# Machine Learning for Healthcare - Final Project

0368-4273

**Submission date: 14/09/23**

## Overview

In this project, you will use observational health data to answer a predictive or descriptive question on a course-related topic. As discussed, we offer you one of the following options:

- **Option 1 - ICU prediction models:** Develop prediction models for 3 ICU outcomes (see section 3).
- **Option 2 - Suggested project:** Any other course-related research question (requires approval) (see section 4).

In both cases, your submission should comply with the submission requirements in section 1.

## 1 Final Submission Requirements

The final submission of all projects should include **(1) a paper** and **(2) your code** as follows:

### 1.1 Paper

The paper should describe your project, including the main analyses and results:

- Use ML workshop format (e.g., ICML or ICLR format). You can find templates on the web.
- Page limit: Up to 4 pages in the main text (not including references). Additionally, you can use an appendix (up to additional 6 pages). Submission in a single PDF file.
- Writing is in English.

It is recommended that your paper contain the following sections:

1. **Introduction:** Introduction and background to your research question.
2. **Cohort Description:** An explanation of the dataset(s), including cohort characteristics (size, comparison between control and positively labeled patients, etc.) and your extracted features.
3. **Methods:**
   (a) **Inclusion and exclusion criteria**
   (b) **Data exploration and preprocessing:** Explain your data exploration process and the subsequent data preprocessing steps, e.g., feature analyses, statistical testing, feature engineering, data imputation, feature scaling, feature selection, outlier removal, etc.
   (c) **Models:** Describe your models. When applying existing models, cite them and briefly explain your modeling choices. If you develop something new – describe it in detail.
   (d) **Evaluation:** Explain how you evaluated your model (e.g., cross-validation, bootstrapping)

4. **Results:** Report and describe your results with relevant figures and/or tables.
   - The figures should be informative with clear legends and captions.
   - You are recommended to explore different aspects of results, e.g., model selection, evaluation over patient subgroups, etc.
   - Mandatory reports:
     - *Option 1:* You must describe your mandatory evaluation performance results detailed in section 3.4), including model performance, calibration, and feature importance.
     - *Option 2:* Provide a comprehensive report of your results with the relevant evaluation metrics.
5. **Discussion:** Summarize your results, clinical insights (if any), limitations, and main conclusions.
6. **Bibliography**

### 1.2 Code:

You should provide the code used for your analysis, as follows:

- Attach a link to your GitHub repository in a footnote of the first paper's page.
- The code should be written in **Python**. You might use open-source libraries such as Numpy, Pandas, Matplotlib, Scikit-learn, TensorFlow, Pytorch, SciPy, and Keras.
- The code should be clear and we will evaluate it for readability, documentation, and code quality.
- Add a README file providing high-level documentation of your project (e.g., describe the purpose of each file).
- Add a *requirements.txt* specifying your required libraries. If none, leave it as is with one empty line.
- For *Option 1*: Your model will be evaluated on unseen data. See requirements in 3.5.

## 2 Project Evaluation

Your project will be mainly evaluated based on:

- Compliance with the instructions
- Novelty and creativity beyond the mandatory requirements. You are recomended to incorporate issues mentioned in the course or from literature such as ethics, discrimination, robustness, etc.
- Final performance results on seen and, for option 1, on unseen subsets (see section 3.1).
- Code quality, readability, and documentation.
- Paper quality: Writing quality, organization, clarity, etc.

### Notes & Tips

- **Plan you work**, e.g., before running fast, you can start simple (develop a baseline model)
- **Understand and explore your data:** Your patient cohort and features (e.g., distributions, outliers) – Preprocess them carefully!
- Try to account for **statistical, analytical, and ethical challenges** discussed in the course
  - E.g., noise, missing data, biases, discrimination, interpretability, generalizability, etc.
- **Literature review:** Look for similar research papers - *Can we do better?*

- Account for **data leakage**
  - Perform data split carefully
  - Perform pre-preprocessing separately in train/test folds
- Provide **uncertainty measures** in your reports (confidence intervals, bootstrap, SD, etc.).

**You are more than welcome to consult with us throughout the project!**

# 3 Guidelines for Option 1: ICU Prediction Models

In this project, you will use a **subsample of MIMIC-III** to predict **3 clinical targets** defined in section 3.2. You should implement **end-to-end pipeline** that includes data extraction, data preprocessing (including target definition), model(s) training, and evaluation, based on methods introduced in the course or from the literature.

## 3.1 Data

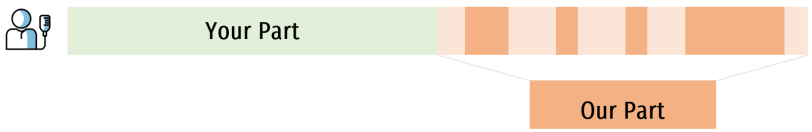
You will develop your pipeline solely based on a predefined patient subsample extracted from MIMIC-III (**your part**). After submission, we will evaluate your trained submitted model(s) on a disjoint **unseen** patient subsample (**our part**) from MIMIC-III with similar characteristics. Specifically:

- **Your part:** *initial_cohort.csv* contains the list subject IDs representing your *initial cohort*.

  **This is the only cohort you can use for the project** – Use it properly to perform inclusion & exclusion criteria, data preprocessing, data partition, model development, and evaluation.

  **Note:** We did not perform any filters on this subset!

- **Our part:** Your trained model will be evaluated on unseen patient data. This entails some technical requirements which are listed in section 3.5.




## 3.2 Targets Definition

Your model(s) should predict each of the following ICU targets:

- **Mortality** during hospitalization or up to 30 days after discharge.
- **Prolonged stay:** length of stay $> 7$ days.
- **Hospital readmission** in 30 days after discharge (not to be confused with ICU readmission within the same hospital admission).

