In this problem you are given data from the SWELL Knowledge Work Dataset containing measurements from different working conditions designed to elicit stress. You will run machine learning experiments to classify between: (i) the various working conditions and (i) the levels of stress.

1. *SWELL_data.csv*: Contains physiological (columns 5-7), user-computer interaction (columns 8-24) data, and facial action unit measures (columns 25-43), as well as labels (columns 3-4) from 3 participants

2. *Notes.csv*: Contains explanations on the features and data

(a) Load data: Load the data and find the number of participants, total number of samples, and number of features for each modality.

(b) Classification between working conditions: The data file contains four different conditions: 1 → relaxation, 2 → neutral, 3 → time pressure, 4 → interruptions. Merge 1 and 2 (potentially low arousal conditions), as well as 3 and 4 (potentially high arousal conditions). **(i)** Count the number of samples per condition. **(ii)** Use a binary decision tree to classify between low and high arousal conditions with a **leave-one-subject-out** cross-validation framework based only on the **physiological** data. Compute the weighted and unweighted recall. **(iii)** Try various tree depths (e.g., 1,2,3,4). Which works best? **Hint:** Use the *tree.DecisionTreeClassifier* and *recall_score* functions from *sklearn*.

(c) Stress level regression: The data further contain participants' self-assessment scores of stress levels. **(i)** Use a regression tree with a **leave-one-subject-out** cross-validation framework to predict stress levels based on the **physiological** data. Compute the Pearson's correlation function between the actual and predicted stress values. **(ii)** Experiment with various tree depths. **Hint:** Use the *tree.DecisionTreeRegressor* from *sklearn* and *pearsonr* from *scipy*.

(d) Stress level classification: **(i)** Plot the histogram of stress level values using 3 bins. What do you observe? Can you bin the stress labels into 3 bins? **(ii)** Use a binary decision tree and a K-Nearest Neighbor classifier with a **leave-one-subject-out** cross-validation framework to classify between three different stress levels based on the **physiological** data. Compute the weighted and unweighted recall. **(iii)** Experiment with various tree depths and number of neighbors. **Hint:** Use the *tree.DecisionTreeClassifier*, *KNeighborsClassifier*, and *recall_score* functions from *sklearn*.

(e) Additional modalities: Perform stress classification using the features related to **user-computer interaction** and **facial action unit** features. What do you observe? How about if different modalities are combined together?