FINAL PROJECT

Computational Machine Learning
Guillem Mirabent Rubinat
Prof. Juan Antonio Rodriguez Aguilar

I. Description:

For this project, you will need to create a Machine Learning pipeline (or two) and use it to predict the values that the target variable takes in a given test set within the data.

This project consists of INDIVIDUAL work. You are expected to upload your own notebook and predictions individually. You will also present individually.

The most important dates for this project are the following (specific times in 24h format):

- **Notebook upload deadline:** 11th of December 2025 @ 08:00
- **Predictions deadline:** 11th of December 2025 @ 18:00
- **IN-PERSON presentation (in front of the rest of the class):** 12th of December 2025 (starting) @ 10:00

II. The Data:

You are given data from the dataset “MIMIC III”, which contains anonymized data on ICU (Intensive Care Unit) patients. The given CSV files are actually a subset of the full “MIMIC III” dataset, which contain:

- **20.885 observations:** Each observation is an individual stay at the ICU.
- **44 variables:** Although some of those are not in the test dataset because they are giveaways of the target variable value, figure them out and drop them!

The variables are either:

- **ID columns:** There are 3 different IDs for each observation.
- **Explanatory Variables:** There are 39 variables that contain information about the specific observation. Bear in mind that you need to drop 3 of those.
- **Target Variables:** There are 2 target variable columns, both of which are in the train dataset, neither of which is in the test dataset.

For a more detailed description of the data, view “mimic_patient_metadata.xlsx”. Most variables are self-explanatory; the rest can be understood by reading the metadata file. If you have any questions even after looking into the metadata description, don’t hesitate to ask either to me or to Guillem.

III. The Task:

You can choose between two similar but distinctly different tasks:

1. **Classification:** Predict the class (0 or 1) of the “HOSPITAL_EXPIRE_FLAG” variable.
2. **Regression:** Predict the “LOS”, or length of stay at the ICU (each observation contains mostly characteristics of the patient measured *when entering the ICU*, therefore, you will be using *ex ante* information in order to forecast how long the patient will stay in the ICU.

You can also choose to work on **both** tasks and upload predictions for both test sets. In this case, your reward if your predictions are top of the class can be, potentially, doubled.

IV. The Submissions:

You are expected to submit the following items:

a. **Notebook:** A Jupyter notebook with your pipeline. It must contain the code to support your submitted predictions and explanations. Clean and organized notebooks will be appreciated (you can use markdown blocks to create sections, within-notebook links, and an index to tie everything together).

b. **Predictions:** You also must submit a .csv file containing your predictions for the test set (included). Note that your test sets do not include the true values for the target variables. These predictions will be evaluated by us using the same metric for everyone (metrics not disclosed to encourage a good methodological approach to generalizability of your models). You will find a sample submission file within the folder for each task. The values in those sample files are randomly generated.

   i. **Classification task:** You should present your predictions in probabilities (using .predict_proba, or a similar method, which most algorithms include) as this will maximize your scores.

   ii. **Regression task:** In this case, remember to make sure that your predictions are de-standardized.

Both submitted documents should be named following the schema

“<surname>_<name>_CML_2025.<csv/.ipynb>”

Where the elements inside each < > pair should be replaced with your actual name and (first) surname, for example, for students Joaquim Serra and Clara Puyol, their submissions should look like this:

“serra_joaquim_CML_2025.csv”
“puyol_clara_CML_2025.ipynb”

Note that we only require you to upload your notebook, not your full repository. Please work in an organized manner and contain your full project inside a main notebook. Additionally, if you use any package that goes beyond the usual ML stack (numpy, pandas, sklearn, matplotlib, seaborn, etc.), you should add a block warning about it, be it in a Markdown cell (```shell \\ uv add <xyz package> \\ ``` where “\\” refers to a line break and is not written in the markdown cell; look at the TA_session_1 notebook for an example), or in a Python cell (!uv add <xyz>).

V. The Presentation:

You are expected to "defend" your code and predictions on a short IN-PERSON evaluation. Each of you will be in front of the class for approximately 5-10 minutes, during which we will be projecting your code and will be asked questions about it.

Anyone who can answer each of our questions with a methodological or practical justification and follow up with a discussion about the modelling decisions they took will be awarded the maximum possible grade for the presentation, which will amount to a total of 9 points (with 1 additional point to be gained from the prediction ranking).

VI. The Evaluation:

Whichever task you choose, you will be evaluated on the following items:

a. **Working code:** Your code must run correctly, work and be appropriate to generate the resulting predictions that you submit.

b. **Understanding of the model:** You must understand your code and must be able to justify the design decisions across your pipeline and model. It's fine to say "This was the optimal option after performing model evaluation" in order to justify selecting a given model or set of hyperparameters, but then there must be evidence in your code of this evaluation step (GridSearch, RandomizedSearch or similar algorithms are your allies here).

   i. Small caveat: This justification is NOT valid to justify strictly code-wise solutions. *Id est*, you also MUST be able to explain what every line of your code does. We don't mind if you use coding assistance (Stack Overflow, classmates, Copilot or other LLMs, etc) since this is not a coding class; however, a lack of understanding of your own code is unacceptable and will be penalized.

c. **Metrics:** Your predictions will be scored using undisclosed metrics. We will create a ranking of prediction metrics. Everyone who is within a reasonable distance from the best predictions will be awarded some points (weighted by distance). Anyone who is unreasonably far away from those (closer to a random guesser than the best predictions in the class) will be awarded zero points. Anyone who submits the best predictions in the class (or a super close second/s) will be awarded 1 full point.

   i. Anyone who submits predictions from both tasks stands to position themselves in both rankings and, potentially, gain up to 2 points from this section. Unfortunately, the maximum grade is 10/10, so the grade of anyone who surpasses the 10/10 limit will be clipped to a straight 10. There are no "honors" in the strict sense, although if you want, we could create a public post on Google Classroom to "honor" anyone who reaches this status. This would only award a prize in "aura", with zero academic or legal repercussion besides a straight 10/10 in the project, though.