**Unsupervised Learning Assignment**

**General Guidelines for Assignment for Unsupervised Learning Component**

You must submit your solutions as a Jupyter Notebook for this assignment in both .html and .ipynb formats, as well as a .csv file that will be called electrophysiology\_data\_kmeans.csv.   
The HTML file will be reviewed for readability, the ipynb file will be used to verify solutions, and the CSV file will be checked for accuracy.

The following instructions outline the key components and guidelines for preparing the notebook:

**Notebook Structure and Content**

1. **Headers and Subheaders**:
   1. Organize the notebook into clearly defined sections using headers and subheaders.
2. **Graphs and Visualizations**:
   1. Include well-labeled graphs with appropriate titles, x and y-axis labels, and scales.
   2. Specify the colors used in each graph to ensure clarity and consistency.
   3. Each graph also includes a subsequent markdown cell to explain the graph's purpose and relevance to the analysis.
3. **Work individually or in pairs**:

We will contact each student to explain the work done individually or in pairs with questions to verify that the student understands the solution submitted in depth. We will only accept assignments submitted jointly by up to two students or submitted individually.

1. **Grading:**

Most of the grading in this section will be on evidence of applied learning in the course. Explain the reasoning behind the methods selected thoughts. Provide visual evidence to confirm the explanations. Achieving a high level of success on performance measures may yield additional evidence of correct thought processes; however, success in performance measures is less significant in grading. Using tools you know or found online that are beyond what was taught in this course is acceptable; however, if you use such tools, please explain them at a level that a freshman majoring in engineering or hard sciences could follow, as we want to follow what you are doing as well. Also, we will ask you verbally about the components of your solutions, so be prepared to demonstrate command of the tools you used.

**The Data**

1. The unsupervised learning component of the homework assignment involves analyzing a dataset structured for behavioral analysis of subject rats in dyadic social experiments. The dataset consists of 8 columns representing some of the features extracted from electrophysiological data refined for this assignment by performing data imputation and cleaning. The features are extracted from electrophysiological measurements collected from brain areas important in mediating social behavior.
2. Link to data:   
   [https://drive.google.com/file/d/1c-VLLYdzG213Pyn0LcYuDNL-73SY1ZL6/view?usp=sharing](https://drive.google.com/file/d/1c-VLLYdzG213Pyn0LcYuDNL-73SY1ZL6/view?usp=sharing%20)

**Assignment Tasks**

1. **Dimensionality reduction (50 points)**:
   * Reduce the dimensions of the data to 2 dimensions using t-SNE, UMAP, and PCA
   * Perform the reduction using the following parameters and visualize each reduction:
     1. t\_SNE: perplexity level: [5, 50] (2 plots)
     2. UMAP: n\_neighbors: [2, 10], min\_dist: [0.1, 1] (4 plots)
     3. PCA (default) (1 plot)
   * Visualize the data for each reduction using a scatter plot
   * Explain how the different parameters affect the reduction
2. **Clustering (50 points):**
   * Cluster the data using KMeans, AgglomerativeClustering, and one more method of your choosing
   * For Kmeans clustering, evaluate the optimal number of clusters using the Elbo method
     1. Fit 10 Kmean models with the number of clusters ranging from 1-10
     2. Extract the inertia of each fitter model
     3. Plot the Elbo plot
   * Add a column to the data frame named "kmeans\_labeles" and assign the labels extracted using the K-means method.