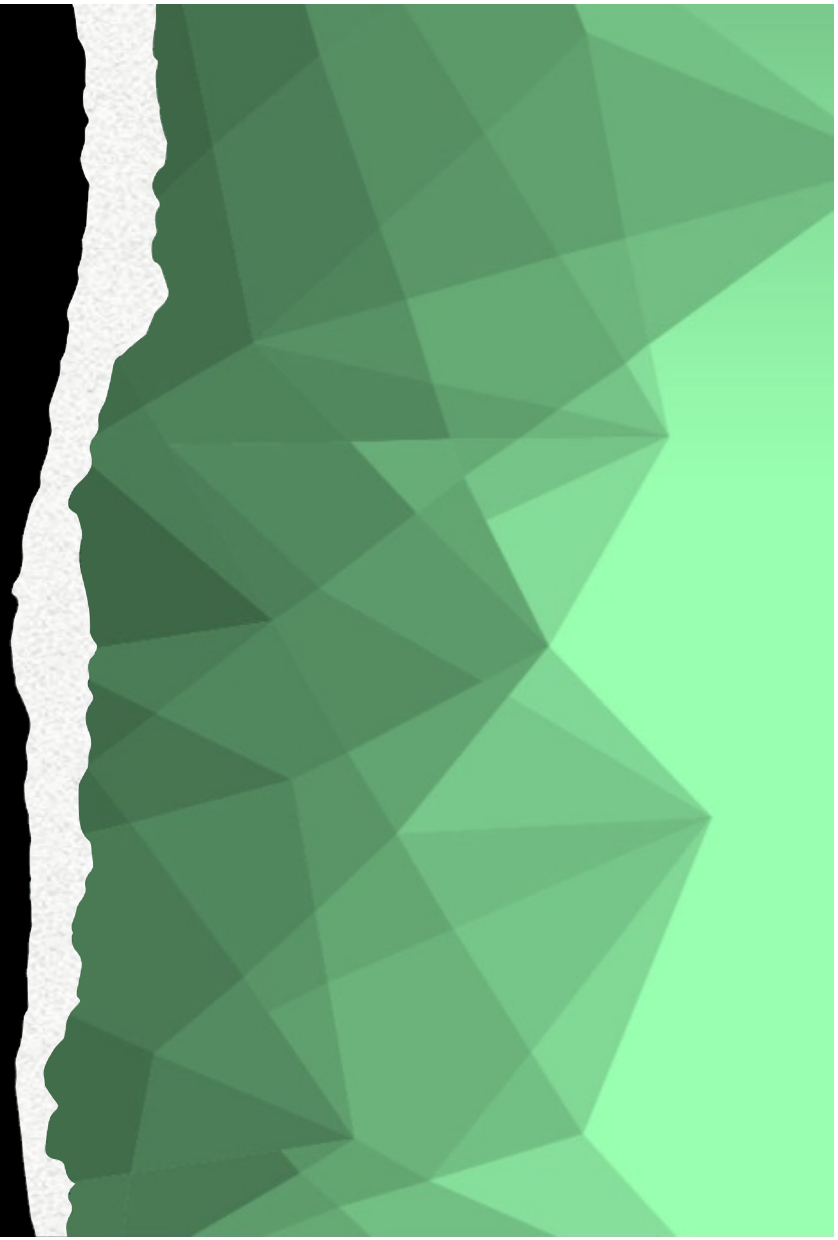
Statistical learning for healthcare data

Indications for the project

Prof. Manuela Ferrario



Project 1 – Suspected Pneumonia database

Database description.
585 patients: 190 had COVID-19, 50 had pneumonia secondary to other respiratory viruses, 252 had other pneumonia (bacterial), and 93 were initially suspected of having pneumonia yet subsequently adjudicated as having respiratory failure unrelated to pneumonia (non-pneumonia controls). 45% of the cohort had an unfavorable outcome defined as discharge to hospice or death.

There are 136 episodes of **community-acquired pneumonia (CAP)**, 214 episodes of **hospital-acquired pneumonia (HAP)**, and 328 episodes of **ventilator-associated pneumonia (VAP)**. 317 episodes were adjudicated to be successfully cured, 131 were adjudicated as indeterminate, and 230 were adjudicated as not cured.

Each patient-ICU-day is presented in a single row, along with admission summary information.
The first 20 columns present demographic and outcome data summarized across the patient's stay, the next 47 columns contain day-by-day values, and the last 5 columns present clinical pneumonia episode adjudication data.

Project 1 – Suspected Pneumonia database

Objectives
The main objective is to develop a model to predict patient mortality.
The team should clearly motivate possible inclusion/exclusion criteria of the patients, how to manage the different ICU length of stay among patients (e.g. predict mortality day by day), how to include the information about the bronchoalveolar lavage (BAL).
Which are the key parameters of the models?

Option
It would be optional to figure out a solution that foreseen a possible automatic update of the model parameters (patients are continuously collected in the hospital and the model could be periodically updated)

Description of the database:
https://physionet.org/content/script-carpediem-dataset/1.1.0/
https://www.jci.org/articles/view/170682

Evaluation criteria of the project

- **Clarity (45%)**: is the document understandable and easy to read (both Jupyter Notebook and report)? is the length appropriate? are all non-obvious design choices made explicit? is the solution/experimental campaign repeatable/reproducible based on the provided description?
- **technical soundness (45%)**: are the problem statement, evaluation criteria, evaluation procedure sound? are design choices motivated experimentally, with references, or by other means? are conclusions and findings actually supported by results?
- **Results (10%)**: does the solution effectively/efficiently solve the problem? is there a baseline which is improved in some way?

Note that the students' solution is not required to exhibit some degree of novelty (i.e., to advance the state of the art of the specific research field). However, student are expected not to simply "cut-and-paste" an existing (research) project.

Oral examination: to test the awareness of methods and techniques used to develop the project.

Check list

This checklist can guide you through your Machine Learning projects. Obviously, you should feel free to adapt this checklist to your needs.

1. Frame the problem and look at the big picture.
2. Get the data.
3. Explore the data to gain insights. correlation, collinearity, ...
4. Prepare the data to better expose the underlying data patterns to Machine Learning algorithms.
5. Explore many different models and shortlist the best ones.
6. Fine-tune your models and combine them into a great solution.
7. Present your solution.

Look at the file ‘End-to-end project.ipynb’ as an example.