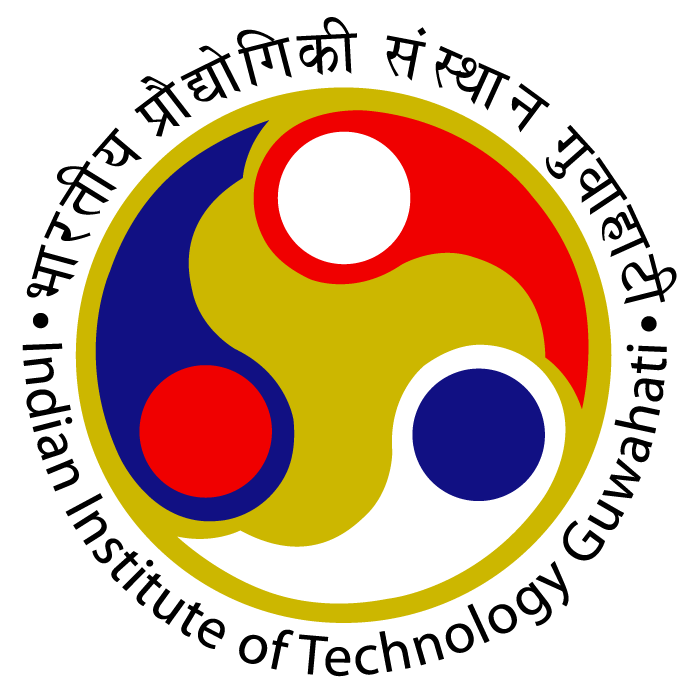
Enzyme Stability Prediction

**Prakash Kumar**

**210107103**

**Submission Date: April 26, 2024**



**Final Project submission**

**Course Name : Applications of Al and ML in chemical engineering**

**Course Code: CL653**

Contents

[1 Executive Summary **Error! Bookmark not defined.**](#_Toc163320085)

[2 Introduction **Error! Bookmark not defined.**](#_Toc163320086)

[3 Methodology **4-6**](#_Toc163320087)**.**

[4 Implementation Plan **Error! Bookmark not defined.**](#_Toc163320088)

[5 Testing and Deployment **Error! Bookmark not defined.**](#_Toc163320089)

[6 Results and Discussion **Error! Bookmark not defined.**](#_Toc163320090)

[7 Conclusion and Future Work **Error! Bookmark not defined.**](#_Toc163320091)

[8 References **Error! Bookmark not defined.**](#_Toc163320092)

[9 Appendices **Error! Bookmark not defined.**](#_Toc163320093)

[10 Auxiliaries **Error! Bookmark not defined.**](#_Toc163320094)

1. **Executive Summary:**

**Overview:** Develop computational model to predict thermostability of enzyme variants from sequence/structure data to enable broader industrial use of enzymes as green biocatalysts.

**Approach:** Multidisciplinary combining biochemistry, molecular biology, and computational science. Train ML models (SVM, Random Forest, XGBoost, deep learning) on melting temperature data.

**Data:** Novozymes Kaggle dataset with protein sequences, pH, data sources.

**Methodology:** Data preprocessing, EDA, model training/tuning, evaluation (Spearman correlation), cross-validation.

**Deployment:** Integrate with LIMS/production systems via APIs. User interfaces. Cloud scaling. GPU acceleration.

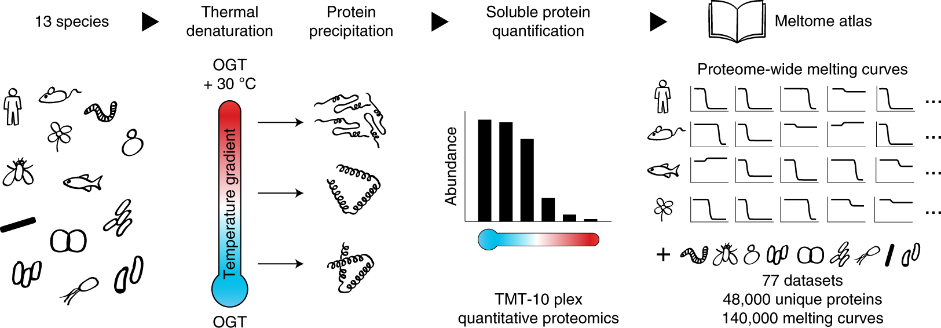
**Applications:** Biofuels, pharmaceuticals, food production - improve efficiency, reduce costs/environmental impact.

