DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

Project 2

Statistical Methods for Artificial Intelligence

(MEIA, 1st Semester, 2nd Quarter, 2025/2026)

Handed out on 12 December 2025.

To be handed back on 9 January 2026.

Consider for **Soils** data set, available in R: **library(carData)**; **data("Soils")**, the subset with all observations for the last 9 variables: (**pH**, **N**, **Dens**, **P**, **Ca**, **Mg**, **K**, **Na**, **Conduc**).

1. Explore/describe this data applying methods taught in this course, in particular using plots and summary statistics (e.g. mean, mad, variance, covariance, generalized/total variance and Mahalanobis distances) and discuss what you have learned from this preliminary analysis.
2. One researcher has rudimentary knowledge about multiple linear regression analysis and wants your help to find a way to explain the variable **pH** with some predictors variables.
   (a) Fit a regression model to this data set. Test for significance of the regression. Discuss the results in terms of the p-value. Compare the test result with the coefficient of multiple determination. Is there any evidence that a subset of the standardized variables should be excluded from the model? Proceed in order to find the best subset of regressors.
   (b) Calculate 97.5% confidence interval (CI) on the mean responses for observation **12** and observation **24**. For the same values of the regressors, and the same confidence level, calculate the prediction interval (PI). Compare and discuss the obtained results.

About the report: The report should not exceed 10 pages (with Annexes). Do not forget to include: introduction, objectives of study, decisions, conclusions and bibliography. The R code and the report must be upload in fenix.