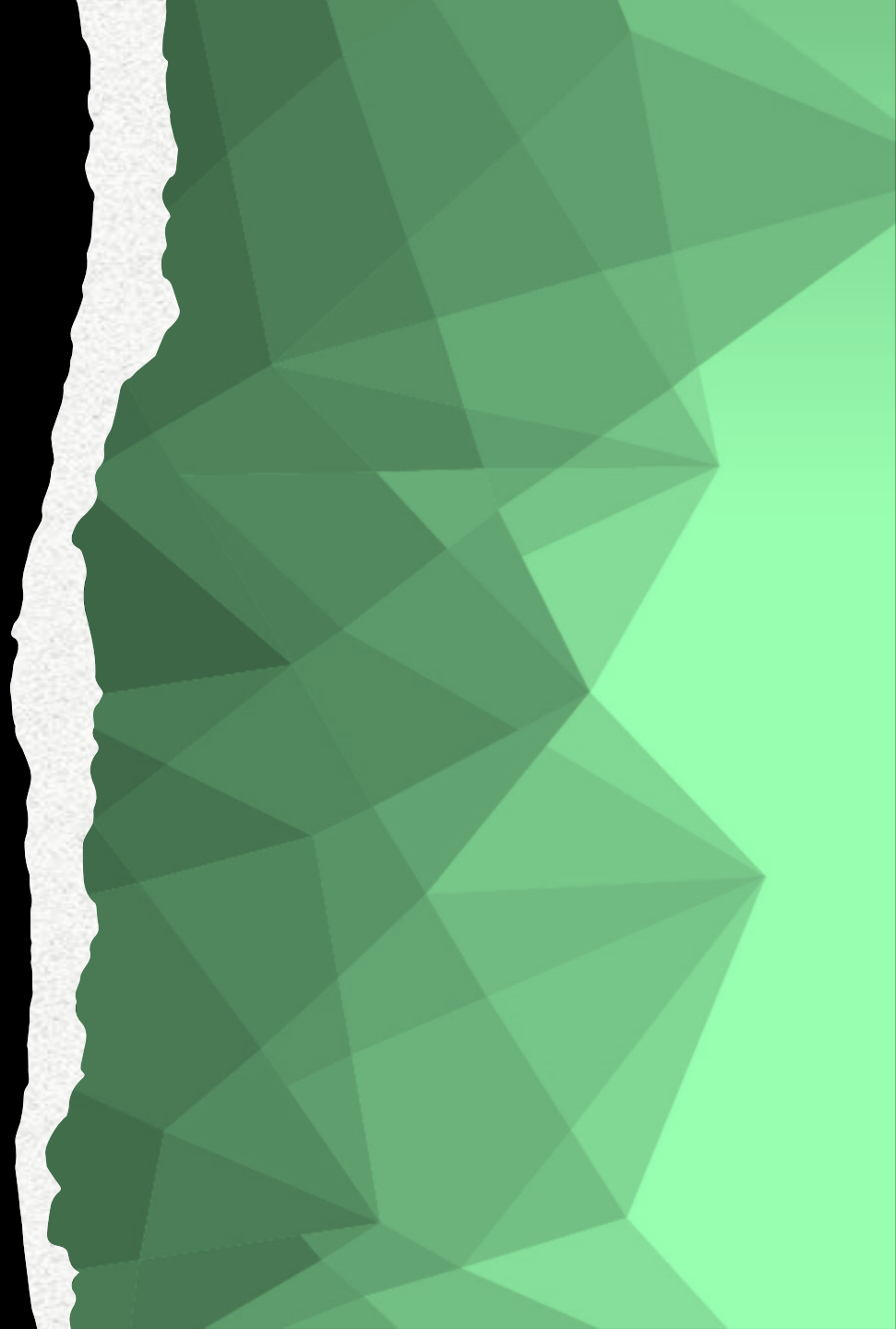
Statistical learning for health data

Indications for the project

Prof. Manuela Ferrario



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.
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.

Some suggestions to report your works

DOME: recommendations for supervised machine learning validation in biology

DOME is a set of community-wide recommendations for reporting supervised machine learning–based analyses applied to biological studies. Broad adoption of these recommendations will help improve machine learning assessment and reproducibility.

Box 1 | Structuring a Methods section for supervised machine learning approaches

Here we suggest a list of questions that authors should address in the Methods sections of manuscripts describing supervised ML approaches, in order to conform to the DOME recommendations and ensure a high quality of ML analysis.