**Impact:** Enable sustainable industrial enzyme catalysis. Accelerate enzyme engineering innovation.

1. **Introduction:**

In the domain of chemical engineering, enzymes play a pivotal role as ***biocatalysts***, driving reactions that are essential for the production of a wide array of chemicals, materials, and pharmaceuticals. Their unique ability to operate under mild conditions—lower temperatures, neutral pH, and atmospheric pressure—contrasts sharply with traditional chemical catalysts that often require harsh environmental conditions. This distinction not only makes enzymes a cornerstone of ***green chemistry and sustainable industrial processes*** but also presents a compelling case for their integration into chemical engineering workflows. However, the application of enzymes in chemical engineering is ***frequently challenged by their inherent stability, particularly under the rigorous conditions*** of industrial processes. Thermostability, or the ability of enzymes to maintain functionality at elevated temperatures, is a critical attribute that can dictate the feasibility of enzymatic processes in chemical engineering.

1. **Methodology:**
   1. **Flow chart for data collection and model development**



A diagram of a program

Description generated with high confidence

ML Regressor Models :

SVM, Random

Forest, XGBoost, Decision Tree, CatBoost

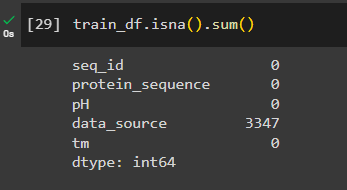
* 1. **Data Preprocessing:**
     1. **Data Source:** Novozymes Enzyme Stability Prediction on Kaggle.
     2. **Data Cleaning:** Handling of missing data points, which may involve imputation techniques such as mean substitution or deletion of records with missing values.
     3. **Feature Engineering:** Explore potential interactions or combinations of features that may provide additional predictive power.
     4. **Exploratory Data Analysis (EDA):** Conduct exploratory data analysis to gain insights into the relationships between features and the target variable. Visualize distributions, correlations, and patterns in the data using techniques like histograms, scatter plots, or correlation matrices. Identify any potential trends or outliers that may impact model training and interpretation.
  2. **Model Development:**

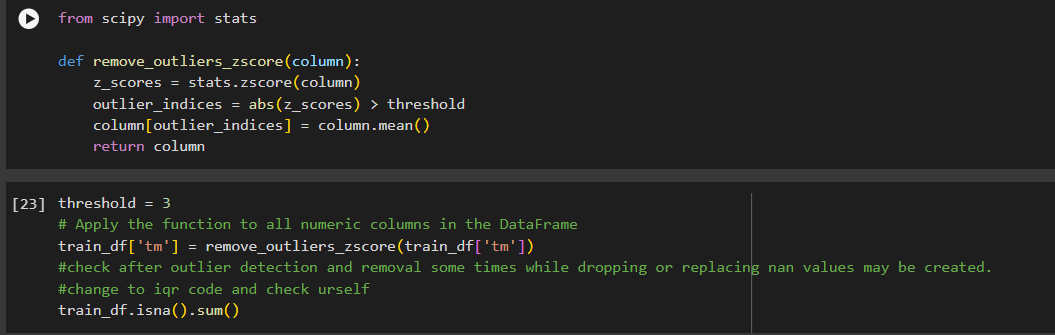
For the project of predicting enzyme thermostability, the AI/ML models under consideration include: ***Support Vector Regressor (SVR), Decision Trees, Random Forest***, ***Gradient Boosting Machine (XGBoost) and Catboost .***

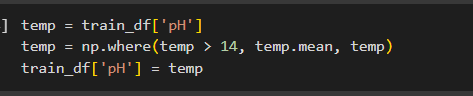
The rationale for choosing these models lies in their ***versatility*, *interpretability***, and ability to handle both numerical and categorical features present in the dataset. Ensemble methods like Random Forest, XGBoost and Catboost is particularly well-suited for handling complex interactions between features.

