**Chapter – 1**

**INTRODUCTION**

* 1. **Overview of the project**

Novozymes finds enzymes in nature and optimizes them for use in industry. In industry, enzymes replace chemicals and accelerate production processes. They help our customers make more from less, while saving energy and generating less waste. Enzymes are widely used in laundry and dishwashing detergents where they remove stains and enable low-temperature washing and concentrated detergents. Other enzymes improve the quality of bread, beer and wine, or increase the nutritional value of animal feed. Enzymes are also used in the production of biofuels where they turn starch or cellulose from biomass into sugars which can be fermented to ethanol. These are just a few examples as we sell enzymes to more than 40 different industries. Like enzymes, microorganisms have natural properties that can be put to use in a variety of processes. Novozymes supplies a range of microorganisms for use in agriculture, animal health and nutrition, industrial cleaning and wastewater treatment.

However, many enzymes are only marginally stable, which limits their performance under harsh application conditions. Instability also decreases the amount of protein that can be produced by the cell. Therefore, the development of efficient computational approaches to predict protein stability carries enormous technical and scientific interest.

Computational protein stability prediction based on physics principles have made remarkable progress thanks to advanced physics-based methods such as FoldX, Rosetta, and others. Recently, many machine learning methods were proposed to predict the stability impact of mutations on protein based on the pattern of variation in natural sequences and their three-dimensional structures. More and more protein structures are being solved thanks to the recent breakthrough of AlphaFold2. However, accurate prediction of protein thermal stability remains a great challenge.

In this competition, Novozymes invites you to develop a model to predict/rank the thermostability of enzyme variants based on experimental melting temperature data, which is obtained from Novozymes’s high throughput screening lab. You’ll have access to data from previous scientific publications. The available thermostability data spans from natural sequences to engineered sequences with single or multiple mutations upon the natural sequences. If successful, you'll help tackle the fundamental problem of improving protein stability, making the approach to design novel and useful proteins, like enzymes and therapeutics, more rapidly and at lower cost.

Novozymes is the world’s leading biotech powerhouse. Our growing world is faced with pressing needs, emphasizing the necessity for solutions that can ensure the health of the planet and its population. At Novozymes, we believe biotech is at the core of connecting those societal needs with the challenges and opportunities our customers face. Novozymes is the global market leader in biological solutions, producing a wide range of enzymes, microorganisms, technical and digital solutions which help our customers, amongst other things, add new features to their products and produce more from less.

* 1. **Goal of the competition**

Enzymes are proteins that act as catalysts in the chemical reactions of living organisms. The goal of this competition is to predict the thermostability of enzyme variants. The experimentally measured thermostability (melting temperature) data includes natural sequences, as well as engineered sequences with single or multiple mutations upon the natural sequences.

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.

* 1. **Objectives of the project**
* The objective is to predict the thermostability of enzyme variants
* improvements to enzyme stability could lower costs and increase the speed scientists can iterate on concepts.
* Novozymes finds enzymes in nature and optimizes them for use in industry. In industry, enzymes replace chemicals and accelerate production processes
* Novozymes finds enzymes in nature and optimizes them for use in industry. and enzymes replace chemicals and accelerate production processes. where they can help our customers make more from less, while saving energy and generating less waste.
  1. **Literature survey**

