, and conclusions.

Extra points (up to 1.5)

- **Best Model:**
  - Best accuracy or least RMSE, and fastest execution (in this order).
  - +0.25 points for each "winner" model (added to the final mark).
  - Models will be tested on a dataset not available to students. Additionally, these points will only be awarded if the remaining deliverables are of sufficient quality, and justify the performance of the model. In the case there are no clear instructions on how to load and use the model, no extra points will be awarded.

Submission

All the work will be submitted on Moodle as a single `.zip` file. The structure of the `.zip` file should be as follows:

```
<student_number>.zip
├── report.pdf
├── code
│   ├── <source_code_files>
│   └── <trained_model_files>
└── presentation.pdf
```

It is fine to have all the code in a single file, or one file for each task. However, in the former case, make sure the code is logically organized (e.g., using appropriately named functions). In the latter case, make sure the code files are named in a way that makes it easy to identify their tasks. In all cases, make sure the code is properly documented.

Name the trained model files in a way that makes it easy to identify (e.g., `task1_model1.pkl`, `task1_model2.pkl`, `task2_model1.pkl`, etc.). Additionally, provide a description for each file in the report, as well as instructions on how to load and use them.

Don't submit the datasets, or any other files not mentioned above. The lecturer will not use any files other than the ones mentioned above. Therefore, make sure your code is self-contained, and that the lecturer can run it without any additional files.

In case you work in a group, only one submission is required, and the name of the `.zip` file should be composed of both student numbers (e.g., `a12345_a67890.zip`).

Deadlines

- **Submission of all material:** 15/jan/2024, 23:59.
- **Oral presentation:** To be scheduled with each student/group individually.

Grading Weights

- **Report:** 8 points.
- **Source Code:** 4 points.
- **Oral Presentation:** 8 points.

Penalties

- **Late Submission:** 1 point/day[1] (Project mark = 0 if more than 2 days).
- **No Report:** 20 points (final mark = 0).
- **Report Over 8 Pages:** 1 point deduction/page over 8.
- **No Source Code:** 20 points (final mark = 0).

Guidelines

- **Report:** Max 8 pages, cite sources, justify decisions, use clear visuals.
- **Presentation:** Max 20 minutes, emphasize results and conclusions.

Academic Integrity

If the lecturer suspects a violation of the Academic Integrity Policy by a student, the case will be reported to the Academic Integrity Committee. This suspicion may arise either during the review of the written report or the oral presentation. In addition, engaging in any of the following actions will result in a grade of 0 for the project:

- **Plagiarism:** Copying from sources without proper citation or utilizing an artificial intelligence system to generate the report or source code.
- **Collusion:** Copying from other students.
- **Contract Cheating:** Have someone to write the report or source code for you.
- **Fabrication:** Falsifying data or results.

Groups

You have the choice of working independently or in pairs. If you decide to collaborate, it's crucial to inform the lecturer via email before November 30, 2023. After this deadline, no changes will be allowed, and individual work will be required. In your email, include the names and student numbers of all group members.

In a group, equal collaboration is expected among all members for all tasks. Since each task comprises two subtasks (i.e., 2 models), you can distribute the workload as follows: one member is responsible for the first subtask, while the other handles the second subtask. However, alternative divisions of work (e.g., one member focusing on the report while the other tackles the code and/or presentation) are not permitted. It's crucial to emphasize that **grades are awarded to the entire group**, meaning that if one member performs poorly, the other member cannot compensate for that. Therefore, make thoughtful decisions when selecting your group members.

---

[1]1 minute later, or 23 hours and 59 minutes later, is considered 1 day.

If a group member decides to drop out after collaboration has begun, the remaining member must proceed with individual work. In this case, the remaining member must notify the lecturer before the submission (and the report will need to be updated accordingly). Failure to notify the lecturer will result in the submission being considered a group submission, and the remaining member will receive a grade of 0.

Questions?

For questions about the project, please use the Q&A forum on Moodle.

References

[1] P. Cortez, Antonio Luís Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. *Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47*, 547-553. 2009.