Project Description:

Due date: 5/1/2025

For your group project, you are requested to do a complete process of machine learning, which includes Feature extraction, classification, and analysis.

You can access the data that we are working with on Canvas -> Modules => Subcellular Data

The project is as follows:

1: You need to get the protein sequences (as it was discussed in assignment 3) and extract relevant features that are representative of the protein sequences. You can also use the features that are already available (e.g., Occurrence and Composition).

2: You will need to prepare the data (properly put them together with labels) and use different machine learning methods on them (e.g., K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes, Artificial Neural Network (ANN), Random Forest, Bagging). As the data suggest, the problem is hand is a multi-class classification task.

Note: The problem is protein subcellular localization. The proteins can function in different locations in the cell. Given a protein, it is important to identify its functioning location. For this problem, we are working with Gram-positive bacterial proteins, which function in 4 locations in the cell. So, the problem is to get the training data, extract relevant features, and build a model to predict the functioning location of a given protein.

3: You then need to analyze and interpret your output using different approaches such as:

- 3.1: Using independent test set
- 3.2: k-fold cross validation
- 3.3: check different accuracy measurement
- 3.4: Discuss your results and interpret the output

Note: You can do steps 2 and 3 in Python programming language.

As the output: You will prepare a report (in Word or text document) that explains what you did in each and every step in detail and present your results, analysis, and discussion (no more than 6 pages). You will also send your codes with proper comments along with your report.

Please do not hesitate to contact me if you have any questions or concerns.

Good Luck!