Project Definition: (Due date: Dec 1st, 2022)

As we discussed in our class, you will be working on a project which has two steps:

1. Working on gas permeability data & apply linear regression
2. Working on a prediction model on polymer_solvent solubility data & apply decision tree model

Note: Please check the project folder on GitHub. (We already added and cleaned the data for the second part of the project.)

How to report the project:

You are expected to turn in a written report (each student should turn in a copy, but the ones for each student in a group should be identical). Your report should include **1 page of written text** that addresses the topics below. Requested figures will make your report longer. Your report should include the following:

- Your group members' names!
- Part one:
    1. Picture of your plots
    2. Discuss the results based on the accuracies and which one is the most effective feature and why?
- Part two:
    1. Report of accuracy metrics of training & test sets
    2. Picture of confusion matrix for the validation set with tick marks labeled.
    3. Discuss the confusion matrix & answer the following questions:
        - What would happen if we changed the test_size from 0.3 to 0.5? how does that affect our result?
        - How can you define the confusion matrix results? What does it tell you about the results of the prediction model?

- Additionally, please submit a .zip file containing (1) the Jupyter notebook along with .pdf file of your report.

Part One:

The dataset includes 8 features (e.g., chemical, mechanical properties of polymers, and transport-related properties) relevant to the 78 polymers.

- Glass transition temperatures
- Melting points
- Decomposition temperatures
- Densities
- Cohesive energy densities
- Fractional free volume
- Tensile strengths
- Young's modulus

The performance of a membrane is an extremely complicated and comprehensive result from many chemical and physical properties as well as properties related to the interactions between the polymer and the guest molecules. However, in this problem, we want to find a feature that is linearly correlated to the selectivity and determine which feature has the strongest linear correlation.

For accomplishing this task, you need to do the following steps:

1. Use the Pandas package and read the csv file "Data on Permeability.csv" from the data folder.
2. Plot the selectivity of O2/N2 vs. 8 features (you can plot the 8 plots at once using the for loop & subplot from the first note)
3. Use Scikit learn package or other option that we discussed to generate the best fit
4. Report the accuracy of each fit using R-squared method
5. Discuss your results

Part Two:

Polymers are widely used in industry applications from membranes to coatings and drug delivery, and in all these applications, selecting the right solvents is critical. The data set will utilize 28 unique features based on experimental conditions, polymer properties, and solvent properties. The dataset contains 5 classes of solubility, and the balance is skewed toward 3 of these classes (identified as 0, 1, and 2) based on the number of member data in each class.

## prediction label meaning:

- 0 - insoluble
- 1 - partially soluble
- 2 - soluble
- 3 - solvent evaporated
- -1 - solvent freeze

You are expected to implement the following steps on the data set in the order they are listed, after you read the excel file using the Pandas package:

1. Using the StandardScaler, normalize the data
2. Using the Hold-out method: divide the dataset to two sets of data & validation set with test_size=0.3, random_state=42.
3. Using the k-fold method: divide the data to training & test set. (k =5)
4. Implement the Decision Tree for prediction on training and test sets & calculate the R-Squared for each set

5. Report the accuracy of the validation set and how the model is performing using the confusion matrix (label the confusion matrix using the prediction labels showed above: insoluble, etc.)

6. Discuss your results, answering following questions:

    6.1. What would happen if we changed the test_size from 0.3 to 0.5. how does that affect our result?

    6.2. How can you define on confusion matrix results. What does it tell you about the results of the prediction model?

Data Science Project Rubric:

(50 pts) Code Functionality:

50 points for following all the steps and well-organized code
40 points for following all the steps and code are not functioning for all the parts! (2 steps are not functioning)
25 points for implementing only one part of the project
15 points for implementing only one or two steps of the project (reading the datasets are excluded)
10 Not functioning code! But tried to follow at least one step of the project

(10 pts) Group evaluation of performance (I will assume all equally participate unless otherwise noted or communicated to me via email)

10 points for approximately equal share
5 points for some participation but significantly unequal
0 points for no participation

(20) Discussion part of the project:

20 points for clear discussion on the results for both part of the project and professionally formatted by discussing the questions.
15 points for mostly clear reasoning and trying to discuss the results and answering the questions.
10 points for some unclear discussion & reasoning or only results for one part of the questions!
5 points for very unclear discussion of the results!

(20 pts) Report

20 points for clear storytelling and writing, appropriate references and all figures requested being added, and professionally formatted
10 points for mostly clear storytelling and writing, mostly appropriate references and figures, mostly professionally formatted
5 points for some unclear storytelling and writing, missing some references and figures, some unprofessional formatting or only one part of the project included!
2 points for very unclear storytelling and writing, many references and figures missing, unprofessional formatting