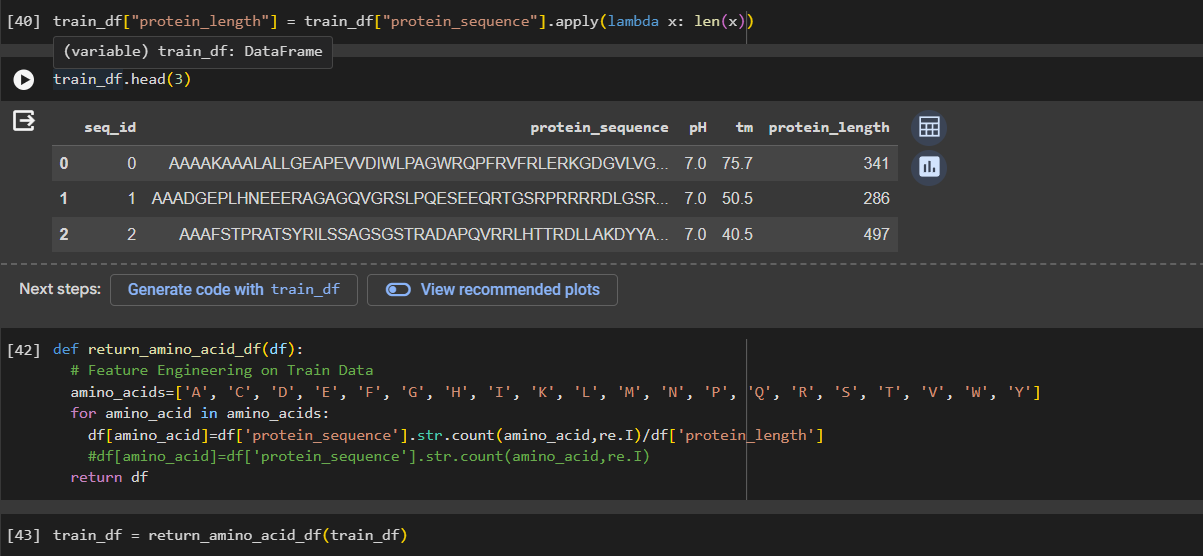
* 1. **Training:**
     1. **Splitting data:** Divide the dataset into training and testing sets, typically using a 70-30 or 80-20 split ratio.
     2. **Model training:** Fit the selected machine learning models (XGBoost, Random Forest, SVM) on the training data using appropriate libraries such as scikit-learn in Python.
     3. **Hyperparameter tuning:** Optimize model hyperparameters using techniques like grid search or random search to improve model performance.
     4. **Cross-validation:** Perform k-fold cross-validation on the training data to assess model generalization and stability.
  2. **Evaluation and Validation:**
     1. **Evaluation Metrics:** For evaluating the model use ***spearman correlation***. **Spearman correlation:** The Spearman correlation coefficient is calculated by first ranking each variable, then applying the Pearson correlation coefficient formula to these ranks. The Spearman correlation coefficient can range from -1 to +1. A coefficient of +1 indicates a perfect positive association of ranks, -1 indicates a perfect negative association of ranks, and 0 indicates no association between ranks.

