Project Directives

Machine Learning 2024 Project Guidelines

The project grade accounts for 70% of the final grade: 30% for the presentation and 40% for the report. The remaining 30% are for the exam. Grades are from 1 to 6 (Absent=0), rounded to the closest 0.1. The final grade (40% report + 30% presentation + 30% exam) is rounded to the closest 0.5.

Organization of Groups

The size of groups is 3 students. You may now register via our moodle page. Because we do not know the final number of participants yet, we may create groups of 4 by the end of the grouping process. By default, only groups of 3 can register for the moment.

Evaluation Criteria

Report

Aspect	Choice of techniques	Implementation	Interpretation	Reporting quality (texts, graphs, tables)	Weight
Data preprocessing: cleaning, scaling, feature engineering, missing data, exploration, representation	2	3	2	3	10

Aspect	Choice of techniques	Implementation	Interpretation	Reporting quality (texts, graphs, tables)	Weight
Model selection and implementation: model choices, splitting methods, metrics, variable selection	7	7	6	5	25
Tuning: hyperparameters, solving overfitting, balancing	6	5	3	1	15
Interpretability: methods and application	7	3	7	3	20
Unsupervised method: clustering, dimension reduction	3	2	4	1	10
Global appreciation: clarity, structure, concept integration, appropriate vocabulary, originality	-	-	-	-	20
Overall weight	-	-	-	-	100

Presentation

Aspect	Weight
Delivery: quality of slides, completeness	25
Clarity: organization, logic, engagement, voice.	15
Time management	15
Q&A session: understanding, correct answer, clarity (no ambiguity)	25
Overall impression	20
Overall weight	100

> **Note**
>
> - Grade (Absent=0, min=1, max=6) weighted average
> - Each graded 1 (absent/non satisfactory) up to 5 (completely); eventually 5.5 (outstanding)

Originality And Style

The project consists of an original analysis of one data set. We encourage the data set to be extracted from the sources provided on the webpage of the course or alternatively use web scraping to gather novel data. If you decide to use another one, please ask us a validation. In any case, the data source must be given.

If the data comes from a previous analysis, then your analysis must have new elements (new models, features, outcome labels, approaches, interpretations, etc.). In such case, this analysis must be explained, and the original aspects of the new analysis must be made clear, for example, being transparent on the important differences between your and the previous analysis.

Completeness

Your analysis must be comprehensive and not leave major aspects of your dataset unaddressed. The project must include at least:

- Cleaning of the data, creation of new features if needed.
- Exploratory Data Analysis.
- Supervised learning analysis: two or more models, two or more meaningful metrics, a tuning of one or more hyperparameters per model, a data splitting (if a training/test set split is enough for the global analysis, at least one CV or bootstrap must be used), addressing of the overfitting problem and unbalance, if applicable.
- Interpretability methods (variable importance and PDP).
- Unsupervised analysis: clustering and/or dimension reduction.

Deliverables

- Report: a full PDF or HTML report of your analysis with reproducible code and figures, not exceeding 30 pages (or equivalent for HTML), including title page and appendices.

- Presentation: the slides of your presentation.

Deadlines

You must meet the following deadlines (by 23h59):

- Project report: Sunday the 19$^{\text{th}}$ of May 2024 at 23h59
- Presentation slide: Thursday the 23$^{\text{rd}}$ of May 2024 at 23h59

Not meeting the deadline (presentation and/or report) penalizes the grade by 0.1 per started hour of delay. No maximum penalty for the project report and presentation slide.

Presentations

Presentations will be organized on site on Monday the 27$^{\text{th}}$ of May 2024 in a 15+5-minutes format (presentation + questions). If the number of groups is large, then either the presentation time will be reduced, or, if not possible, another oral presentation session will be organized later (possibly during the exam session).