1. Advancement and comparative profiles in the production technologies using solid-state and submerged fermentation for microbial cellulases. RR Singhania, RK Sukumaran, AK Patel has done Enzyme and Microbial on Novozymes in 2010.
2. C Axel, E Zannini , EK Arendt - Critical Reviews in food science 2017 - Taylor & Francin.In 2011, Novozymes surveyed over 4000 bread consumers throughout Europe.
3. S Li, X Yang, S Yang, M Zhu, X Wang - Computational and structural in 2012 - After two generations of Cellic® release in 2009 and 2010.
4. W Wei, C Sun, X Wang, Q Jin, X Xu, CC Akoh, X Wang - Engineering, Silica gel Novozymes (Denmark) Lipozyme RM IM Rhizomucor miehei Ion-exchange resin in 2020
5. JO Metzger, U Bornscheuer - Applied microbiology and biotechnology, Currently, Betapol is manufactured by interesterification of tripalmitin with oleic acid using a lipase from Rhizomucor miehei (Novozyme RM IM) in 2006.
6. N Vaishnav, A Singh, M Adsul, P Dixit… - Bioresource Technology in 2018 - Elsevier Its importance can be judged by the fact that Novozyme Cellic Ctec2 which is superior version of cellulase series than celluclast; is different in having additional β-glucosidas
7. U Biermann, U Bornscheuer, MAR Meier… - Angewandte Chemie …, 2011 - Wiley Online Library Betapol is manufactured by transesterification of tripalmitin with oleic acid using a lipase from Rhizomucor miehei (Novozyme RMIM).
8. C De Villiers, PCK Hsiao - Sustainability accounting in 2017 - taylorfrancis.com Novozymes’ CEO described the decision to prepare an integrated report as “a natural consequence of business and sustainability moving ever closer together
9. AA Al-Ghanayem, B Joseph - Applied microbiology and biotechnology in 2020 - Springer Lipoclean® is a cold-active lipase produced by Novozymes that is used for triglyceride stains. It is active at low temperature (≃ 20 C), and stable with other enzymes
10. U Bornscheuer, K Buchholz… -Angewandte Chemie in 2014 - Wiley Online Library reesei) and Novozyme 188 (both products from Novozymes, Denmark). 13b This method is attractive owing to its cost-saving potential and reactor construction (low-corrosion potential
    1. **Problem definition**

**Novozymes Enzyme Stability Prediction:** It help identify the thermostable mutations in enzymes. This is to develop models that can predict the ranking of protein thermostability (as measured by melting point, tm) after single-point amino acid mutation and deletion.

**Chapter – 2**

**PROPOSED SYSTEM**

**2.1 Understanding the data**

In this competition, you are asked to develop models that can predict the ranking of protein thermostability (as measured by melting point, tm) after single-point amino acid mutation and deletion. For the training set, the protein thermostability (experimental melting temperature) data includes natural sequences, as well as engineered sequences with single or multiple mutations upon the natural sequences. The data are mainly from different sources of published studies such as [Meltome atlas—thermal proteome stability across the tree of life](https://doi.org/10.1038/s41592-020-0801-4). Many other public datasets exist for protein stability; please see the competition [Rule 7C](https://www.kaggle.com/competitions/novozymes-enzyme-stability-prediction/rules) for external data usage requirements. There are also other publicly available methods which can predict protein stabilities such as [ESM](https://doi.org/10.1073/pnas.2016239118), [EVE](https://doi.org/10.1038/s41586-021-04043-8) and [Rosetta](https://doi.org/10.1016/B978-0-12-381270-4.00019-6) etc., without using the provided training set. These methods may also be used as part of the competition.

The test set contains experimental melting temperature of over 2,413 single-mutation variants of an enzyme (GenBank: KOC15878.1), obtained by Novozymes A/S. The amino acid sequence of the wild type is:

VPVNPEPDATSVENVALKTGSGDSQSDPIKADLEVKGQSALPFDVDCWAILCKGAPNVLQRVNEKTKNSNRDRSGANKGPFKDPQKWGIKALPPKNPSWSAQDFKSPEEYAFASSLQGGTNAILAPVNLASQNSQGGVLNGFYSANKVAQFDPSKPQQTKGTWFQITKFTGAAGPYCKALGSNDKSVCDKNKNIAGDWGFDPAKWAYQYDEKNNKFNYVGK

**2.2 Data set description**

**Dataset 1:**

**Train.csv** - the training data, with columns as follows:

* seq\_id: unique identifier of each protein variants
* protein\_sequence: amino acid sequence of each protein variant. The stability (as measured by tm) of protein is determined by its protein sequence. (Please note that most of the sequences in the test data have the same length of 221 amino acids, but some of them have 220 because of amino acid deletion.)
* pH: the scale used to specify the acidity of an aqueous solution under which the stability of protein was measured. Stability of the same protein can change at different pH levels.
* data\_source: source where the data was published
* tm: target column. Since only the spearman correlation will be used for the evaluation, the correct prediction of the relative order is more important than the absolute tm values. (Higher tm means the protein variant is more stable.)

