Data Analytics for Digital Health Project

Analysis of hospitalized patients

A.Y. 2024/2025

A **project** consists of data analysis based on data mining tools. The project has to be performed by a team of 2 students. It has to be performed by using Python. The guidelines require addressing specific tasks, and results must be reported in a unique paper. This paper's total length must be **25 pages** of text including figures. The students must deliver both: paper and well-commented Python notebooks. The Python notebooks need to be executable, hence all the necessary dependencies and libraries need to be stated at the beginning of each notebook.

Dataset description

The dataset is composed of 4 csv files and additional files in .zip format.
The data combines ECG recordings with clinical information. Essentially, it provides a table (**patients.csv**) that links ECG recordings to the discharge diagnoses, coded in ICD-10-CM, for the corresponding patients in the ECG dataset.
The dataset, called patients_main_table.csv, was created by aligning the times of ECG recordings with patient admission and discharge times and diagnostic codes in ICD-10-CM (to know more about them: here).
The main table (**patients.csv**) includes various columns, such as:

- **ts_ecg_path**: path to the waveform
- **study_id**: study ID within MIMIC-IV-ECG
- **ed_stay_id**: identifier for ED stays
- **ed_hadm_id** and hosp_hadm_id: identifiers for hospital admissions
- **diagnosis**: set of ICD-10-CM codes for hospital discharges
- **subject_id**: patient ID within MIMIC-IV-ECG
- **ecg_time**: time of ECG recording
- **gender**: patient gender
- **age**: patient age at the time of ECG recording
- **dod**: date of death, if applicable
- Boolean variables (**ecg_taken_in_ed**, **ecg_taken_in_hosp**, **ecg_taken_in_ed_or_hosp**) indicating where the ECG was taken.

These identifiers also allow users to retrieve additional clinical information from the other csv files, beyond the basic demographic data included in the table.
The additional data are:

1. **Triage**: The dataset contains information collected from patients during triage in the emergency department (ED). Triage is the initial assessment process used to determine a patient's health status and reason for visiting the ED. The dataset includes the following columns:

- **subject_id**: patient ID (also available in the other .csv files)
- **stay_id**: ID for the ED visit
- **temperature**: body temperature
- **heartrate**: heart rate
- **resprate**: respiratory rate
- **o2sat**: oxygen saturation level
- **sbp**: systolic blood pressure
- **dbp**: diastolic blood pressure
- **pain**: patient-reported pain level
- **acuity**: severity level assigned by the care provider, ranging from 1 (highest severity) to 5 (lowest severity)
- **chiefcomplaint**: the patient's reported reason for coming to the ED

Vital signs collected during triage (temperature, heart rate, respiratory rate, oxygen saturation, systolic and diastolic blood pressure) are documented numerically. The pain column represents the patient's reported pain level, and the chiefcomplaint column contains a free-text description of their reason for visiting, often with multiple complaints separated by commas. Any personally identifiable information (PHI) in chiefcomplaint has been replaced with three underscores ("___").

2. **Vital sign**: this table contains vital signs that were documented at different times during a patient's stay in the hospital. It includes the following columns:

- **subject_id**: patient ID
- **stay_id**: ID for the stay
- **charttime**: time when the vital signs were recorded
- **temperature**: body temperature
- **heartrate**: heart rate
- **resprate**: respiratory rate
- **o2sat**: oxygen saturation level
- **sbp**: systolic blood pressure
- **dbp**: diastolic blood pressure
- **rhythm**: patient's heart rhythm
- **pain**: patient-reported pain level

The vital signs recorded in this table are similar to those in the triage table, but the vital_sign table also includes a rhythm column to capture the patient's heart rhythm, and the charttime column indicates when the measurements were taken. These vital signs are documented at multiple points during the patient's stay, rather than just during the initial triage.

