**Subject: Midterm Assignment - Machine Learning Project**

Dear Students,

As part of the midterm evaluation, you are required to complete a machine learning project. This project will involve applying the concepts we have learned in the course so far, including clustering, binary classification, and model selection.

**Assignment Details:**

The midterm assignment needs you to complete a clustering and a binary classification task.

**1 Clustering Task:**

You have been provided with two single-cell datasets:

[SpinalCord\_clustering\_rawCounts.csv]: Raw counts matrix of single-cell gene expression.

[SpinalCord\_clustering\_CLR\_Transform.csv]: CLR transformed matrix of single-cell gene expression.

The single-cell dataset of the spinal cord contains 1116 rows of cells and 3000 columns of variable genes plus one column of the original label (first column). Central Log Ratio (CLR) transformation of the expression of each gene in a cell can be described as the logarithmic transformation of the ratio of gene expression to the geometric mean of all gene expressions, so the mean of gene expression of the cell will be centralized to 0. You need to perform a clustering analysis on the Spinal Cord single-cell dataset before and after CLR centralization.

The report requirements or questions that you have to answer in the clustering part:

Describe the overall methodology of your clustering. (5 points)

Discuss the differences in the clustering results before and after centralization. (15 points)

What insights did you gain from these analyses? (10 points)

**2 Binary Classification Task:**

You have been provided with a training dataset and a testing dataset:

[L1\_train\_dataset.csv]: The training dataset of the transposon L1 insertions.

[L1\_316\_BG\_clone19\_Test\_dataset.csv]: The testing dataset of the transposon L1 insertions should be used in the finalized model.

The L1 dataset contains 38 columns of features plus one column of target. Training dataset has 15000 true cases (column ‘target’ labeled as 1) and 100000 false cases (column ‘target’ labeled as 0). Testing dataset has 1962 true cases and 25565 false cases. So, you are going to predict the column "target" by using the rest of the 38 columns.

1. train at least two different binary classification models on the training dataset
2. Evaluate the performance of these models and select the best one using appropriate metrics. (Diagnostic, Hyperparameter......)
3. Use the selected model to make predictions on the test dataset

The report requirements or questions that you have to answer in the binary classification part:

Describe the overall methodology. (5 points)

Which classification models did you use and why? (10 points)

How did you handle missing values in the dataset? (10 points)

What other preprocessing steps did you take? (5 points)

How did you evaluate their performance? (10 points)

Which model performed the best, and why do you think this is the case? (5 points)

How well did your selected model perform on the test dataset? (10 points)

Were there any issues or challenges? (5 points)

How might the model be improved? (10 points)

**Submission Guidelines:**

Please submit your code in a Jupyter notebook (".ipynb" file). Your notebook should include comments explaining your code and decisions throughout the project.

In addition to the code, you are required to submit a report (Word or PDF) discussing your findings. Your report show provides and describes any diagnostic plots or tables used in your analysis.

You should write a 3-6 page report, and the deadline for this assignment is **4th Nov**. This assignment will test your understanding of the course material and your ability to apply these concepts to real-world data. If there are any questions, please reach out to us.