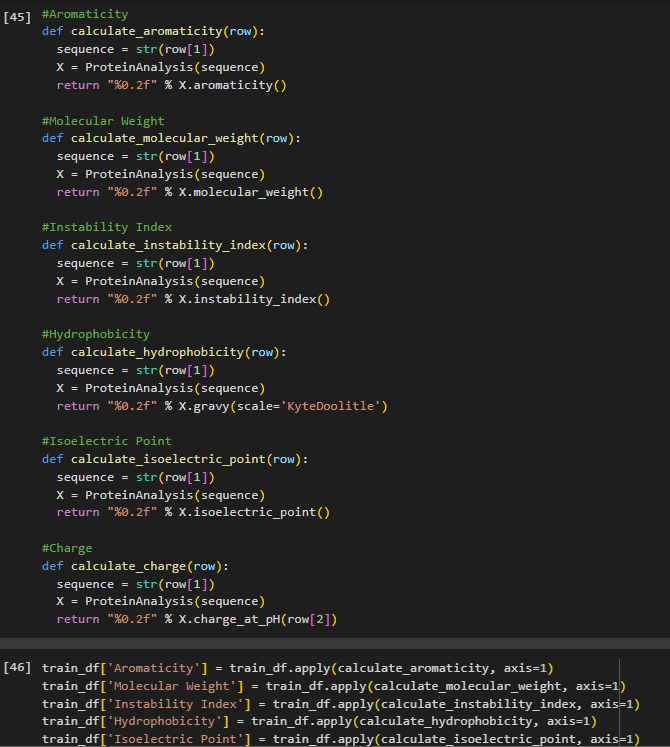
1. **Implementation:**

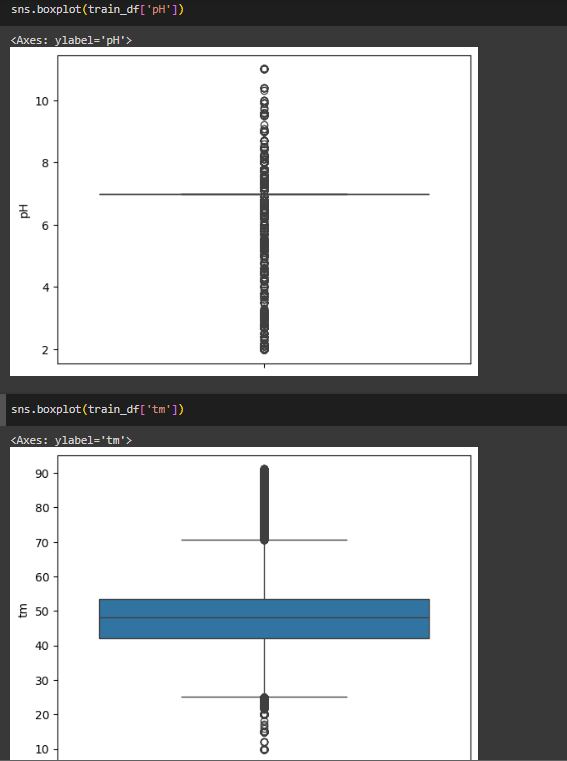


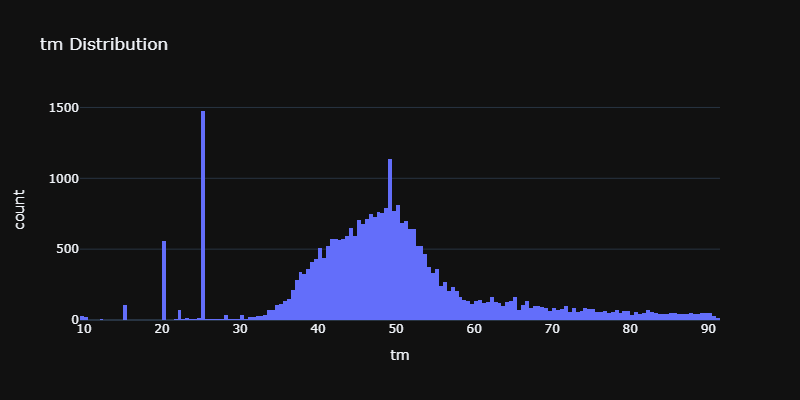


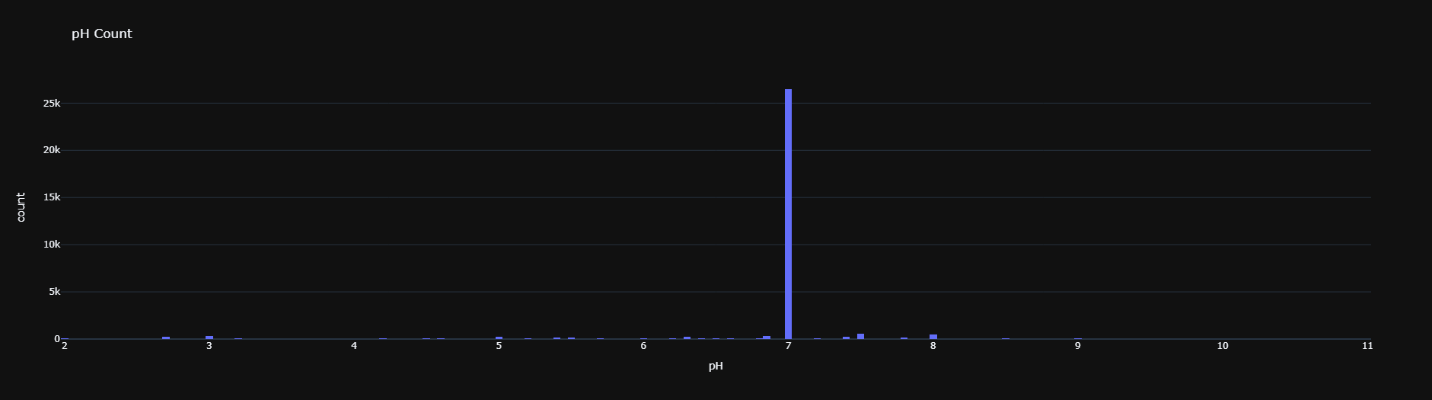


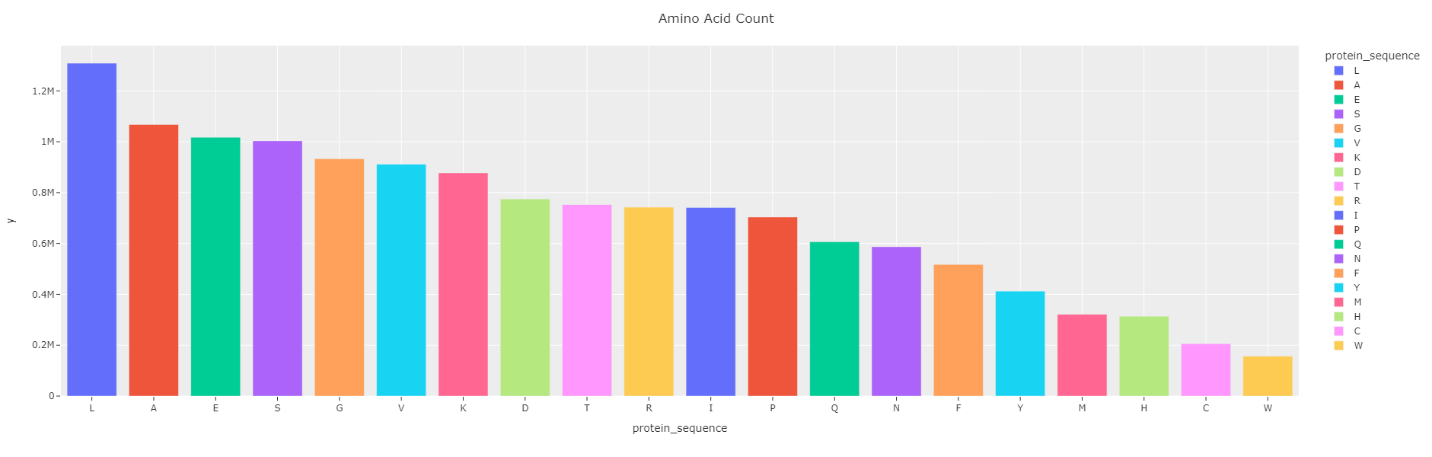


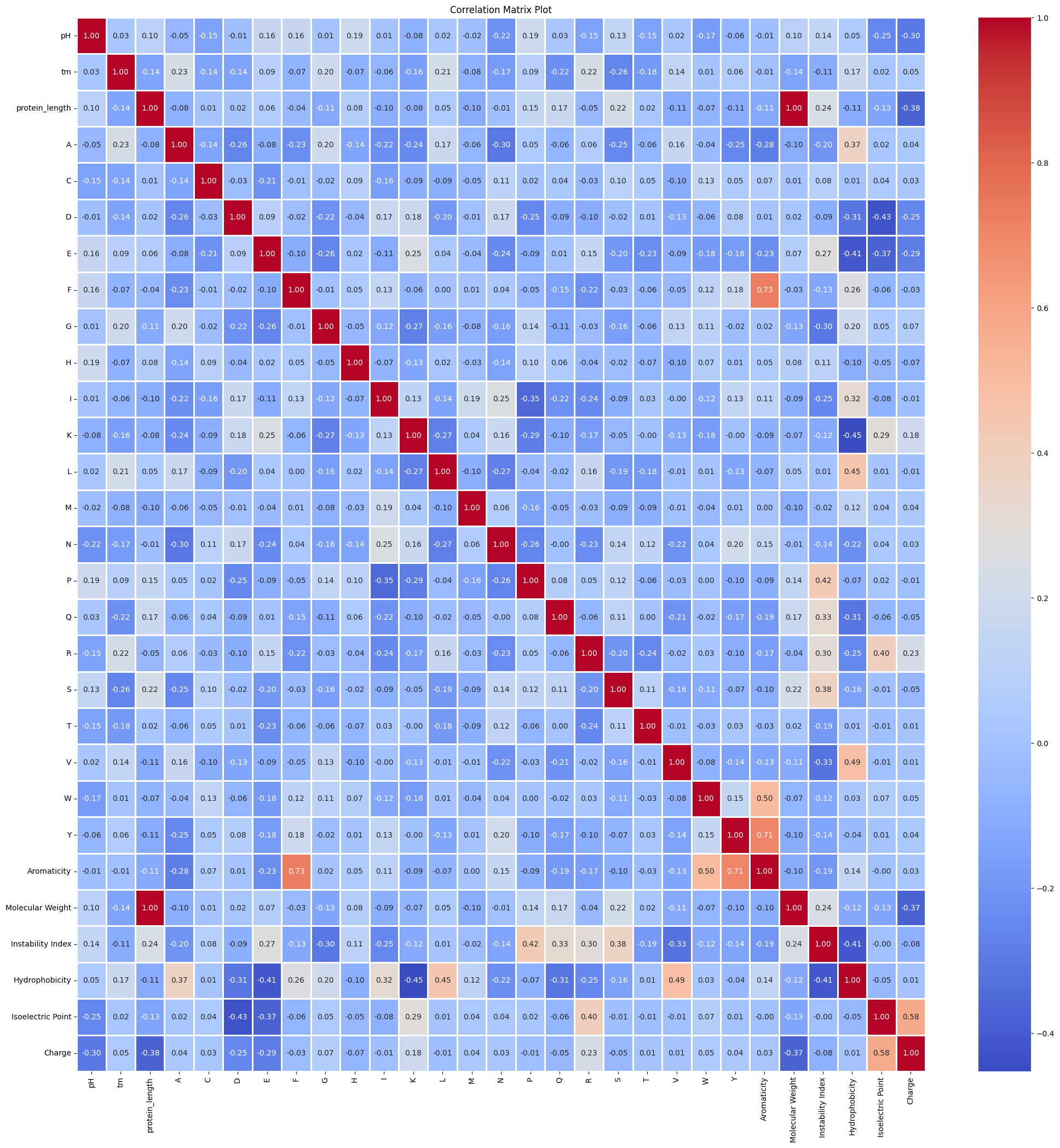




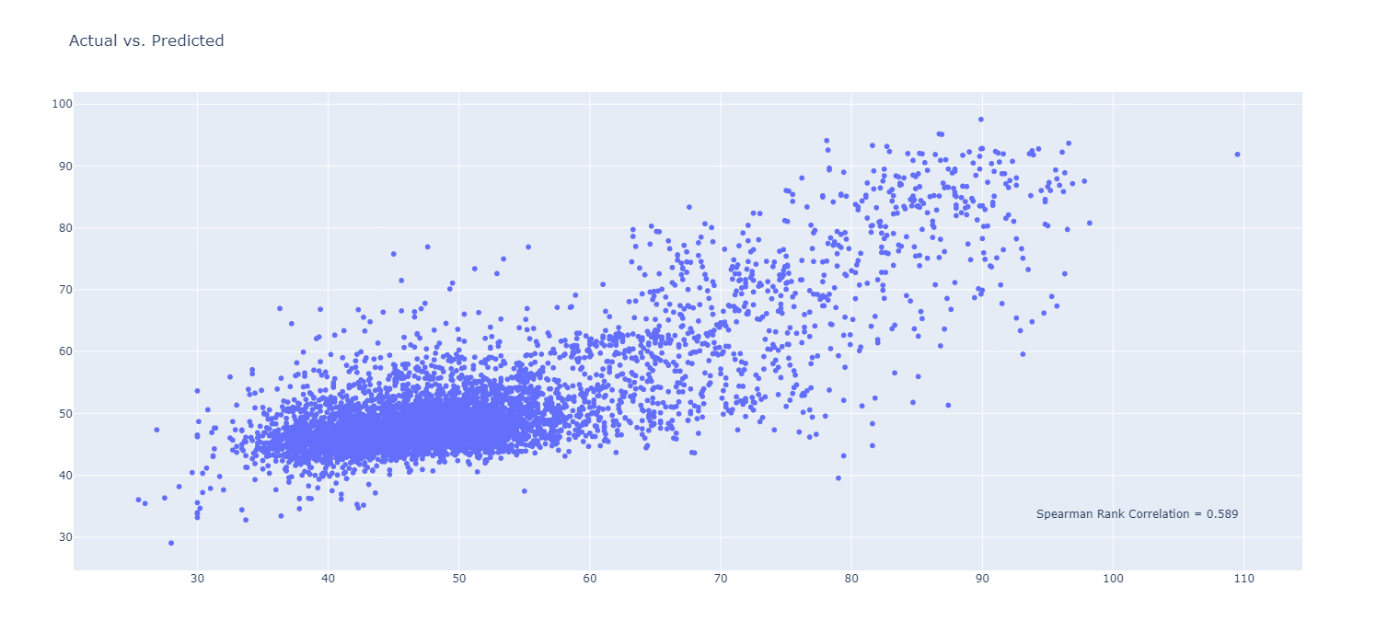


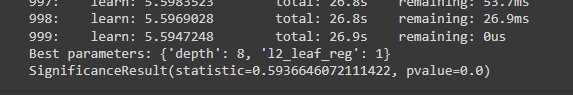




****

1. **Testing And Deployment:**
   1. **Testing Strategy:** Testing was done with train-test split.
   2. **Deployment Strategy:**
      1. **Integration with systems:** Enable easy access to the model's predictions through APIs or web services integrated with laboratory information management systems (LIMS) or production control system.
      2. **User Interface:** Design a user-friendly interface that allows scientists, engineers, and technicians to interact with the model. Incorporate visualization tools within the interface to help users interpret the model's predictions, such as graphs showing enzyme stability under different conditions. Provide training and documentation to ensure that users can effectively utilize the model and interface.
