TÉCNICO+

DASH: Unsupervised Analysis (2026)

Project

Goal: principled application of data science techniques to acquire novel and relevant knowledge from two distinct real-world problems.

In this context, participants are asked to *explore* 2 datasets (problems) and, in accordance with their profiling and goals, adequately select and *learn* descriptive models from the available data, and further *evaluate* the descriptors. Students should be able to critically assess the acquired results, identify limits for knowledge acquisition and, accordingly, hypothesize causes and identify opportunities to improve the descriptive modeling process.

Guidelines

Projects to be developed in groups of 2 or, upon agreement with your faculty host, 3 trainees.

Datasets: Two data sources need to be selected for this project. A list of possible datasets from different domains is provided in the course webpage. The groups can autonomously propose alternative data sources. The datasets can follow any data structure – tensor, temporal, event, spatiotemporal, and relational data structures are welcomed.

Programming language: We recommend the use of *Python* for the course's project. Nevertheless, students are allowed to use other languages.

Objectives: The unsupervised exploration of the datasets must be done along three stages: i) *data profiling/representation*, ii) *clustering*, and iii) *pattern analysis* or *outlier analysis*.

Tasks

Regarding **data profiling** and **representation**:

- the groups should perform a statistical analysis of the datasets in advance, and summarize relevant implications in the report, such as the underlying distributions, relevant statistics and hypothesized forms feature dependency;
- adequate data representations should be pursued for projects with non-tabular data structures and unstructured variables using domain-guided transformations or representation learning (*embeddings*);
- the underlying complexity of the datasets can be assessed via summarization capacity using linear methods, such as PCA, and non-linear methods, such as autoenconders;

- in accordance with the profiled characteristics, and the targeted problems, the trainees can further apply data processing/transformation techniques to aid the learning (e.g. feature engineering, dimensionality reduction, imputation, normalization, discretization).

Regarding **clustering analysis**:

- the adequacy of the distance functions, methods and number of clusters should be experimentally discussed. In particular, the trainees should explore the results from two different clustering algorithms (e.g. hierarchical, model-based, partitioning);
- a careful intrinsic evaluation should be pursued. In the presence of target variables, extrinsic evaluation should be further undertaken against each of these variables;
- trainees can opt to either include or exclude variables of interest, justifying the underlying reasoning in accordance with the pursued descriptive goals;
- the trainees should pursue a visualization of the most promising clustering solutions by projecting the data into a two-dimensional representation by either selecting the most informative/discriminative features or extracting the top principal components. The median and medoid center of the found clusters can be further recovered for descriptive purposes.

Regarding **pattern analysis**:

- trainees can opt to pursue association rule mining (ARM) and/or biclustering for pattern analysis;
- handling numeric variables in ARM, imbalanced categories, homogeneity criteria in biclustering searches, and pattern redundancies may be relevant;
- evaluation statistics, including the statistical significance of the patterns or the lift of association rules should be computed, along with domain relevance;
- in the presence of variables of interest, the discriminative power of patterns and biclusters can be assessed for each outcome of interest. The ARM search can be further adapted to discover association rules with classes in the antecedent/consequent.

Regarding **outlier analysis**:

- trainees should perform (unsupervised) multivariate outlier analysis, paying attention to the selected distance, density or statistical threshold criteria to flag anomalies;
- feature relevance using domain knowledge or, in alternative, redundancies using correlation analysis should be taken into consideration to guide the detection;
- the trainees should further calculate and analyze outlier scores for knowledge acquisition or, in alternative and if applicable, compare the applied unsupervised stance with a supervised stance using an interpretable classifier or regressor.

Delivery

Report: the report should follow the given template (Latex or Word) and is limited to 10 pages (6 pages suggested). The report can be written in Portuguese or English, and submitted in PDF format. The report should clearly describe the placed choices, the applied

parameterizations, the major results, and their critical discussion. Additionally, it should include a comparison of the results achieved in both problems, considering the differences between the selected data sources.

Please note that the evaluation of the project is primarily guided by the contents available in the report, so do not underestimate the necessary time to write a good report.

Functional project: Trainees should further provide the underlying programming code. A reference Jupyter notebook for projects in Python or R is available in the course webpage. Provide a .ZIP with all files (without the data sources), ensuring its standalone functionality.

Excellence

A project that applies the suggested data mining techniques over the given datasets and provides a sound analysis of the collected results is not necessarily an excelling project. Excelling projects have four major characteristics:

1. show an acute understanding of the data characteristics and their impact on the learning. Excelling projects formulate data-driven hypotheses behind the observed behavioral differences;
2. the report is written in clear, structured, and succinct language;
3. show creative thinking on ways of improving the learning, e.g. how the principled use of alternative data processing principles could aid the learning ;
4. robust assessments go beyond simple performance indicators. Excelling projects draw (parameter-varying) plots, inquiry the findings, and test hypotheses.

Evaluation Criteria

The project will be evaluated as a whole. Nevertheless, we provide below a decomposition of the total project score for the purpose of guidance and prioritization. Reference scores:

1. Data profiling/representation/transformation (20%)
2. Clustering analysis (45%)
3. Pattern/outlier analysis (35%)

The evaluation of the clustering and pattern/outlier analysis components is further divided into the following major criteria:

i. Methodology (40%) – soundness and adequacy
ii. Assessment (25%) – robustness and completeness
iii. Visualization (10%) – adequate presentation of clustering and pattern solutions
iv. Domain findings and implications (25%) – critical analysis

END