Nicola De Cristofaro (Matr. 0522500876) **Cloud Computing Curriculum**

Course Project: Predictive Models derived from Electronic Health record EHR

Introduction & Objectives

The healthcare sector is one of the fundamental sectors of our society, and unfortunately it is often one of the most overloaded and full of problems. Especially in times like the one we are experiencing, the responsibility for this sector increases considerably and all the weak points are beginning to be noticed.

For this reason, one of the points from which we must start again to defeat the pandemic in progress is the strengthening of hospitals and all the systems that manage them.

One way that certainly manages to bring tangible benefits to the hospital sector is the use of Machine Learning. Machine Learning could be used to predict illness and treatment to help physicians and payers intervene earlier, predict population health risk by identifying patterns and surfacing high risk markers and model disease progression. Machine learning could also help pathologists make quicker and more accurate diagnoses as well as identify patients that might benefit from new types of treatments or therapies. These are just a few examples of the large set of benefits that Machine Learning can give to this sector.

In this project, making use of the EHR (Electronic Health Record) which is a systematic collection of information on the health of individuals in digital format, we focus on defining predictive models that allow us to predict:

- **Hospitalization time**: length of stay in hospital
- **In-hospital mortality**: death occurred during the hospital stay
- **Readmission within 30 days**: readmission to hospital within 30 days of last admission
- **Readmission within 90 days**: readmission to hospital within 90 days of last admission
- **Readmission within 365 days**: readmission to hospital within 365 days of last admission

Dataset

The dataset used in this project is MIMIC (Medical Information Mart for Intensive Care) openly accessible, created by the Laboratory of Computational Physicology of The Massachusetts Institute for Technology (MIT) with the goal of providing tools for the creation of clinical knowledge through the application of data analysis techniques.

MIMIC-III is aversion from 2016, it is a large, freely-available relational database comprising **deidentified** health related data associated with over forty thousand patients who stayed in Intensive Care Units at Beth Israel Deaconess Medical Center (Boston, Massachusetts). The data spans June 2001 - October 2012.

MIMIC-III is structured in a relational manner and contains 26 linked tables through patient identifier:

- Each table contains each patient record (at each row) with specific field (columns).
- Tables start with 'D_' are dictionaries and provide definitions for identifiers.

There is also an updated version **MIMIC-IV** (2020), which incorporates contemporary data and improves on numerous aspects of MIMIC-III. MIMIC-IV adopts a modular approach to data organization, highlighting data provenance and facilitating both individual and combined use of disparate data sources.

In this project we use MIMIC-IV to study how the data regarding patients are evolved through years since there are already a lot o papers and study on MIMIC-III.

General Information on tables

- Each patient is unique with its own **"subject_id"**
- Each hospital admission of a patient is unique with **"hadm_id"**
- Each ICU (Intensive Care Unit) stay of a patient is unique with **"icustay_id"**

This means that:

- One **subject_id** can be associated with multiple **hadm_ids** when a patient had multiple admissions.
- One **hadm_id** can be linked to multiple **icustay_id** when a patient had a multiple ICU stays during an admission. (e.g., transferring between multiple ICUs)

For the complete descriptions of the table's structure here is the associated documentation: **MIMIC-IV module's structure** -> https://mimic.mit.edu/docs/iv/modules/

Data Exploration and Preparation

The first step was exploring the tables, find useful insights, then merge the tables to compose as baseline dataset where to start the analysis for our goals.

Below the process to extract and prepare the data:

View FILE -> *01_data_exploration_preparation.pdf*

We use Jupyter Notebook to describe these steps in order to show also the python code used to reach our target.

Now, for each prediction we want to get, starting from the baseline dataset created at the previous step, we go through the following steps:

1. Problem Statement
2. Type of model to use for prediction
3. Metrics used for validation
4. Features distribution and features engineering
5. Data cleaning
6. Features Selection
7. Split in training/test set
8. Prediction Model choice
9. Prediction Model validation
10. Parameter Tuning
11. Results discussion

Hospital LOS (Length-of-Stay)

View FILE -> *02_length_of_stay.pdf*

In-Hospital Mortality

View FILE -> *03_in-hospital-mortality.pdf*

Readmission within 30, 90, 365 days

View FILE -> *04_in-hospital-mortality.pdf*