**Dataset 2:**

**Train\_updates\_20220929.csv** - corrected rows in train, please see this [forum post](https://www.kaggle.com/competitions/novozymes-enzyme-stability-prediction/discussion/356251) for details

**Dataset 3:**

**Test.csv** - the test data; your task is to predict the target tm for each protein\_sequence (indicated by a unique seq\_id)

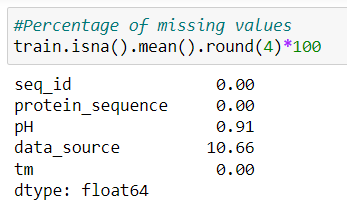
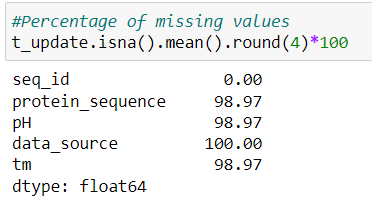
**Dataset 4:**

**Sample\_submission.csv** - a sample submission file in the correct format, with seq\_id values corresponding to test.csv

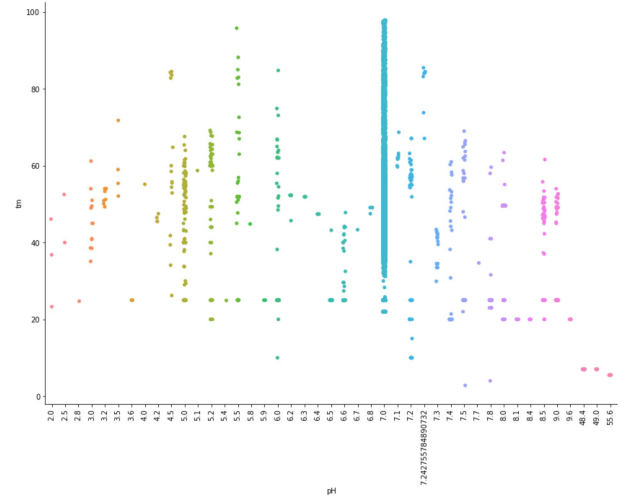
**2.2.1 Attributes of the Train dataset**

|  |  |
| --- | --- |
| **Attributes** | **Description** |
| seq\_id | unique identifier of each protein variants |
| protein\_sequence | amino acid sequence of each protein variant |
| pH | the scale used to specify the acidity of an aqueous solution |
| data\_source | source where the data was published |
| tm | stability of each protein variant is measured by tm |

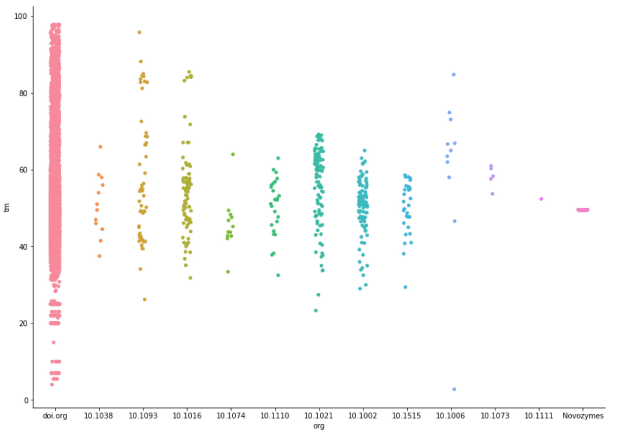
**2.2.2 Percentage of null values in dataset**

