Final Project: Classification of Climate Model Simulation Crashes

- **Instructions:** While discussion with classmates are permitted and encouraged, feel free to work on the project in a group or independently (using R or Python). For a beginner R-programmer, a sample R-script is provided in D2L. Please practice the sample code before starting the project.

- We consider the Climate Model Simulation Crashes data available at UCI machine learning repository:
http://archive.ics.uci.edu/ml/datasets/Climate+Model+Simulation+Crashes

  This dataset contains records of simulation crashes encountered during climate model uncertainty quantification (UQ) ensembles. Ensemble members were constructed using a Latin hypercube method in LLNL's UQ Pipeline software system to sample the uncertainties of 18 model parameters within the Parallel Ocean Program (POP2) component of the Community Climate System Model (CCSM4). Three separate Latin hypercube ensembles were conducted, each containing 180 ensemble members. We shall assume that all these ensemble members are independent of each other. Forty-six out of the 540 simulations failed for numerical reasons at combinations of parameter values. The goal is to use classification to predict simulation outcomes (variable outcome with values 'fail' or 'succeed') from input parameter values (columns 3-20), and to use sensitivity analysis and feature selection to determine the causes of simulation crashes. Columns 3-20 contain numerical values of 18 climate model parameters scaled in the interval [0, 1].

  Follow the steps below to conduct your analysis

  1. *(Data preparation)* Bring in the data D into R. Inspect if there is any missing values and, if so, handle them with imputation.
  2. *(Exploratory Data Analysis)* Explore the data with EDA and present at least THREE interesting findings.
  3. *(Data partition)* Randomly split the data D into the training set D1 and the test set D2 with a ratio of approximately 2:1 in sample size. Use **set.seed()** to fix the random seed so that the results are easily reproducible.
  4. In the steps to follow, we will train several classifiers with D1 and then apply each trained model on D2 to predict whether a simulation fails. For each approach, obtain the misclassification rate (with default cutoff point 0.50), the ROC curve and the corresponding AUC based on the prediction on D2. **Compare the performance of all these classifiers and summarize the results.**
     (a) *(Logistic Regression)* Fit a regularized logistic regression model as one baseline classifier for comparison. You may use either LASSO or SCAD or any other penalty

function of your choice. Explain how the optimal tuning parameter is determined. Present the final logistic model and interpret the results.

(b) *(Random Forest)* Fit random forests as another baseline for comparison. Also, obtain the variable importance ranking from RF.

(c) *(Artificial Neural Network)* Fit at least three different artificial neural network (ANN) models, e.g., with different numbers of layers and different number of units.

- **Timeline:**
  - A final report is due at 11:59 pm on Wednesday, May 05, 2021. All members of a group do not need to individually submit their project. Please submit just one report for the group but ensure that the names of all group members are listed on the report. The report (word/pdf file) should contain the outputs (image, table, etc.), **and interpretation of your analysis**. Please choose any of the following to submit:

    1. R script/Python script and a doc/pdf report, or
    2. A pdf file generated by R-markdown that includes both your analysis and code, or
    3. A notebook file in Jupyter with R/Python that includes both your analysis and code.