Welcome to Your Fourth Step of the Challenge!

Total Points: 100

In this challenge, we will generate the proper pipeline for training and evaluation of models, do tuning, evaluate on unforeseen protein data, and submission to online competition. The dataset for constructing your models is provided as `metadata_org_w_features.csv` and the data for the competition is `testing_data_w_features.csv.` Note that the testing data does not give you the Protein Class as you will need to evaluate this from the model you constructed. The only way to see how well your models performed is that you will need to upload your results and log in with your JHU email to:

- Website: https://jhucompetition.streamlit.app/
- User: your jhu email
- Password: EN.605.656.8VL.SP25

This challenge is divided into two parts that should be addressed in the Notebook:

1. Coding: Focused on addressing the challenge of Machin Learning. (60 points)
2. Questions: Designed to explore and address the subcategories within the challenge. (40 points)

Data Overview:

The dataset contains the following columns:

1. Entry and Protein Class
2. Sequence and Sequence Length
3. Selected PDB
4. Amino Acid Frequencies
5. All possible dipeptide frequencies
6. Reduced Amino Acid Alphabet Frequencies
7. N-Gram Profiles of Reduced Amino Acid Alphabet
8. Protein Properties

Questions to Explore

1. Constructing the Dataset

- Why is it essential to ensure that the features used for training a machine learning model match the features present in the data used for evaluation? Look into approaches and write a paragraph with citations of potential approaches for selecting and constructing your data. (6 points)
- Code (10 point)

2. Creating a Model

- Why is it essential to analyze the performance of different algorithms, and what insights can be drawn from comparing their results? What are the key criteria for selecting the most appropriate algorithm for a given task? Explore methodologies and summarize potential strategies. Look into approaches and write a paragraph with citations of potential approaches for selecting and constructing your data. (6 points)
- Coding section (10 points)

3. Training and Tuning Model

- How do you choose the type of data that a model would need, and what considerations should guide the selection of data for model training to ensure optimal performance? How does data quality influence outcomes, and what preprocessing techniques, such as normalization and scaling, should

be employed for specific machine learning models? Additionally, how can fine-tuning be leveraged to refine a model's performance and adaptability for nuanced tasks? Explore their importance and application. Look into approaches and write a paragraph with citations of potential approaches. (12 points)
- Coding section (20 points)

4. Evaluate your Model and Predict for the Unforeseen Data

- What are the key evaluation metrics for assessing model performance, and how do these metrics help in understanding a model's suitability for handling unseen data? How can techniques like cross-validation, bootstrapping, or holdout sets provide insights into a model's ability to generalize to new data? What strategies can be employed to ensure that a model remains robust and accurate when exposed to unforeseen data in real-world scenarios? Look into approaches and write a paragraph with citations of potential approaches. (10 points)
- Coding section (10 points)

6. Save, Create your Model Output, and Competition Time

- Why is it important to save trained models and their parameters, and what strategies can ensure the reproducibility of model outputs? What techniques can be employed to optimize a model's output for accuracy and relevance, especially when addressing real-world applications? How can benchmarking against competing models or industry standards drive innovation and improve a model's performance? Look into approaches and write a paragraph with citations of potential approaches. (6 points)
- Coding section (10 points)