**2.3 Visualisation of Data**

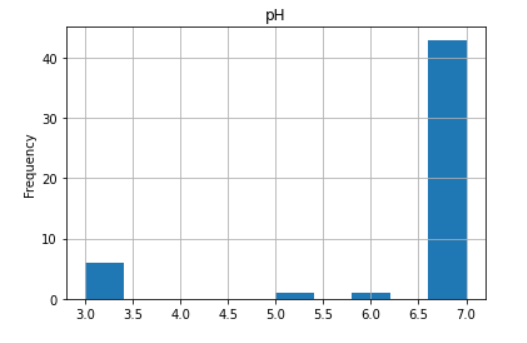
****

Comparing all pH values with target variable

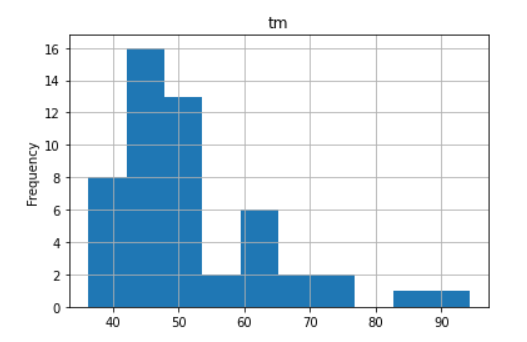
****

Comparing all org values with target variable

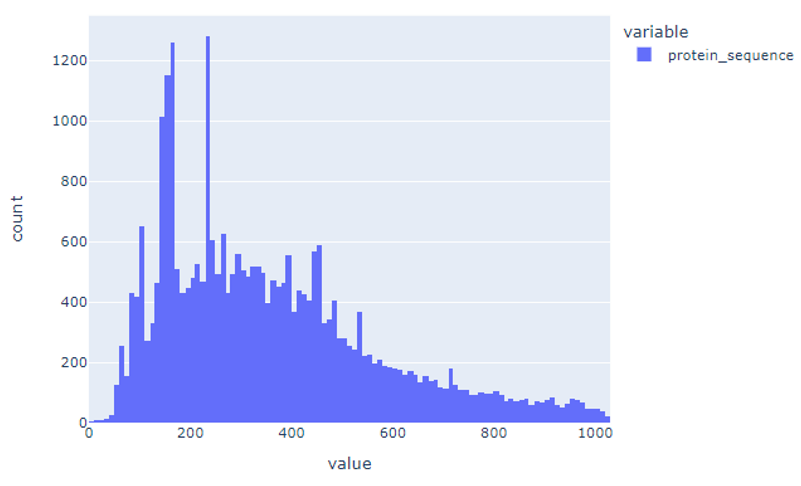
**2.3.1 Analysis of Data distribution**

****

Finding frequency of ‘pH’ value



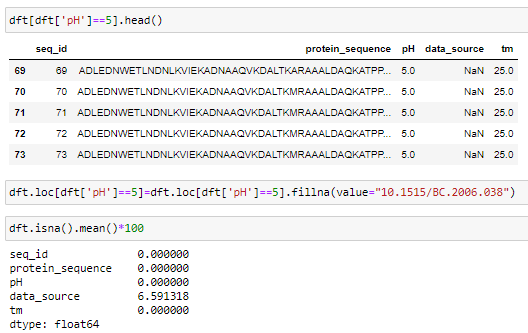
Finding frequency of ‘tm’ value

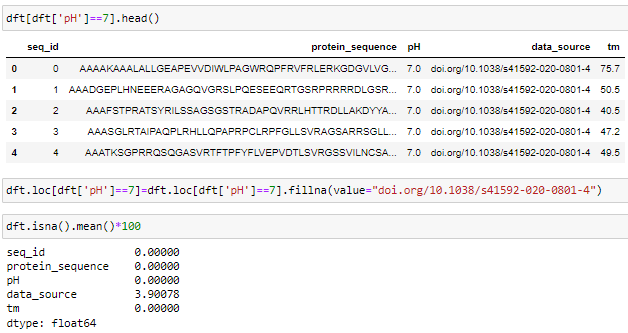


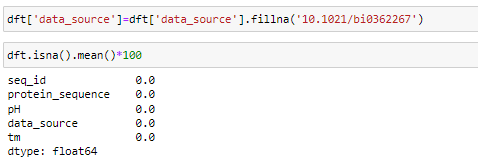
Finding frequency of ‘protein sequences’ value

