SPH 6004: Advanced Statistical Learning – Assignment 1 (Individual Assessment 40%)

For this assignment, all students will be given a pre-extracted dataset from MIMIC-IV, a real-world EMR dataset, and all students will be tasked to solve the same prediction problem. The objective of this assignment is test whether the students can:

1) Develop and implement predictive model(s) on a static structure dataset
2) Develop and implement feature selection methods to enhance prediction model(s) 3) Analyze and explain the differences in performance from different models

Note: NUS takes plagiarism very seriously. One should NOT copy from seniors' previous years' assignments nor his/her own assignments for other subjects.

Regarding the use of AI tools, such as ChatGPT, Gemini, or their equivalents, students are ONLY permitted to employ these tools as supportive aids solely for the purpose of proofreading their final reports. It is expressly prohibited for students to utilize AI tools to generate project code or create content for their reports. In the event of any suspected violations, individual interviews with the students will be conducted, and if any fraudulent activity is substantiated, appropriate disciplinary actions will be taken against the student involved.

Both assignments 1 and 2 will be based on the MIMIC-IV (**M**edical **I**nformation **M**art for **I**ntensive **C**are IV) database. The MIMIC database is a large, freely-available database comprising of de-identified health-related data of over forty thousand patients who stayed in critical care units at the Beth Israel Deaconess Medical Center. The database includes information such as demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital). **MIMIC-IV version 2.2**, which is the latest version of the dataset just released this year.

For this assignment, you will work on an extracted dataset of patients' snapshot of health status at admission that comprises of many features/variables.

You will develop a strategy to select the right combination of features to build a predictor that estimates the patients' risk of **kidney failure** in the Intensive Care Unit (ICU). You will be assessed on the performance of the following tasks:

Individual Assessment (40% of the total marks)

1. 1) Register for the data access to MIMIC (2% of the total marks).

   Deadline: 28th Jan 2024
   Obtain up to 2% when student:

- Successfully registers for data access
Note: Indicate Dr. Feng as your supervisor for the application.

2. 2) Propose and implement a strategy to select the right features for the predictive model (15% of the total marks)
   Deadline: 7th Mar 2024
   Obtain up to 15% when student:
   - - Clearly explains the rationale of the proposed strategy (5%)
   - - Clearly explains how the proposed strategy works (5%)
   - - Demonstrate experimentally how the proposed strategy manages to minimise the feature dimensions of the pre-extracted MIMIC data (5%)

3. 3) With the selected features, implement the predictive models taught in first half of the course and compare their performances in estimating the patients' chance of kidney failure in the ICU (15% of total marks)
   Deadline: 7th Mar 2024

   Obtain up to 15% when student:

   - Implements all predictive models correctly (5%)
   - Evaluates and compares the performance of the models (5%) - Clearly explains the better performance of some models (5%)

4. 4) Summarise results and findings from 2) and 3) in a written report, no more than 8 pages and upload the report to Canvas Student Submission Folder through the Turnitin platform. All the relevant codes should be uploaded to github.com, and github URL needs to be included in the report as well. (8% of total marks) **Deadline: 7th Mar 2024**
   - - **Obtain up to 8% when student:**
   - - Clearly presents and explains the problem statement (2%)
   - - Clearly explains the proposed ideas and experience setups (2%)
   - - Effectively summarises and clearly illustrates the key findings design with appropriate use of graphics and tables (2%)
   - - Compiles into a formal report (2%)