Data (this section should be repeated separately for each dataset)

If yes, why was it chosen over better known alternatives?

- *Meta-predictions*: Does the model use data from other ML algorithms as input? If yes, which ones? Is it clear that training data of initial predictors and meta-predictor are independent of test data for the meta-predictor?
- *Data encoding*: How were the data encoded and preprocessed for the ML

- *Output*: Is the model classification or regression?
- *Execution time*: How much time does a single representative prediction require on a standard machine (for example, seconds on a desktop PC or high-performance computing cluster)?
- *Availability of software*: Is the source code released? Is a method to run the

Box 1 suggests you the key information should be reported in the methods description

https://www.nature.com/articles/s41592-021-01205-4

Project 1 – Heart Failure

Heart failure is a common reason for hospitalization in the elderly and it is associated with significant mortality and morbidity.
The data consist of a retrospective heart failure dataset created by using *electronic health data* collected from patients who were *admitted to a hospital* in Sichuan, China between 2016 and 2019. The dataset includes 168 variables for 2,008 patients with heart failure.

Objective: to predict the readmission at 6 months after hospital discharge by including also information about drug therapy. Are the drugs important features for the outcome? Any additional features can be added in the model?

Description of the database
https://www.physionet.org/content/heart-failure-zigong/1.2/

References about the topic
https://www.ahajournals.org/doi/epub/10.1161/CIRCHEARTFAILURE.121.008335
https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/414374

Project 2 – Myocardial Infarction

The proposed database can be used to solve two practically important problems: predicting complications of Myocardial Infarction (MI) based on information about the patient (i) at the time of admission and (ii) at the third day of the hospital period.

Objective1: to predict the complications of MI patients (no: alive, yes: lethal events)
Objective2: to predict if a MI patient will become a chronic heart failure patient but without lethal events.

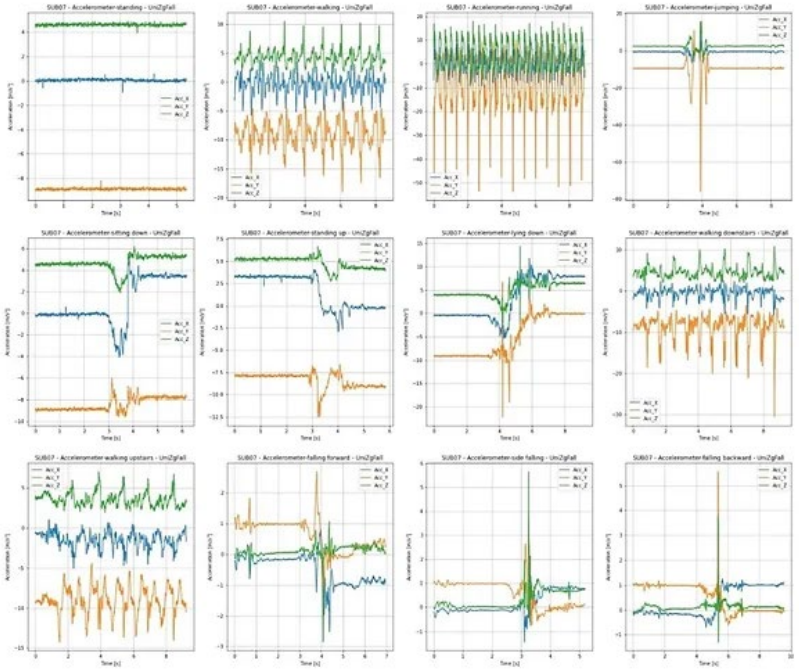
Description of the database
https://archive.ics.uci.edu/ml/datasets/Myocardial+infarction+complications

References about the topic:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7835562/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4229030/

Challenge 1

UniZgFall dataset

A dataset for accelerometer-based fall detection is used in this Scientific Challenge. For the purpose of data acquisition 16 young healthy subjects were recruited to perform 12 types of activities of daily living (ADL) and 3 types of simulated falls while wearing an inertial sensor unit (Shimmer sensing, Ireland) attached sideways to their waist at belt high. A 2 cm thick tatami mattress was used to cushion the impact during fall simulations.



Objective: to develop a classification system, able to classify each record/activity in one of distinct classes:

1) moving
2) falls
3) others

Evaluation criteria of the project

- **Clarity (45%)**: is the document understandable and easy to read? 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