Final Projects Statistical Methods 2025/2026

Nicola Torelli, Gioia Di Credico

General Instructions

The final projects must be discussed during the winter session 2025/2026.

The presentation should include the following:

- a description of the project's aim
- an exploratory data analysis, including a possible data-cleaning phase
- selection, description, and possibly comparison of the most suitable statistical models
- comments on results.

For some data files, you will likely find some analyses on the net carried out by others. You can look at them for inspiration, but we ask you to find your original key to the analysis.

Each presentation must last 20/25 minutes. All group members must be aware of all the project's parts and take part in the presentation. **The coherence of results across the project sections will be evaluated**.

You are free to choose the kind of programming language and file extension to organize your presentation, e.g., slides, report. You have to upload on Moodle the script project and the report that will be presented the day before the exam.

Students that did not participate in homework and/or partial tests, that have to pass an oral examination, have to send an email to professors to have assigned a final project.

Operational Instructions

1 Free aggregation into groups, with a maximum of 4 members per group.
2 Select the project on moodle: each dataset has its description and a link. Before starting the analysis, carefully read the description of the dataset and variables. All group members must select the same project. You are free to choose another dataset of your liking, subject to consultation with the lecturers.
3 Register on esse3 for the desired session: all group members must register for the same session. Those registered for the session are warmly invited to listen to the presentations of the other groups registered for the same session.
4 Submit the presentation on Moodle before the session.

Do not hesitate to contact us for any issue.

List of Selected Projects

A - 'Airlines Performances Data' Dataset

- **Source:** https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7
- **Description:** Includes arrival and departure details for all commercial flights in the USA, from October 1987 to April 2008. Variables include departure and arrival delays, date/time information, origin and destination airports, and so on.
- **Objective:** Build a regression model to investigate relationships between delays and other variables and to predict flight delay times.

B - 'Bank Marketing' Dataset

- **Source:** https://www.kaggle.com/datasets/pkdarabi/bank-marketing-dataset
- **Description:** Analyze patterns from the previous marketing campaign for a financial institution to improve the effectiveness of future campaigns.
- **Objective:** Use the **y** variable (positive/negative customer response to the campaign) as the response variable. Predict the outcome and understand which variables best explain the probability of responding positively to the campaign.

C - 'Bike-Sharing' Dataset

- **Source:** Moodle
- **Description:** Two-year historical data (2011 and 2012) related to the bike-sharing rental process, which is highly correlated with environmental and seasonal settings (weather conditions, precipitation, day of week, season, hour of the day, etc.).
- **Objective:** Build regression models to predict the **number of bike rentals** based on environmental and temporal factors.

D - 'Cardiovascular Disease' Dataset

- **Source:** https://www.kaggle.com/datasets/akshatshaw7/cardiovascular-disease-dataset
- **Description:** Dataset on the presence or absence of cardiovascular disease.
- **Objective:** Build a classification model to predict the **presence/absence of cardiovascular disease** (binary variable) and evaluate the importance of variables (e.g., blood pressure, cholesterol, smoking).

E - 'Churn' Dataset

- **Source:** Moodle
- **Description:** Data related to customer churn in a telephone company.
- **Objective:** Predict the probability that a customer will churn (**churn**, binary variable), identifying the most significant risk factors.

F - 'Flight' Dataset

- **Source:** https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction
- **Description:** The price of an airline ticket is influenced by several factors, such as flight duration, days remaining until departure, arrival and departure time, etc.
- **Objective:** Build regression models for the **price** variable and explain which variables most influence the ticket price.

G - 'Health Care Australia' Dataset

- **Source:** Moodle
- **Description:** Contains data related to the use of healthcare services in Australia.
- **Objective:** Model and predict the **number of doctor visits**) using appropriate regression models.

H - 'Malnutrition in Zambia' Dataset

- **Source:** Moodle.
- **Description:** Data on the nutritional condition of children aged 0 to 5, measured by the standardized measure **Z-score**.
- **Objective:** Model and predict the **Z-score** as a function of available demographic, socioeconomic, or health variables.

I - 'Pollution' Dataset

- **Source:** Moodle
- **Description:** Dataset used to model the impact of pollution-related variables (e.g., atmospheric pollutant concentrations) on mortality.
- **Objective:** Build a regression model to quantify the association between pollution levels and mortality.

J - 'Rent Modelling in Munich' Dataset

- **Source:** Moodle and also in `gamlss.data` package (`rent99`).
- **Description:** Data on rents in a Munich area in 1999.
- **Objective:** Build a regression model to predict the **rent price** based on property characteristics (size, location, year of construction, etc.).

K - 'Student Performances' Dataset

- **Source:** Moodle
- **Description:** Data for predicting secondary school student performance. The dataset includes demographic, social, and school-related variables. Performance is measured by final grades.
- **Objective:** The final math grade was collected by using school reports and questionnaires. The data attributes include student grades, demographic, social and school related features). Attempt both the regression approach (to predict the exact grade) and the classification approach (to predict success).

L - 'Swiss Labor' Dataset

- **Source:** Moodle and also in `AER` package (`SwissLabor`).
- **Description:** Data related to women's labor force participation in Switzerland.
- **Objective:** Predict whether a woman **participates in the labor force** (binary variable) based on socioeconomic and demographic covariates.

M - 'Travel_ticket_cancellation' Dataset

- **Source:** https://www.kaggle.com/datasets/pkdarabi/classification-of-travel-purpose
- **Description:** Every cancellation results in a fine for the ticket registration website by the airline. It is crucial to identify tickets likely to be canceled to manage cancellation risk effectively.
- **Objective:** Develop a model to predict if users will cancel their tickets. The response variable is **Cancel** (0 if not canceled, 1 if canceled).

N - 'Videogames' Dataset

- **Source:** https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings
- **Description:** Reports sales as of December 2016 for major video games on the market. Each row expresses the characteristics of a given video game.
- **Objective:** Use the **Global_Sales** variable as the response variable and use statistical models to predict and model this quantity as a function of the available variables.

O - Proposed Dataset

P - Proposed Dataset