Statistical Modelling
Degree on Artificial Intelligence

Project Deliverables

Lecturers: Jordi Cortés (jordi.cortes-martinez@upc.edu)
Dante Conti (dante.conti@upc.edu)

Schedule course 2024-25				
ID	**Task**	**Date to deliver**	**Deliverables**	**Via**
D1	Team definition	12/9/2025	email to lab teacher	e-mail
D2	Data document	19/9/2025	Zip folder with: 1) Pdf data document (2 pages); and 2) Metadata Excel file	Atenea
D3	Work plan	26/9/2025	Pdf document (2 pages)	Atenea
D4	Pre-project	29/10/2025	Zip folder containing: 1) pdf report; 2) raw datasets, 3) preprocessed datasets; and 4) R scripts	Atenea
		31/10/2025	Pdf presentation	
D5	Final project	16/12/2025	Zip folder containing: 1) pdf report; 2) raw datasets, 3) preprocessed datasets; and 4) R scripts	Atenea
		17/12/2025	Pdf presentation	

D1. Team definition

1. The groups should be formed by **4 persons (group 12) or 5 persons (group 11)**.
2. As far as possible there should be **different genders** in the teams.
3. Send an **email to both lecturers** providing the name of all group components.
    a. If you do not have a group of 4 or 5 people, send the email mentioning the number of people you are or simply mentioning that you don't have any group. In any case, please, indicate your names.
4. The lecturers will provide the **composition of the final groups and an identifier of the group** a few days after receiving the emails.

D2. Data document. Every group must present a two-page report with the following information:

1. Data source including the url or urls involved
2. One paragraph explaining the process to get your data (basic download, more sophisticated processes when used).
3. One paragraph explaining what data are about
4. Basic structure of data matrix (half page): One table with:
    a. nr of records (better if it is bigger than 500, if you are working with countries in the world or other situations, this might be reconsidered)
    b. nr of variables
        i. nr of numerical variables (minimum of 5 numerical variables)
        ii. nr of binary variables (minimum of 2 binary variables)
        iii. nr of qualitative variables (minimum of 5 categorical variables)
        iv. nr of variables referring to dates (minimum 1 variable to apply time series analysis). Remark: if it is not possible to have a single variable related to dates, find another dataset with a related topic to perform a time series analysis.
    c. number and % of missing data per each variable
    d. % of missing data in the whole data matrix.

    Remark: In some cases, you will need an additional dataset for the **time series** part: The data should have: i) At least 100 observations; ii) Some correlation between variables."
5. Pre-processing (half page). Describe the methods to do perform the pre-processing of data:
    a. Missing data. How to deal with missing data (remove, impute,…)?
    b. Outliers. How to deal with outliers (remove, impute,…)?
    c. Errors. How to detect errors in the variables (feasible range, check categories…)?
    d. Group categorical data. When to group categories (e.g, >5 categories?)?
    e. Transform data. When to transform data?

In addition an Excel file with metadata information should be presented (name of variables, description, type, range,…)

Statistical Modelling
Degree on Artificial Intelligence

D3. Work plan. A two-page document explaining the organization of the team for performing the tasks related to the project (see the document *WorkPlan.pdf*). There should be a **project manager** in each team who will be in charge of communicating the progress of the project to the professor. It should include:

1. Gantt diagram. Include the tasks of the project
2. Assignment of tasks. Assign the tasks between the members of the team.
3. Brief risk contingency plan. Anticipate possible contingencies during the project.

D4. Pre-project. Provide a pdf presentation and a zip folder containing: 1) pdf report; 2) raw datasets, 3) preprocessed datasets; and 4) R scripts. The structure of the report should be:

1. Introduction.
    a. Problem
    b. Data (from D2)
    c. Work Plan (from D3)
    d. Objectives

2. Initial Statistical analysis
    a. Preprocessing.
        i. **Steps** of the preprocessing process are used with your particular data:
        ii. List and justify all **decisions** taken for each preprocessing step
    b. Descriptive statistics
        i. **Basic initial univariate and bivariate descriptive statistics** of preprocessed variables. Compare raw and preprocessed variable distribution when relevant
        ii. Half page describing the dataset according to the **main conclusions** of the univariate and bivariate statistics

3. Generalized Linear Models (GLMz)
    a. Model fitting
        i. Identify a **numerical response** (discrete or continuous) and fit a GLMz. Justify the quality of the model
        ii. Identify a **binary response** variable and fit a valid binary-response regression model. Justify the quality of the model
    b. Model validation. Perform the validation of the two fitted models.

4. Time series (TS)
    a. Find a chronological variable in your dataset and build a significant and valid time series model. If your dataset does not contain chronological data, find a chronological variable linked to your main dataset (i.e. if you work with a salaries database, find the series of average salaries per month or year for building the time series).

The structure of the pdf presentation should be similar but shorter.

D5. Final project delivery. Provide a pdf presentation and a zip folder containing: 1) pdf report; 2) raw datasets, 3) preprocessed datasets; and 4) R scripts. The final report should contain all the information already included in D4 and the following information:

5. PCA analysis for numerical variables:
    a. *Screeplot*. Specify how many principal components are selected
    b. Factorial map visualization.
    c. Interpretation

6. Clustering
    a. Hierarchical Clustering on original data:
        i. Precise description of the data used (which variables have been included in the analysis)
        ii. Clustering method used: metrics and aggregation criteria used
        iii. Dendrogram.
        iv. Discuss about how to get the final number of clusters
        v. Table with a description of the clusters size
    b. Partitional Clustering
        vi. Precise description of the data used (which variables have been included in the analysis)

vii. Clustering method used: K-means, K-medoids,…
viii. Visual representation
ix. Table with a description of the clusters size

7. Profiling of clusters.
    a. Use class variable as a response variable to analyze conditional distributions of variables to clusters and eventual statistical tests to assess which variables are significant in each cluster.
    b. Detect commonalities of each cluster and differences between clusters. What is intrinsic of each cluster? What distinguishes clusters among them?
    c. Profiling graphs, multiple boxplots, bivariate barplots, descriptive by groups, etc...
    d. For selected relevant variables, you can also add specific profiling tests to complete clusters interpretation
    e. Synthesize the result of the classes' interpretation process into a set of templates characterizing the clusters, one template per cluster
8. Discussion and conclusions
9. Planned Gantt and task distribution grid. Final real executed Gantt, tasks assignment grid and real risks addressed along the project with explanation of the changes regarding the initial plan

The structure of the pdf presentation should be similar but shorter.

Failure to comply with these rules may be penalized.

If someone cannot attend to the presentation days, please contact the lecturer by mail in advance to agree on a solution (send an email to jordi.cortes-martinez@upc.edu and dante.conti@upc.edu). **Non-notified absences to presentation days will be penalized.**