Data Analytics for Digital Health Project

Analysis of hospitalized patients

A.Y. 2025/2026

A **project** consists of data analysis based on data mining tools. The project has to be performed by a team of 2-3 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 consists of four CSV files, each describing a different aspect of patient records collected during hospital stays. All files can be linked through the hadm_id field, which uniquely identifies a hospital admission. The dataset integrates clinical notes, laboratory results, microbiology tests, and procedure codes. It is delivered in four separate files only to facilitate handling and reduce computational complexity.

In the following, a description of the content of each dataset.

1. Heart Diagnoses

This table contains clinical notes describing patients' conditions and examination results. Diagnostic codes are also included.

Columns:

- note_id: Unique identifier of the clinical note;
- subject_id: Unique patient identifier (also available in other CSV files);
- hadm_id: Hospital admission identifier (also available in other CSV files);
- note_type: Kind of note (discharge summary, progress note, …);

- note_seq: Sequence number, only if multiple notes are recorded for the same admission id;

- charttime: Timestamp when the note was written;

- storetime: Timestamp when the note was stored. It may be at the same time of charttime or later;

- HPI: History of Present Illness (text);
- information about the subject, such as the age or the date of death;

- physical_exam: Textual description of the findings of physical examination;

- chief_complaint: Principal patient complaint reported at admission;

- invasions, X-ray, CT, Ultrasound, CATH, ECG, MRI: Indicators for examinations or procedures performed;

- reports: Reports, separated by "|" if more are available;

- icd_code: Diagnostic code (ICD-9 or ICD-10);

- long_title: Description of the diagnosis code.

2. Laboratory Events

This table contains results of laboratory tests recorded during hospital admissions. It provides both numeric and textual information, together with information about measurement units and abnormality indicators.

Columns:

- hadm_id: Hospital admission identifier;

- charttime: Timestamp when the laboratory test was recorded;

- value: Raw test result (in free text);

- valuenum: Numeric value of the test result, if available;
- valueuom: Unit of measurement (mg, dL, …);
- ref_range_lower - ref_range_upper: Lower and upper bounds of the reference range;
- flag: Abnormality indicator;
- label: Name of the laboratory test;
- fluid: Type of fluid analyzed (blood, urine, …);
- examination_group: Higher-level group for the test (Chemistry, Hematology, …);
- analysis_batch_id: Identifier of the processing batch;
- qc_flag: Quality control flag;
- ref_range: Reference ranges;

3. Microbiology Events

This table records results of microbiology examinations, including specimen type and antibiotic susceptibility testing.

Columns:

- subject_id: Patient identifier;
- hadm_id: Hospital admission identifier;
- charttime: Timestamp of the result or sample collection;
- spec_type_desc: Description of the specimen type (blood, urine, …);
- test_name: Name of the microbiological test performed;
- org_name: Name of the organism identified;
- ab_name: Name of the antibiotic tested;
- dilution_text: Text;
- dilution_comparison: Operator associated with the dilution value (<, >, …);

- dilution_value: Numeric value of the dilution;
- interpretation: Interpretation of the result (Sensitive, Resistant, …);
- technician_id: Identifier of the technician;
- qc_flag: Quality control flag.

4. Procedure Codes

This table contains information on procedures carried out during hospital admissions, represented by standardized codes and textual descriptions.

Columns:

- subject_id: Patient identifier;
- hadm_id: Hospital admission identifier;
- seq_num: Sequence number of the procedure within a hospital admission. The value 1 indicates the primary procedure, while larger values correspond to additional procedures performed during the same admission;
- chartdate: Date when the procedure was recorded;
- icd_code: Procedure code (ICD-9, ICD-10);
- long_title: Textual description of the procedure.

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.

There is a wide variety of data available and we strongly encourage you to link the information across the different CSV files provided.

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:

- Highest value of glucose recorded for the patient during the admission?
- Total count of laboratory events linked to the admission?
- Ratio between the number of tests flagged as abnormal and the total number of tests?
- Total count of microbiology examinations for the admission?
- Total count of procedure codes linked to the admission?

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! Please add these substaks in your final report.

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

Based on the features extracted in the previous task, explore the dataset using the clustering techniques presented during the lessons. You should explore the data after the creation of your patient profile, hence after Data Understanding and Pre Processing. Carefully describe your decisions for each algorithm and which are the advantages provided by the different approaches.