**2.4 Data Pre-processing**

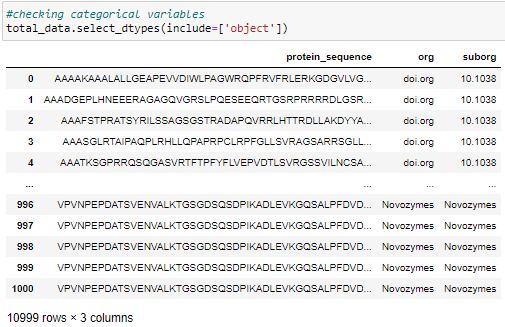
**Filling null values with fillna() method**



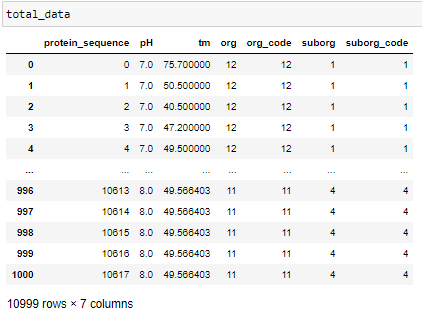


**s**

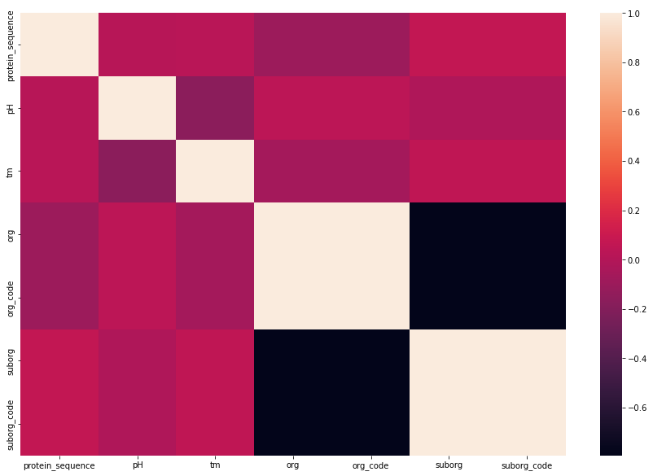
**Finding categorical value**



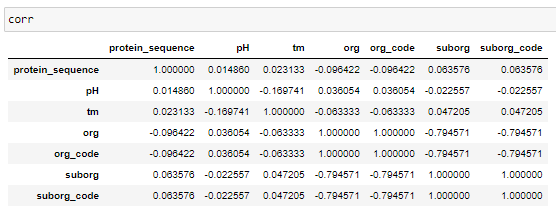
**Converting categorical to numerical**



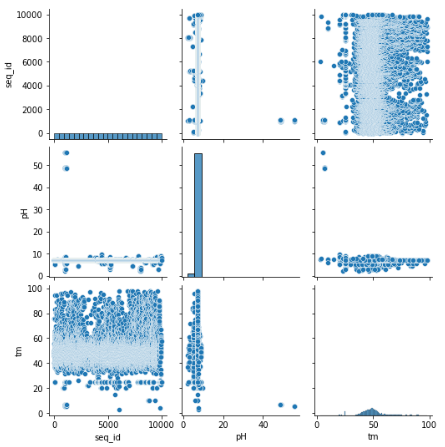
**Finding the co-relation of the attributes**



