DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA
M. Rosário Oliveira

Project – Multivariate Analysis

(MECD, Minor-CD, & MMAC, 1^{st} Semester, 2023/2024)

Handed out on September 26, 2023.

To be handed back by **December 18** , 2023.

Presentations of Data Preliminary Analysis on **November 17**, 2023.

Presentations and discussion on **January 3**, 2024.

1. Find a labelled dataset that you have an interest in its analysis.
2. Make a preliminary analysis of the data and discuss what you have learned from this analysis.
3. Solve your classification problem using supervised learning methods. Have in mind that some of the input variables may be irrelevant to the classification problem and that you may need to do some preprocessing methodologies of your data set e.g. dimensionality reduction techniques.
4. Apply unsupervised methods to your data, ignoring the class variable. Interpret the results. Compare the obtained partitions with the true class each object belongs to.
5. Repeat the classification study using for classes the partition clusters obtained in (4). Compare the results with the ones obtained in 3. Discuss the potential advantages and drawbacks of each strategy. Which would you recommend to analyze your dataset?

   Include in your discussion all options that you have made, the advantages and disadvantages of each alternative.
6. Imagine you are going to meet the researcher who contacted you. Report to him/her what you have learned about the problem. Discuss the limitations of the analysis you have done and provide suggestions for future work.

- **About the datasets:**
  - Students should find an interesting dataset to be analyzed.
  - A bonus of almost one point is given for the originality of the chosen dataset.
  - Datasets libraries examples:
    * Kaggle
    * Pordata
    * INE
    * Eurostat
    * UCI Machine Learning Repository
    * re3data
    * DataCite Metadata Search
    * UK Data Archive
    * KEEL
    * Google Dataset Search
    * Machine Learning and AI, Carnegie Mellon University Libraries
    * List of datasets for machine-learning research, Wikipedia
    * Elite Data Science
    * BioDiversity4All
    * iNaturalist
    * Journal of the Royal Statistical Society: Series A (Statistics in Society) Datasets
    * Journal of the Royal Statistical Society: Series B (Statistical Methodology) Datasets
    * Journal of the Royal Statistical Society: Series C (Applied Statistics) Datasets
  - Report the source of the chosen dataset.
  - A bonus of at most 1.0 points (out of 20) are assigned for the dataset originality.
- **About the groups:**
  - Students should organize themselves in groups of 5 persons, and fill their groups in the following Google form: https://forms.gle/2Wd726SFEy7TDYt78 until **6 October 2023**.
- **About the presentation:**
  - Duration of 10 minutes plus 5 minutes for discussion.
  - Slides and commented code must be handed back with the report, on the due time.
- **About the report:**
  - The report should not exceed 10 pages, in the form of a scientific paper.
- Do not forget topics such as:
  1. Description of the problem under study;
  2. Objectives;
  3. Estimation and validation methods;
  4. Discussion of the results and interpretation of the findings;
  5. Conclusions;
  6. References.
- The **commented R code**, the **presentation** as well as the **report** must be uploaded to the Fenix webpage.