Data Science Lab: Process and methods
Politecnico di Torino

Project Assignment
Summer Call, A.Y. 2023/2024

Last update: June 9, 2024

1 Project dates

Start date: June 9, 2024 at 23:59 (CET)
Due date: June 24, 2024 at 23:59 (CET)

Due date is a **strict deadline**.

2 Problem description

Predicting patient survival rates in critically ill populations is a vital aspect of modern healthcare, with profound implications for clinical decision-making and patient management. Accurate predictions can inform timely and appropriate interventions, enhance the quality of life for patients, and aid in resource allocation within healthcare systems. By anticipating the likely outcomes of severe medical conditions, healthcare providers can prioritize care, implement palliative measures when necessary, and support patients and their families in making informed decisions about treatment options. This predictive capability is crucial for addressing the growing concerns surrounding end-of-life care, ensuring that patients receive compassionate and effective care that aligns with their wishes and medical needs.

The primary goal of this project is to develop a machine learning model that accurately predicts the binary outcome "death" for these patients, which signifies whether a patient survives or dies within a specified time frame.

2.1 Dataset

This dataset comprises 9105 individual critically ill patients across 5 United States medical centers, accessioned throughout 1989-1991 and 1992-1994. Each row concerns hospitalized patient records who met the inclusion and exclusion criteria for nine disease categories: acute respiratory failure, chronic obstructive pulmonary disease, congestive heart failure, liver disease, coma, colon cancer, lung cancer, multiple organ system failure with malignancy, and multiple organ system failure with sepsis. The goal is to determine if these patients survived based on several physiologic, demographics, and disease severity information. It is an important problem because it addresses the growing national concern over patients' loss of control near the end of life. It enables earlier decisions and planning to reduce the frequency of a mechanical, painful, and prolonged dying process.

The features reported in the dataset are:

- `id`: an identifier of the sample.
- `age`: Age of the patients in years
- `sex`: Gender of the patient. Listed values are male, female.

- `dzgroup`: The patient's disease sub category amongst ARF/MOSF w/Sepsis, CHF, COPD, Cirrhosis, Colon Cancer, Coma, Lung Cancer, MOSF w/Malig.
- `dzclass`: The patient's disease category amongst "ARF/MOSF", "COPD/CHF/Cirrhosis", "Cancer", "Coma".
- `num.co`: The number of simultaneous diseases (or comorbidities) exhibited by the patient. Values are ordinal with higher values indicating worse condition and chances of survival.
- `edu`: Years of education
- `income`: Income of the patient. Listed values are "$11-$25k", "$25-$50k", "$>50k$","$under$11k".
- `scoma`: SUPPORT day 3 Coma Score based on Glasgow scale (predicted by a model).
- `charges`: Hospital charges
- `totcst`: Total ratio of costs to charges (RCC) cost
- `totmcst`: Total micro cost
- `avtisst`: Average TISS score, days 3-25, where Therapeutic Intervention Scoring System (TISS) is a method for calculating costs in the intensive care unit (ICU) and intermediate care unit (IMCU).
- `race`: Race of the patient. Listed values are asian, black, hispanic, missing, other, white.
- `sps`: SUPPORT physiology score on day 3 (predicted by a model).
- `aps`: APACHE III day 3 physiology score (no coma, imp bun, uout for ph1)
- `surv2m`: SUPPORT model 2-month survival estimate at day 3 (predicted by a model)
- `surv6m`: SUPPORT model 6-month survival estimate at day 3 (predicted by a model)
- `hday`: Day in hospital at which patient entered study.
- `diabetes`: Whether the patient exhibits diabetes (Com 27-28, Dx 73) as a comorbidity (Y) or not (N).
- `dementia`: Whether the patient exhibits dementia (Comorbidity 6) as a comorbidity (Y) or not (N).
- `ca`: Whether the patient has cancer (yes), whether it has spread out (metastatic), or if it is healthy (no).
- `prg2m`: Physician's 2-month survival estimate for patient.
- `prg6m`: Physician's 6-month survival estimate for patient.
- `dnr`: Whether the patient has a do not resuscitate (DNR) order or not. Possible values are dnr after sadm, dnr before sadm, missing, no dnr.
- `dnrday`: Day of DNR order (<0 if before study)
- `meanbp`: mean arterial blood pressure of the patient, measured at day 3.
- `wblc`: counts of white blood cells (in thousands) measured at day 3.
- `hrt`: heart rate of the patient measured at day 3.
- `resp`: respiration rate of the patient measured at day 3.
- `temp`: temperature in Celsius degrees measured at day 3.
- `pafi`: PaO_2/FiO_2 ratio measured at day 3. The ratio of arterial oxygen partial pressure (PaO2 in mmHg) to fractional inspired oxygen (FiO2 expressed as a fraction). Widely used clinical indicator of hypoxaemia, though its diagnostic utility is controversial. Specific ranges of values can be associated with different levels of mortality.

