1. Dataset Download

All required raw data files can be obtained from MIMIC-III database, an account will be required to gain access to the files.

<https://physionet.org/works/MIMICIIIClinicalDatabase/files/>

These are the required data files for our project:

* ADMISSIONS.csv
* CPTEVENTS.csv
* DIAGNOSES\_ICD.csv
* LABEVENTS.csv
* NOTEEVENTS.csv
* PRESCRIPTIONS.csv

Place all downloaded files into the folder ‘rawdata’.

1. Data pre-processing

Next, we will filter all data and events according to patients’ index date.

Run ‘1\_DataPreprocessing.ipynb’ in Jupyter Notebook to filter the datasets. Descriptions are provided in the notebook.

After executing the notebook, 8 csv files will be exported to ‘3\_code/data’.

1. Process unstructured data

Now, we will get unstructured n-gram features out of case\_notes.csv and control\_notes.csv

Run ‘2\_Notes\_to\_Features.ipynb’ in Jupyter Notebook to get a csv file with all the n-gram features for all patients.

After executing the notebook, the n-grams feature csv file will be exported to ‘3\_code/data’.

1. Process structured data, and consolidate features at patient level

We will now use Spark for the next step of data preparation.

The Spark project is contained in the folder ‘3\_code’.

Before compiling and running the project, ensure that there are 9 csv files in the ‘3\_code/data’ folder, all of which were generated from the previous steps. (case\_notes.csv, case\_notes\_no\_text.csv, control\_notes.csv, control\_notes\_no\_text.csv, filter\_diag.csv, filter\_lab.csv, filter\_pres.csv, filter\_proc.csv, formatted\_nf.csv)

Also ensure that ‘features’, ‘mapping.txt’, and ‘patients.svmlight’ are not present in the ‘3\_code’ folder (these are files to be generated in the current step).

Compile the project and run the code in Docker or any preferred environment.

Next, ensure you are in the ‘3\_code’ directory, then run the following code to generate svmlight file:

cat features/\* > patients.svmlight

1. Predictive Modelling

Run ‘4\_PredictiveModeling.ipynb’ in Jupyter Notebook to build Machine Learning models to predict heart failure. Descriptions are provided in the notebook, all results are also generated in the notebook.