**Exercise 3-2: Using supervised methods to predict health outcomes**

In this exercise, you will implement a full prediction pipeline using a sample dataset to predict readmission for myocardial infarction. No .RMD file is provided.

Description of Dataset

The dataset (mi.data.final) consists of  an ID variable, 14 features describing clinical tests and comorbidities, and an indicator for whether the individual was readmitted to the hospital within 30 days of discharge.

The variables are as follows:

ID: identifier

Age: age at initial MI (years)

Sex: Reported by patient 0-Male, 1-Female, 2-Non-binary/Other

sodium: serum sodium (mmol/L)

ALT: liver enzymes (IU/L)

WBC: white blood cell count (billions/L)

ESR: erythrocyte sedimentation rate

SBP: systolic blood pressure at intake (mmHg)

DBP: diastolic blood pressure at intake (mmHg)

Pulm.adema: Pulmonary adema (1=Yes, 0=No)

FC: functional class of angina pectoris in the last year

* 1: there is no angina pectoris   
  2: I FC   
  3: II FC   
  4: III FC   
  5: IV FC

Arrythmia: Presence of arrythmia (1=Yes, 0=No)

Diab: Presence of diabetes (1=Yes, 0=No)

Obesity: Presence of obesity (1=Yes, 0=No)

Asthma: Presence of asthma (1=Yes, 0=No)

readmission: Readmitted to hospital within 30 days (1=Yes, 0=No)

Instructions

For this exercise, your pipeline should include tasks for data preparation, any partitioning or resampling you deem necessary, any tuning of hyperparameters, and explicit evaluation metrics you will examine in order to choose your optimal algorithm. You should choose 2 of the supervised algorithms covered and one should be an ensemble algorithm.

Even though a starting .RMD file is not provided, remember you have all of the code from demonstration files within Modules 2 and 3. Please feel free to post questions to me or each other on slack if you get stuck anywhere. Part of this exercise is working through what you would do in a “real” analysis.