Project Title: Novozymes Enzyme Stability Prediction

Project Goal:

Enzymes are proteins that act as catalysts in the chemical reactions of living organisms. In industry, enzymes replace chemicals and accelerate production processes. They help our customers make more from less, while saving energy and generating less waste. However, many enzymes are marginally stable, limiting their performance under harsh application conditions. Instability also decreases the amount of protein that can be produced by the cell. The goal of this project is to predict the thermostability of enzyme variants based on their amino acid sequence. Understanding and accurately predict protein stability is a fundamental problem in biotechnology. Its applications include enzyme engineering for addressing the world’s challenges in sustainability, carbon neutrality and more. Improvements to enzyme stability could lower costs and increase the speed scientists can iterate on concepts. Therefore, the development of efficient computational approaches to predict protein stability carries enormous technical and scientific interest.

Dataset:

In this project, we will be using a dataset provided by Novozymes through their Kaggle competition: [Novozymes Enzyme Stability Prediction | Kaggle](https://www.kaggle.com/competitions/novozymes-enzyme-stability-prediction/overview/description). This dataset will provide the experimentally measured thermostability (melting temperature) data, natural enzyme sequences, as well as engineered sequences with single or multiple mutations upon the natural sequences. In this project, we will develop a model to predict/rank the thermostability of enzyme variants based on experimental melting temperature data

Stakeholders:

* Biotech Companies
  + Biologics
  + Industrial Enzyme production

Key Performance Indicators:

* Identify three predictive frameworks to explore project
* Perform exploratory data analysis with each framework and prepare a summary report on the performance of the frameworks
* Select the best operating framework, after initial optimization, utilizing Normalized Root Mean Squared Error and Spearman Correlation Coefficient
* Optimize the selected framework through cross-validation and prepare a final project report with visualizations