Subtasks for tabular data

- 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/

Task 3: Time series analysis (30 points)

The time series data provided for this project is composed of ECG signals with 12 channels for each record. We suggest building one time series for each patient: create one time series for each subject_id and focus on simple, preferably univariate series (each series should represent one specific type of information at a time). To achieve this, you should pre-process the time series with the tools and approaches seen during the lessons. Then, you can decide to approximate your time series or to extract features from it.

Note: the series can be short, and this is not a problem, just focus on the consistency of your chosen subtask rather than the length of the sequence.

- Always start from the patient: create one time series for each subject_id containing the timestamps and the corresponding values for the chosen event type.

- For diagnostic data, use the codes from the heart_diagnoses table, selecting those associated with each subject_id. These codes, as well as all the other information extracted from the tabular data, can enrich your analysis also from the time series point of view.

Task 4: Classification analysis (30 POINTS - 32 with optional subtask**)

From the original csv files, in particular from heart_diagnoses_1.csv, each patient is associated with one or more cardiovascular diagnostic codes (ICD-9/10, chapter "I").
To enable a supervised classification task, we convert these diagnoses into a binary label that distinguishes between **ischemic** and **non-ischemic** cardiovascular conditions.
To obtain a binary classification task, we need to group some codes. The rule used is the following. For **Class 1 (Ischemic)** we group:

1. **I20** – Angina pectoris
2. **I21** – Acute myocardial infarction
3. **I22** – Subsequent myocardial infarction
4. **I24** – Other acute ischemic heart diseases
5. **I25** – Chronic ischemic heart disease

All other cardiovascular codes in the dataset, such as arrhythmias, myocarditis, valve diseases or heart failure, are assigned to **Class 0 (Non-ischemic)**. Please note that in the original csv file there are some patients with more than one diagnosis. In this case, if a subject has **at least one** ischemic diagnosis (I20–I25), the subject is labeled as Class 1. Otherwise, the subject is labeled as Class 0.

Using the patient profile(s) created in the previous task, each group should:

- *Create the label for each subject.* Derive the binary label for every subject_id, following the rule described above. In addition, a .py file is provided that extracts the correct labels for all subjects in the original CSV that you can use as a reference.
- *Train and evaluate predictive models.* Train the classification models discussed during the lessons (e.g. logistic regression, decision trees, random forests, gradient boosting, etc.). Train at least 3 classification models. Compare their performance and justify the choice of models.
- *Apply and discuss pre-processing steps.* Apply all necessary preprocessing steps to make the dataset suitable for classification, such as: encoding categorical variables, selecting features, addressing possible class imbalance, problems of correlation.
- *Evaluate and analyse results.* Report performance metrics on both training and test sets (e.g. accuracy, balanced accuracy, precision, recall, F1-score, ROC-AUC, confusion matrices). Compare the performance of the different models and discuss which configuration appears most appropriate for this task.

Optional (2 points): As an optional task, perform a **multiclass classification** task by grouping the ICD codes into 3, 4 or 5 **clinically meaningful categories**. The selection of the number of classes and the meaning of each class is up to you.
An example of a possible mapping may be:

1. **Ischemic heart disease** (With angina, acute myocardial infarction, subsequent infarction, chronic ischemic disease, like the label for the binary classification.)
2. **Arrhythmias and conduction disorders** (With conduction blocks, supraventricular and ventricular arrhythmias, atrial fibrillation/flutter, cardiac arrest.)
3. **Valvular and pump failure disorders (**With mitral, aortic, and tricuspid valve disorders,heart failure and cardiomyopathy.)
4. **Inflammatory and structural heart disease (**With pericarditis, endocarditis, myocarditis, structural abnormalities.)

Task 5: Time series Clustering (30 POINTS - 32 with optional subtask**)

Using the time series pre-processed in the previous task, explore them using the clustering techniques presented during the lessons. You should explore the data after the pre-processing of your time series. In this part, it is expected that you take into account only the time series (and not the tabular data). You can work with them in the form of time series (univariate is suggested but not mandatory) or in feature representation. You should apply at least 2 different kinds of clustering techniques.

Optional: explore the possibility of adding the information about the patients extracted from the tabular data.

Task 6: Time series Classification (30 POINTS - 32 with optional subtask**)

Given the pre-processed time series, limit your analysis to those for which corresponding tabular data are available (merge on the patient id). In this way, you can use the same labels employed in the binary classification task of the tabular dataset.
NB: only the time series should be used as input for the classification model. The connection to the tabular data serves only to obtain the labels.

Optional*:* Explore whether patient information derived from the tabular data can be incorporated into the analysis.

Deadline 5 January 2026.