Welcome to Your Third Step of the Challenge!

Total Points: 100

In this challenge, we will generate the proper pipeline for training and evaluation of models, do cross validation and make ensembles for protein classification. The dataset provided is `metadata_org_w_features.csv`.

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

- If you do not know how to construct the data what can you do? 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 relevant to evaluate multiple models and what information can be attained from evaluating multiple models? How do you go about choosing models? Look into approaches and write a paragraph with citations of potential approaches. (6 points)
- Coding section (10 points)

3. Training a Model

- How do you choose the type of data that a model would need? When and where should we leverage normalizations approaches? Why would you need to consider different normalization approaches for different model? What factors do you need to think about as you train the model? Look into approaches and write a paragraph with citations of potential approaches. (6 points)
- Coding section (10 points)

4. Evaluate your model

- List several approaches for evaluating models and describe what information each approach would give? What information can you get when you evaluate different models? Look into approaches and write a paragraph with citations of potential approaches. (6 points)
- Coding section (10 points)

5. Cross Validation

- What is the point? List pros and cons of such approaches? Look into approaches and write a paragraph with citations of potential approaches. (6 points)
- Coding section (10 points)

6. Ensemble Section

- Why are smaller weak leaners potentially a good idea? How do you determine the optimal way to combine predictions from different models in an ensemble, and what challenges arise when trying to balance diversity and accuracy among the models? Look into approaches and write a paragraph with citations of potential approaches. (10 points)
- Coding section (10 points)