**Note:** Each patient can experience more than one clinical target.

## 3.3 Prediction Timeline

For each patient, we will perform a single prediction of each of the clinical targets for the first admissions, using data collected at the first 42 hours of hospitalization.

Implement your pipeline according to the following timeline:

- Focus only on **first** hospital admissions in cases of multiple admissions available.
- Consider only patients with **at least 48 hours** of hospitalization data.
- To preserve a **6-hour prediction gap**, use only data collected in the first 42 hours for prediction. Specifically, the **prediction time** for each patient is 42 hours since admission ($t = 42$), using only data collected up to that time point.

In terms of input representation and modeling, e.g., how to represent the 42-hours input - this is your choice.

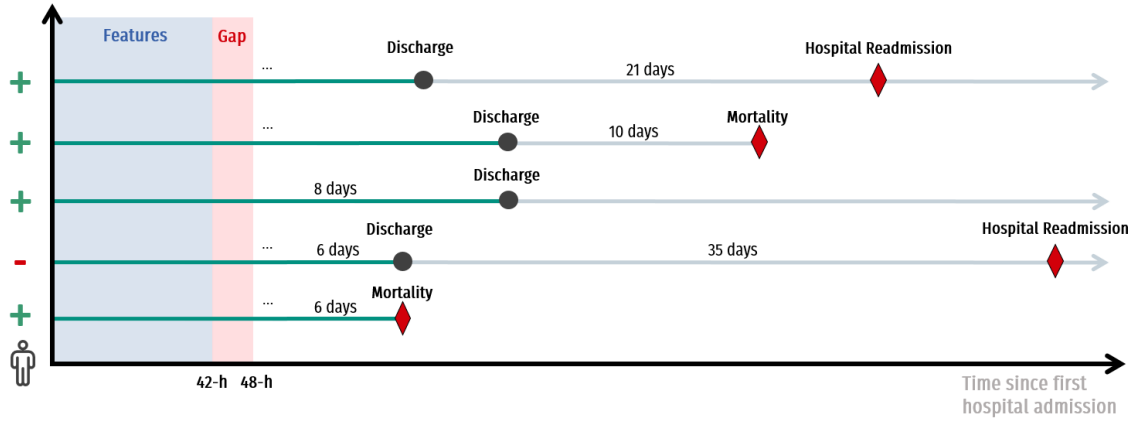



Figure 1: Example of 4 labeled (*y-axis*) patient timelines (*x-axis*). '+', '-': positively and negatively labeled, respectively. Green line: during hospitalization, gray line: after discharge.

## 3.4 Pipeline

Your pipeline must contain some **mandatory pre-defined pipeline requirements**, described below. Besides, you are recommended to design and implement your ideas beyond the requirements, using methods introduced in the course and related literature (with credit). You can find some suggestions in the *course summary* slides on Moodle.

**Pipeline requirements:**

1. **Data Extraction:**
   You must extract and preprocess at least several features of each of the following types:
   (a) Demographic features (e.g., age, gender).
   (b) Vital signs
   (c) Laboratory test results
   (d) At least two additional modalities:
   - Explore MIMIC and pick at least **two additional modalities** (e.g., microbiology, habits, data extracted from notes, non-vital chart events, past medical history, inputs, outputs, prescriptions, etc.).
   - Extract a few features per modality (e.g., a few background diseases, medications, etc.).
   - Use MIMIC documentation and a broad literature review to preprocess the new features and include them in your analysis. Explain your choices and analyses in the paper.

   **Notes:**

- If you wish, you can later perform automatic feature selection to reduce feature dimensions.
- If you need some help with defining new queries you can contact us.
- You may use to code provided in HW2 for medical history processing, but not the partial CSV file (which is incomplete).

2. **Target Definition:** According to section 3.2.
3. **Data Partition:** Your choices.
4. **Data Preprocessing:** Your choices.
5. **Modeling:** Your choices.
   - Your final model must return **calibrated probabilities**.
6. **Evaluation:** Your model must be evaluated using **at least** each of the following aspects:
   - **Classification performance:** e.g., plot ROC and PR curves.
   - **Calibration performance:** e.g. plot calibration curve.
   - **Feature importance:** e.g., analyze the important features using an interpretability method (e.g., SHAP).

### 3.5 Evaluation on Unseen Data & Related Code Requirements

We will evaluate your trained submitted model(s) by running it on unseen data (see section 3.1) that is disjoint from your training data, but with similar characteristics.

Your final repository should include a folder named *project* that contains:

- **Your code:** All the files and resources required for running your pipeline (e.g., PY files, saved models, training parameters for preprocessing, etc.)
- ***__init__.py*** (given).
- ***requirements.txt***
- **unseen_data_evaluation.py**: To be able to run your pipeline on our unseen subset, we provided a file named *unseen_data_evaluation.py* that defines the function *run_pipeline_on_unseen_data*. This function should take subject IDs as input (a *test* set), extract their data, apply your pre-trained model, and generate prediction probabilities of each patient to have each of the outcomes (at time $t = 42$-h). **Note:** It should also perform the necessary data transformation and processing steps to the extracted data.

  See the function's documentation and **implement** *unseen_data_evaluation.py* accordingly.

Tester notebook: We've provided a notebook running your final pipeline. **Use it to make sure your code is ready for testing!**

To run it:

- Create a copy of the notebook and insert your BigQuery project id in the relevant place.
- Upload your zipped *project* folder.
- Upload the *test_example.csv* and run the code.

# 4 Option 2: Suggested Projects

After approval, you are free to go. **Note** to comply with the final submission guidelines regarding the paper scope and code quality.

## Good Luck!