Co-relation of all the attributes



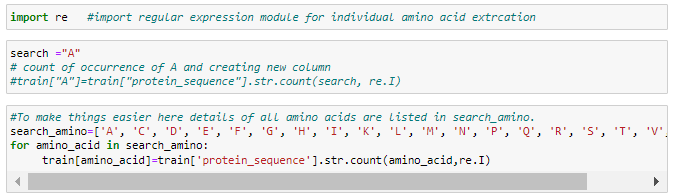
**Pair plot graph for train data**



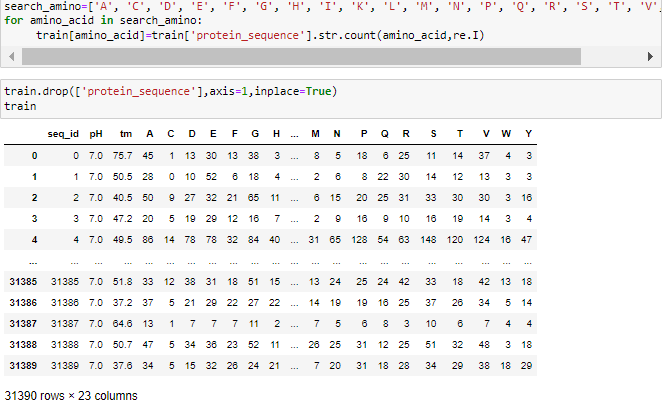
Pair plot of train dataset

**2.5 Initial Approach**

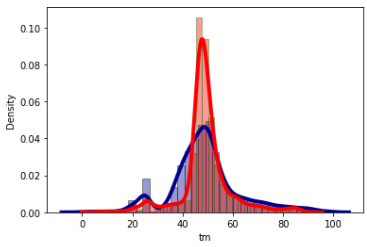
* import regular expression module for individual amino acid extraction
* count of occurrence of A and creating new column
* train["A"]=train["protein\_sequence"].str.count(search, re.I)

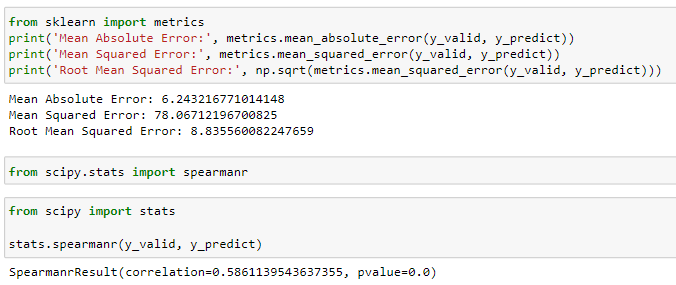


* To make things easier here details of all amino acids are listed in search\_amino.



* Density plot and Histogram of all arrival delay





**2.6 Proposed methodology**

Since its labelled dataset we are going to use supervised machine learning algorithm. As we can see the target variable 'tm' in pair plot values are not classified. It contains random values so considering this we are using below models.

**Data models used to predicate the accuracy.**

1. Random Forest regressor model
2. KNN regressor model
3. XG Booster model
4. Pipeline model

**2.6.1 Proposed model 1: Random Forest regressor model**

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predict accuracy of our dataset. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

**2.6.2 Proposed model 2: KNN Regressor model**

It breaks down a dataset into smaller and smaller subnets while at the same time an associated decision tree is incrementally developed. Regression based on k-nearest neighbors. The target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set.

**2.6.3 Proposed model 3: XG Booster Model**

This provides accurate results and a highly scalable training method that avoids overfitting. **XGBoost** is an optimized distributed gradient boosting library designed to be highly **efficient**, **flexible** and **portable**. It implements machine learning algorithms under the [Gradient Boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

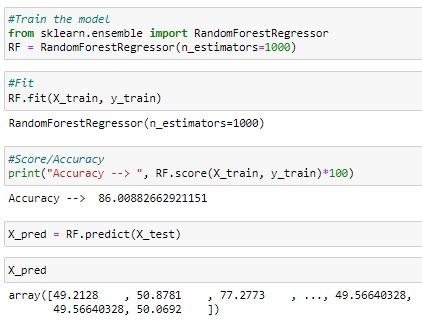
**2.6.4 Proposed model 4: Pipeline Model**

A machine learning pipeline is a way to codify and automate the workflow it takes to produce a machine learning model.

