Prerequisites

- Apache Spark 1.3.1
- SBT
- (Optional) Docker (Bigbox docker image : http://sunlab.org/teaching/cse6250/spring2018/lab/env-local-docker/)

Downloading MIMIC-III Dataset

1. Complete CITI training "Data or Specimens Only Research on: https://www.citiprogram.org/index.cfm?pageID=154&icat=0&ac=0
2. Request access to MIMIC-III data on: https://mimic.physionet.org/gettingstarted/access/
3. Once approved, download MIMIC-III dataset from https://physionet.org/works/MIMICIIIClinicalDatabase/

Loading Data in PostgreSQL and Pre-processing

1. Download and install latest version of PostgreSQL https://www.postgresql.org/download/

2. Download MIT's MIMIC Code repository from Github. We will be using these SQL scripts to load data in PostgreSQL and to generate SAPS score and Comorbidities. https://github.com/MIT-LCP/mimic-code

3. Open Postgres PSQL command line and run the following commands to create a user.
   CREATE USER mimic;
   ALTER USER mimic superuser;

4. Create a database named mimic with newly created user.
   CREATE DATABASE mimic OWNER mimic;

5. Using command prompt, run the following SQL scripts from mimic-code repository to create tables and load data.
   psql -f postgres_create_tables.sql -U mimic
   psql -f postgres_load_data.sql -U mimic -v mimic_data_dir='<location of mimic csv files>'

6. Next, run the scripts below to add indexes, add constraints, generate views and derive comorbidities. Example, "psql -f <sql script location> -U mimic"
   postgres_add_indexes.sql
   postgres_add_constraints.sql

urine-output-first-day.sql
ventilation-durations.sql
ventilation-first-day.sql
vitals-first-day.sql
gcs-first-day.sql
labs-first-day.sql
sapsii.sql
elixhauser-ahrq-v37-with-drg.sql

7. Export SAPSII and Comorbidities features to a .csv file from database using command line command:
   \copy (SELECT * FROM sapsii) TO '<output location>/SAPSII.csv' DELIMITER ',' CSV HEADER;

   \copy (SELECT * FROM ELIXHAUSER_AHRQ) TO '<output location>/EHCOMORBIDITIES.csv' DELIMITER ',' CSV HEADER;

8. Export NOTEEVENTS from database at the same time removing special characters using SQL command. Run the following commands from command line.
   \copy (SELECT row_id, subject_id, hadm_id, chartdate, charttime, storetime, category, description, cgid, iserror, regexp_replace(text, E'[\n\r]+', ' ', 'g') as text FROM noteevents) TO '<output location>\NOTEEVENTS.csv' DELIMITER ',' CSV HEADER;

9. For this project are using only the following files. Copy these files to the data directory in root of the project.
   EHCOMORBIDITIES.csv
   ICUSTAYS.csv
   NOTEEVENTS.csv
   PATIENTS.csv
   SAPSII.csv

Running Project

1. From command prompt, navigate to root of the project directory
2. Execute command "sbt compile run"
3. Output result will be printed on console

Note, the default settings will run prediction models for In ICU mortality. To predict for 30 Days or 1 Year after discharge mortality change the value of variable **labels** on line 92 in file Main.scala. The possible values are **labelsInIcu**, **labelsIn30Days** and **labelsIn1Year**.