3. **Omr**: this table contains miscellaneous health information from the Online Medical Record, which is useful for understanding an individual's health. It includes the following information:

- Blood pressure

- Height
- Weight
- Body Mass Index (BMI)
- Estimated Glomerular Filtration Rate (eGFR)

These values are collected from both inpatient and outpatient visits. In many cases, a "baseline" value is also available, which represents a patient's health status before hospitalization.

Task1: Data Understanding and Preparation (30 points)

Task 1.1: Data Understanding

Explore the various dataset with the analytical tools studied and write a concise “data understanding” report assessing data quality, the distribution of the variables and the pairwise correlations.
We strongly suggest performing 2 different data understanding: one for the tabular data, another one for the ecgs. Only after having analyzed the data separately, you may consider merging some information.
In addition, pay attention to the fact that the datasets are not small, hence loading all of them at the same time may be unfeasible, depending on the laptop you are working with. We suggest analyzing them in different stages, cleaning and saving the data along the way, keeping in memory only what is necessary.
For the data understanding you may exploit the information in the column diagnosis: you have a list of all the diagnoses associated to the patient under analysis. To exploit them, you need to refer to the ICD-10 codes from here. The structure of these codes is fundamental and you are going to use them throughout your studies and maybe even during your work, so spend time understanding them. A possible solution for this project may be to map the codes in strings with meanings (just a hint: mapping, but you can find even better solutions!).

Task 1.2: Data Preparation

Improve the quality of your data and prepare it by extracting new features interesting for describing the patients. Therefore, you are going to describe the information patient wise and examples of indicators to be computed are:

- What is the highest heart rate registered?
- What is the lower heart rate registered?
- How many times did the doctors perform blood analysis?
- What is the mean for the temperature of the patient?
- What is the entropy of the respiratory rate?
- Is the respiratory rate increasing or decreasing during the stay of the patient?

Note that these examples are not mandatory. You can derive indicators that you prefer and that you consider interesting for describing the patients.

It is MANDATORY that each team defines some indicators. Each of them has to be correlated with a description (in which should be clearly stated the objective of the variable derived) and when it is necessary also its mathematical formulation.
The extracted variables will be useful for the clustering analysis (i.e., the second project's task). Once the set of indicators is computed, the team has to explore the new features for a statistical analysis (distributions, outliers, visualizations, correlations).

Subtasks of DU:

- Data semantics for each feature (min, max, avg, std) above and the new one defined by the team
- Distribution of the variables and statistics
- Assessing data quality (missing values, outliers, duplicated records, errors)
- Variables transformations
- Pairwise correlations and eventual elimination of redundant variables.

Nice visualization and insights can be obtained, explore the web to get more ideas!

Task 2: Clustering analysis (30 POINTS - 32 with optional subtask**)

**For groups (3 or more people) the optional part is mandatory.

Based on the features extracted in the previous task, explore the dataset using various clustering techniques. Carefully describe your decisions for each algorithm and which are the advantages provided by the different approaches.

Subtasks

- Clustering Analysis by K-means on the entire dataset:
    1. Identification of the best value of k
    2. Characterization of the obtained clusters by using both analysis of the k centroids and comparison of the distribution of variables within the clusters and that in the whole dataset
    3. Evaluation of the clustering results
- Analysis by density-based clustering:
    1. Study of the clustering parameters
    2. Characterization and interpretation of the obtained clusters
- Analysis by hierarchical clustering:
    1. Compare different clustering results got by using different version of the algorithm

2. Show and discuss different dendrograms using different algorithms
- Final evaluation of the best clustering approach and comparison of the clustering obtained
- **Optional (2 points):** Explore the opportunity to use alternative clustering techniques in the library: https://github.com/annoviko/pyclustering/

How to work with your data

In this project, two types of data are considered: tabular data and time series data (the ECG for each patient). During the lessons, we examined various techniques for assessing similarity among time series, as well as several methods applicable to tabular data. For the clustering task, you have the option to analyze these two data types either separately or together. We recommend exploring both approaches, discussing them in your report, and selecting the approach that yields the best results, discussing them.