- `alb`: serum albumin levels measured at day 3.
- `bili`: bilirubin levels measured at day 3.
- `crea`: serum creatinine levels measured at day 3.
- `sod`: serum sodium concentration measured at day 3.
- `ph`: Arterial blood pH. The pH of blood is usually between 7.35 and 7.45. Abnormal results may be due to lung, kidney, metabolic diseases, or medicines. Head or neck injuries or other injuries that affect breathing can also lead to abnormal results.
- `glucose`: Glucose levels measured at day 3.
- `bun`: Blood urea nitrogen levels measured at day 3.
- `urine`: Urine output measured at day 3.
- `adlp`: Index of Activities of Daily Living (ADL) of the patient, filled out by the patient. Higher values indicate more chance of survival, measured at day 3.
- `adls`: Index of Activities of Daily Living (ADL) of the patient, filled out by a surrogate (e.g. family member), measured at day 3. Higher values indicate more chance of survival.
- `adlsc`: Imputed ADL Calibrated to Surrogate.
- `death`: Death at any time up to National Death Index (NDI) data on 31 of December of 1994. Some patients are discharged before the end of the study and are not followed up. The authors looked up the information about death. **This is the binary target variable to be predicted**.

The dataset is located at this URL.

Within the archive, you will find the following elements:

- **development.csv** (development set): a comma-separated values file containing the 7,285 patients' information for the development set. This portion has the "death" feature to be predicted for each event. This feature is the target value to be used to train and validate your models.
- **evaluation.csv** (evaluation set): a comma-separated values file containing the 1,820 patients corresponding to the evaluation set. This portion does not contain the "death" target variable.

2.2 Task

You are required to build a binary classification pipeline to predict the target variable, i.e., "death" columns of development.csv file.

2.3 Evaluation metric

Your submissions will be evaluated in terms of the F1 score. Here you can find the function used to evaluate your submissions.

3 Submit your result

Submission file To get your results evaluated, you have to upload a result file on our submission competition platform, Data Science Lab Environment (DSLE). The submission file must be a `CSV` file formatted as follows:

```
Id,Predicted
0,0
1,0
2,1
3,1
4,0
...
```

The submission file must contain a header line and a row for each record in the Evaluation collection. Each row must have two fields:

- the `Id` of the corresponding record in the Evaluation set. It corresponds to the column `Id` in the evaluation CSV file.
- the `Predicted` binary outcome. It must be in numerical form, e.g., 0 or 1, as the development set provides.

You can find a sample submission file in the project material (see 2.1).

Submission platform The submission platform is the same one you used during the course laboratories. Therefore, you have to use your personal key. If you do not have the key or have problems using it, please send an email to `lorenzo.vaiani@polito.it`. Please refer to the guide on the course website to go through the submission procedure.

You can find the DSLE platform at http://trinidad.polito.it:8888

4 Upload the report and the software

The report and the software have to be submitted by the due date reported in Section 1. This is a strict deadline.

Submission All the required files (i.e., for the report and the software) must be included in a **single ZIP archive**. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section. Please use as description: **report_exam_summer_2024**.

Info: A ZIP archive is a ZIP archive, not a RAR, a 7z, or a tarball archive, nor any of those renamed with a trailing `.zip` extension.

Formatting rules The formatting rules for both the report and the software are described in the document with exam rules. You can find it on the course website.

5 Fill in the LLM usage form

As discussed in the exam rules, adoption of Large Language Models (e.g. ChatGPT) is allowed for the production of the report (**not** for the implementation of the solution). Each team **must** provide information about whether they used, and to which extent they did, LLM-based tools.

To do so, please fill in this form by the due date of this project. Failure to do so will result in a void project.

Warning: This is an additional requirement that was not required in past years. Make sure you remember to fill in the form by the due date, or your project will not be considered valid!