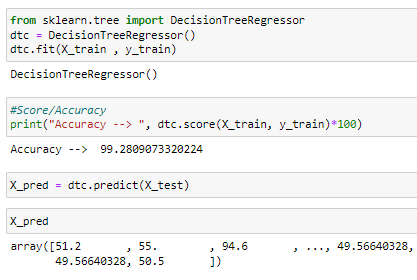
**Chapter – 3**

**IMPLEMENTATION**

**3.1 Proposed model 1: Random Forest regressor model**

****

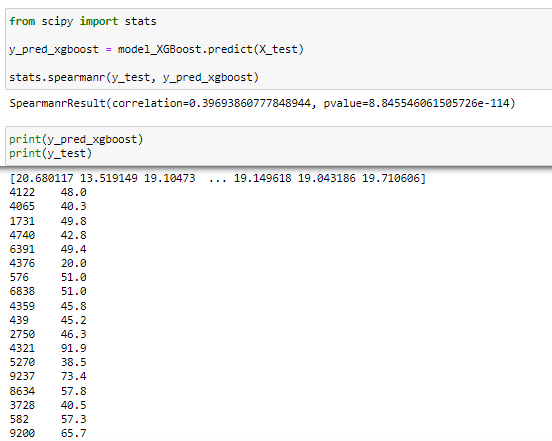
**3.2 Proposed model 2: KNN Regressor model**

****

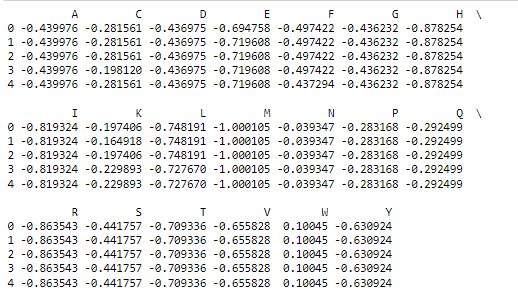
**3.3 Proposed model 3: XGBooster Model**

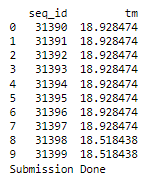
****

**Spearman’s rank correlation coefficient**

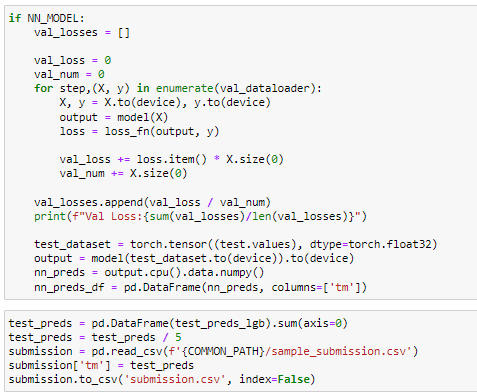
****

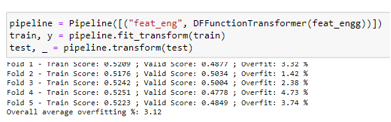
**Prediction of XGBooster Model**

****

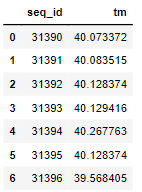
****

**3.4 Proposed model 4: Pipeline Model**

****

****

**Prediction of Pipeline Model**

****

**Chapter – 4**

**RESULTS AND DISCUSSION**

**M**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No.** | **Model Name** | **Description** | **Accuracy** |
| 1 | XG Booster | This provides accurate results and a highly scalable training method that avoids overfitting | 80.9284 |
| 2 | Random forest regressor | Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predict accuracy of our dataset | 86.00882662921151 |
| 3 | KNN Regressor | It breaks down a dataset into smaller and smaller subnets while at the same time an associated decision tree is incrementally developed | 99.2809073320224 |
| 4 | Pipeline Model | A machine learning pipeline is a way to codify and automate the workflow it takes to produce a machine learning model. | 40.0835 |

**4.2 Efficient result**

KNN Regressor model provides efficient result with 99.2809073320224% accuracy. Therefore, the KNN Regressor model is the most efficient and suitable model for the enzyme prediction.

**Chapter – 8**

**CONCLUSION AND FUTURE SCOPE**

We have trained the dataset by using four models which are XGBoost model, KNN model, pipeline regression model, random forest model and by bias variance trade off where we have reduced the overfitting of the model in which we have predicted KNN model is the best model for our dataset

**Chapter – 9**

**REFERENCES**

* <https://www.kaggle.com/c/novozymes-enzyme-stability-prediction>
* <https://www.kaggle.com/code/seyered/eda-novozymes-enzyme-stability>
* <https://www.geeksforgeeks.org/xgboost/>
* <https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/>
* <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
* <https://scikit-learn.org/stable/modules/feature_extraction.html>