# Block 2 laboratory deliverable guidelines

**Advanced Machine Learning, Master in Data Science, Fall 2023, UPC**

---

This document contains the guidelines for the block 2 deliverable. Please read with care!

---

**General information**

This project is meant to give you the opportunity to apply the techniques seen during this block to a real-world dataset. The project should cover all aspects of the modeling methodology seen in previous courses: from preprocessing to generating a final predictive model together with an assessment of its prediction quality and how it could be improved.

It is expected from you to:

- **The project is to be done in teams of two persons**; **singles are not allowed**.
- Select a dataset to work on. You can choose one from the data repositories we are going to provide to you (see below), or propose your own. Take into account that the learning algorithms presented in this course can be used for problems more complex than those that are common in the Machine Learning literature.
- In any case, you are expected to hand in a written report:
  - This document should describe the work carried out, its motivation, the problems encountered, the actions taken to solve the problems and the solutions found together with final results and conclusions of your study.

To carry out your analysis you can use Python, R or both. Remember that there are many useful packages that extend its basic functionality. If you use code or ideas or any kind of resource from elsewhere you should cite it appropriately. Plagiarism will be prosecuted.

**First task**

Submit a 1-page project proposal:

- It should include the composition of your team.
- The problem that you are going to work on, the reason why you chose it, together with any references of previous work on this problem if applicable.
- A title for your project.
- Include information on the data such as number of rows, columns, and nature of variables (e.g. if categorical, how many levels).

**Please hand this in no later than October 16th through the racó.**

**Report submit**

The final report should include:

1. A brief description of the work and its goals, data available, and any additional information that you may have used.
2. Related previous work (if applicable).
3. The data exploration process, including: pre-processing, feature selection/extraction, visualization, clustering, etc.
4. **Modeling**:
   a. A kernel base model ("kernelized" method).
   b. A baseline model that must be the not kernelized version of the method used in a).

    c. (Optional) One or more of any other learning method, for example: A decision tree, a random forest, etc.

    *Justify all the methods used.

5. Results obtained with each method used (along with the best set of parameters) and comparison of results.
6. Final model chosen and an estimation of its generalization performance.
7. Scientific and personal conclusions
8. Possible extensions and known limitations.

Note that the report should not describe explanations seen in class; every table or plot should be appropriately described. The style of the report should resemble what you encounter in a scientific publication.
Your code should be reproducible; that means using “seeds” if your code is stochastic.
For the final delivery, make sure you include in a compressed file the following:

1. The written report (pdf document). **It should not exceed 14 pages**; if you need more space, consider placing the secondary information in a separate appendix file.
2. Any script or code you have used (R/python notebooks, scripts, or any other code)
3. A flat text file with precise instructions on how to execute and reproduce your results.

Code and report should be submitted separately (namely, a python notebook is not a report).

**Evaluation**
The following are the conditions that will be taken into account in the grading of this part of the laboratory:

1. The proper use of techniques and methods presented in class.
2. The care and rigor for obtaining the results (resampling protocol, quality metrics, statistical significance).
3. The quality of the obtained results (generalization error, simplicity, interpretability).
4. The quality of the written report (conciseness, completeness, clarity).

**Key dates**

- **18 October**: proposal.
- **12 November**: Report and code.

**Data repositories**
The following sites contain a number of very diverse datasets; many correspond to real-world problems.
They vary in domain (biology, medicine, economy, etc.), and also in size, type of variables, type of problem (classification or regression), among other things. Please choose one that interests you and where the studied methods can work well in theory.

- Open ML [https://www.openml.org/search?type=data]
- UCI Repository [http://archive.ics.uci.edu/ml/index.php]
- UCI KDD Archive [http://kdd.ics.uci.edu/summary.data.application.html]
- Statlib [http://lib.stat.cmu.edu/datasets/]
- Delve [http://www.cs.utoronto.ca/~delve/data/datasets.html]
- School of Informatics (U. of Edinburgh) repository [http://www.inf.ed.ac.uk/teaching/courses/irds/miniproject-datasets.html]

You can also use a dataset from Kaggle (**datasets section, not from competitions**) but take into account the following:
The vast amount of data is very heterogeneous and so there could be times that the data can not be used for modeling or that the modeling techniques required are beyond this course (time series, NLP, images, etc.). Some problems are easier in some aspects and more difficult in others. Therefore, the selection of the particular problem does not have a lot of importance for the grade. In particular, it is not at all advisable that you start to test problems to see how they behave. It is recommended that you make the decision by the interest that it raises in you.
The main warning we give is to be aware of the large computational needs that some (or many) of the chosen methods may have. If you choose a problem with a number of rows in the order of 100000 or more, you will surely get into serious demands of CPU/GPU and RAM ... Unless you do something to reduce these needs, it is wise to keep this number in between 1000 to 100000 rows.

**Others**
Only one member of each team should submit information (always via the Racó)