      3. **Maintenance and Updates:** Set up a system for continuous learning, where the model can be updated with new data from ongoing enzyme stability experiments and production processes.
      4. **Security:** Ensure compliance with data privacy regulations and industry standards for data protection.
2. **Result and Discussion:**



****

Got a spearman coeff. of 0.54.

Challenges and Limitations:

* Data Quality: Ensuring the quality and representativeness of the dataset posed challenges, particularly in obtaining comprehensive data covering a wide range of operating conditions and scenarios.
* Feature Engineering: Extracting informative features from raw data and determining their relevance to CHF prediction required careful consideration and domain expertise.
* Model Complexity: Balancing model complexity and interpretability was a challenge, as more complex models tended to provide better predictive performance but were harder to interpret and explain.

1. **Conclusion and Future Work**

The enzyme stability prediction project represents a significant innovation in the field of biotechnology and chemical engineering. By leveraging advanced AI/ML models, the project aims to accurately predict the thermostability of enzyme variants, which is a critical factor in their industrial application. The key points of the project can be summarized as follows:

* **Addressing a Critical Need:** The project focuses on solving the challenge of enzyme instability under industrial conditions, particularly at high temperatures, which is a major bottleneck in the widespread use of enzymes in various industries.
* **Advanced Computational Models:** A range of AI/ML models, including Support Vector Machines, Random Forest, XGBoost, and deep learning architectures, have been considered for their ability to handle the complex patterns in enzyme sequence and structure data.
* **Real-World Application:** The predictive model has practical applications in industries such as biofuels, pharmaceuticals, and food processing, where it can guide the selection and engineering of enzymes for improved stability and efficiency.
* **Environmental and Economic Impact:** By enabling the use of more stable enzymes, the project contributes to more sustainable industrial processes, reducing energy consumption and waste, and leading to cost savings.
* **Accelerating Innovation:** The ability to predict enzyme stability rapidly and accurately can significantly speed up the enzyme engineering process, facilitating the development of new and improved biocatalysts.

