Instructions

- This homework assignment is worth 54 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: September 23, 2022 by 11:59 pm.**

Exercise 1

(4 points) Given 1000 records in a dataset, 1000 models are trained with 999 records as part of the training sample and the remaining 1 sample for testing, and the error rate is averaged out, this validation technique is called

(a) validation set

(b) k-fold cross validation

(c) LOOCV

(d) Bootstrapping

(e) None of the above

Exercise 2

(4 points) In k-fold cross validation technique, the value of k being small could lead to which of the following in relation to the error rate

(a) low bias and low variance

(b) low bias and high variance

(c) high bias and low variance

(d) high bias and high variance

(e) All of the above

Exercise 3

(4 points) In k-fold cross validation technique, the value of k being large could lead to which of the following in relation to the error rate

(a) low bias and low variance

(b) low bias and high variance

(c) high bias and low variance

(d) high bias and high variance

(e) All of the above

Exercise 4

(6 points) Explain what regularization is and why it is useful.

Exercise 5

Consider the `framingham.csv` data file. The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients? information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

- Demographic:
  - Sex: male or female (Nominal)
  - Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- Behavioral
  - Current Smoker: whether or not the patient is a current smoker (Nominal)
  - Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
- Medical (history)
  - BP Meds: whether or not the patient was on blood pressure medication (Nominal)
  - Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
  - Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
  - Diabetes: whether or not the patient had diabetes (Nominal)
- Medical (current)
  - Tot Chol: total cholesterol level (Continuous)
  - Sys BP: systolic blood pressure (Continuous)
  - Dia BP: diastolic blood pressure (Continuous)

  - BMI: Body Mass Index (Continuous)
  - Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
  - Glucose: glucose level (Continuous)

- Predict variable (desired target)
  - 10 year risk of coronary heart disease CHD (binary: "1", means "Yes", "0" means "No")

In **Python**, answer the following:

(a) (4 points) Using the pandas library, read the csv data file and create a data-frame called `heart`.

(b) (3 points) Remove observations with missing values.

(c) (25 points) Perform a 5-folds cross validation with the goal of measuring the performance, in terms of F1-score, of two competing models:

- Using `age`, `currentSmoker`, `totChol`, `sysBP`, `diaBP`, `BMI`, `heartRate`, and `glucose` as the predictor variables, and `TenYearCHD` as the target variable build a logistic regression model under the 5-folds cross validation framework. Compute and store the F1-score for each iteration.
- Using `age`, `currentSmoker`, `totChol`, `BMI`, `heartRate`, and `glucose` as the predictor variables, and `TenYearCHD` as the target variable build a logistic regression model under the 5-folds cross validation framework. Compute and store the F1-score for each iteration.

Use 25% as threshold to change the likelihoods to labels. Make sure to scale the input variables of both models to 0-1 range (see `MinMaxScaler`) before you run the 5-fold cross validation framework. Also, you can use the `f1_score` function to compute the F1-score.

(d) (4 points) Report the average F1-score of each of the models. What model would you use to predict `TenYearCHD`? Explain.