Data Analytics in Healthcare and Connected Care
Vrije Universiteit Brussel
Project Description
2022-2023

March 14, 2023

Lecturer/s: Bart Jansen, Jef Vandemeulebroucke, Nikolaos Deligiannis

Teaching Assistant/s: Redona Brahimetaj, Joris Wuts, Soha Sadat Mahdi

1 Outline

Healthcare institutions collect everyday enormous volumes of data from a variety of systems and devices (i.e.: patient portals, electronic health records, wearable devices, diagnostic systems, etc). Analyzing current and previously collected data, has the potential to improve clinical diagnosis, reduce treatment costs, avoid preventable diseases and improve the patient care quality in general. However, the data collected exists in different formats (clinical notes, medical images such as CT scans, digital notes, etc) and in many times also unstructured which makes the analysing process a challenging one. ***The purpose of this project is to make you understand how real-life data from a healthcare system is acquired, stored, processed, visualized and used to provide meaningful insights for clinicians.***

2 Dataset

During the project, you will use the *MIMIC-III database* which contains 61,532 intensive care unit stays: 53,432 stays for adult patients and 8,100 for neonatal patients who stayed within the intensive care units at Beth Israel Deaconess Medical Center from June 2001 - October 2012. This database comprises twenty-six different tables with clinical notes, demographics, vital signs, laboratory tests information, etc. To acquire access to the database, *every student* needs to fulfill a course on ethics on CITI: “Data or Specimens Only Research” and submit a request for access. MIMIC-III database is vast in size so, it can not be simply processed on a local computer. During the project, the data is stored centralised in a google BigQuery table. After the 14th of March, students with approval from Physionet will get access to the MIMIC-III BigQuery database.

3 Project Content

Different skills related to the course are developed during the project: (a) hands on experience with ethical approvals; (b) learning how to query big data from cloud based platforms; (c) doing data analysis (d) and ending with machine learning. All aspects described below will contribute to the final evaluation.

3.1 Clinical research

Heart failure (HF) is a syndrome caused due to a heart function disorder and is a life-threatening disease. Depending on its severity, HF patients demands for medical care vary. A high amount of HF patients may require advanced, high-technology, life-saving care that is available only in intensive care units (ICUs).

In this project, you will develop a mortality prediction model of patients in ICU with HF disease where you should use all/partial in-hospital mortality related data among the ICU-admitted HF patients. For this project, you should start by understanding and brainstorming about the clinical relevance of the research topic. Several mortality prediction models exists in the literature using the MIMIC-III dataset, we suggest you to focus in this research study for further inspirations.

3.2 Successfully acquiring/storing/processing data

The purpose of this section is to create a subset of the whole dataset that is relevant for your research question. Not all of the data available in MIMIC-III is needed. So, you will need to query the relevant information using SQL. As already mentioned, since the dataset is too big to be processed locally, you will be working from google BigQuery to process it. There you can view, inspect and extract the relevant data that is needed for further analysis in your project using SQL. During the 4th lab session (March 14th), an introduction will be given to BigQuery. Please keep in mind that during the project, the queries a student can execute per month are limited to 1TB, so think carefully before submitting a query. This reflects the real life costs related to querying big data.

3.3 Data analysis and visualisations

During this task, students are asked to extract relevant knowledge from the raw data they will extract from the full database. This analysis can be a combination of visualisations, hypothesis testing, aggregated features, etc. Data manipulation should be done in google colab, using Python (as done also during the lab sessions).

3.4 Supervised Machine Learning

As an ending point, a classifier used on the MIMIC-III dataset should be developed. Classifiers should be build with the scikit-learn package in google colab. Following points should be taken into account:

- A correct algorithm choice that fits you data and project goal.
- Data pre-processing (normalising, missing data, feature selection,...).
- Properly split into train/validation/test sets.
- Techniques used to improve the performance (boosting, ensembles,...)
- Metrics to report the performance of the algorithm (sensitivity, specificity, confusion matrix,...)

3.5 Report and code

Each group needs to submit the final project work (pdf report and your code) by 15th of May. This report should be maximum 8 pages and it should contain at least the following parts:

- Description of the research question and its medical relevance.
- Explanation of the used data (sub)sets and how this contributes toward achieving you goal.
- Description of the used methods and why they are suited for this purpose.
- The obtained results(visualisations, statistics, graphs, classifier,...)
- Discussion
- Description of work division/teamwork and issues to report

All code (clean and commented) should also be submitted (notebooks or .py files). Additionally, a text-file containing the used queries to extract the datasets from BigQuery should be provided as well.

4 Feedback session

In April 18th, every group will be asked to prepare a small presentation consisting of five slides to show the work that is already done (max 5 minutes per group). By this point, halfway during the project, you should already have preliminary results for the research topic (query and data analysis/visualizations). The slides will be shown to the teaching assistants. The main purpose is to get feedback on the project to further improve it before the final submission. The progress made up to this point, counts for 10 % of the final scoring of the project.

5 Timeline and deadlines:

February 21	Request for access to MIMIC-III on Physio-net
March 14	Tutorial on BigQuerry and kick-of the project
March 21	Outline the group members
April 18	Feedback moment
May 2	Optional Feedback
May 15	Final submission of code and report

6 Additional information:

- Each ***group*** should consist ***of 4 students***. Each student within the group, must work actively and contribute equally to the project.
- The group leader should ***send an email*** to ***Redona Brahimetaj, Joris Wuts, Soha Sadat Mahdi*** outlining the group members by March 21st, 23:59:59 (GMT+1)).
- We will add you to a group on Canvas and all the personal communications and feedback will be provided there. For general questions on the project which might concern also other students, please use the chat of “DAHCC - lab sessions” link.

- For your final submission, you should *submit your code (*.ipynb/*.py)* (with comments), ***and a written PDF report*** (8-10 pages maximum) with ***clear description about what functionalities you have implemented on your project and the contribution of each group member.*** All components of the project (including the report) should be combined into *a single file (*.zip)* named: *[Name1Lastname1_Name2Lastname2_Name3Lastname3_Name4Lastname4.zip]* and submitted via Canvas from your group leader under the *Assignments* section.

- Your project together with the graded practicals contributes ***30% of the final course score***.

- **Important**: We carefully check your code and being caught plagiarising has severe consequences for your evaluation/course.

- Useful links: *Course notes*, *Python*