In conclusion, the enzyme stability prediction project stands out for its potential to transform industrial biocatalysis. It combines cutting-edge computational techniques with practical applications, offering a pathway to more sustainable and efficient manufacturing processes. The project's success could have a lasting impact on multiple sectors, driving innovation and sustainability in equal measure.

1. **References**

* Anna Jarzab, Nils Kurzawa, Thomas Hopf, Matthias Moerch, Jana Zecha, Niels Leijten, Yangyang Bian, Eva Musiol, Melanie Maschberger, Gabriele Stoehr, Isabelle Becher, Charlotte Daly, Patroklos Samaras, Julia Mergner, Britta Spanier, Angel Angelov, Thilo Werner, Marcus Bantscheff, Mathias Wilhelm, Martin Klingenspor, Simone Lemeer, Wolfgang Liebl, Hannes Hahne, Mikhail M. Savitski & Bernhard Kuster,***” Meltome atlas—thermal proteome stability across the tree of life”***, Nature Methods (13 April 2020)

1. **Auxiliaries**

Data taken from Kaggle.

**10. Appendices**

**DataSource:**<https://www.kaggle.com/c/novozymes-enzyme-stability-prediction>

**Python file:** <https://colab.research.google.com/drive/1fpM8VCaUQUv0yXTyMAQ5Kx5i5DLs4Mnm?usp=sharing>