Degree in Applied Data Science
Machine Learning
2024/2025

Practical Assignment
Data Analysis using Machine Learning

This assignment proposes an in-depth investigation into the field of data exploration and analysis using machine learning techniques. The main goal is to enable students to apply appropriate methodologies to explore and understand datasets, perform preprocessing, apply unsupervised analysis techniques, and evaluate the performance of various supervised machine learning algorithms.

To achieve this, a Jupyter Notebook should be created, structured into sections that encompass the steps of the analysis and provide a concise explanation of the procedures performed and decisions made throughout the analysis. Additionally, groups will present their work at the end of the semester.

Students will address one of the following types of problems:

- Binary classification
- Multiclass classification
- Regression

The dataset selection is the responsibility of the group. Some suggestions are provided at the end of this document.

The work groups should consist of 3 members.

In general, the following steps should be completed throughout the assignment:

1. **Initial Data Exploration and Preprocessing:**
    a. Review of the available documentation about the dataset.
    b. Exploratory analysis of the dataset.
    c. Data preprocessing, including handling missing values and possibly generating and selecting features.
    d. Description of the dataset's characteristics and justification for preprocessing choices.
    e. Inclusion of initial exploratory charts illustrating the dataset's main characteristics.

2. **Unsupervised analysis:**
    a. Use of dimensionality reduction techniques.

b. Analysis of the results obtained from dimensionality reduction techniques and data visualization.
c. Use of clustering methods.
d. Analysis of the results obtained from clustering algorithms.

3. **Supervised Machine Learning:**
    a. Comparison of the performance of various machine learning models/algorithms.
    b. Use of ensemble methods.
    c. Calculation of error metrics and application of appropriate error estimation methods.
    d. Hyperparameter optimization.
    e. Selection of the best model achieved and interpretation of the results where possible.
    f. Critical analysis of the results obtained in this stage.

Important Dates:

- Final selection of datasets and work groups:
    - 28th of February 2025
    - By this date, the chosen datasets must be discussed with the professor for approval.
- Presentation: 16th of May 2025
- Final Submission: 30th of May 2025

Evaluation:

The assignment will be graded based on the following criteria:

- Presentation: 30%
- Final Notebook: 70%

Note: Individual contributions will be considered, and grades may vary within a group if justified.

Some resources for choosing datasets:

- UCI Machine Learning Repository: https://archive.ics.uci.edu/datasets
- Kaggle datasets: https://www.kaggle.com/datasets
- Awesome public datasets repository: https://github.com/awesomedata/awesome-public-datasets/tree/master