Openclassroom

Master's degree

Ingénieur Machine Learning

Rapport de la compétition kaggle pour la fin de la formation

Prédiction de la stabilité des enzymes Novozymes

Mentor: Nicolas Tisserand
Email: ntisserand@gmail.com

Etudiant : Ousmane LO
Email : ousmanelo78@gmail.com

Table des matières

I. Introduction 2

II. Présentation du projet 2

1. Contexte 2
2. Objectif 3

III. La base de données 3

1. Origine et description des données 3
2. Préparation des données 4

IV. Exploration des données 5

V. Modélisation 7

1. Modèle et performance 7
2. Importance des features 8

VI. Conclusion 10

Table des figures

Figure 1: Structure tridimensionnelle de l'enzyme 4
Figure 2: Vue d'ensemble de la solution 4
Figure 3: Histogramme des pH 5
Figure 4: Distribution du pH 6
Figure 5: Distribution des tm 6
Figure 6: Analyse des résidus 7
Figure 7: Importance des features via shap 8
Figure 8: Impact des features 9

I. Introduction

Pour terminer la formation, il est demandé de participer à une compétition Kaggle en cours. Parmi les compétitions ouvertes au début de ce projet, celles servant à l'apprentissage, et donc les compétitions très connues comme la classification des chiffres manuscrits ou la prédiction de survivants sur le Titanic sont écartées. Par conséquent, puisque je m'intéresse plus particulièrement aux données biologiques, la compétition « Novozymes Enzyme Stability Prediction » (lien ici) a été choisie. La récompense attribuée à cette compétition s'élève à 25k$.

Le temps attribué pour ce projet étant assez court, une étude complète n'est pas envisagée. Le choix a été fait de s'inspirer des travaux réalisés par d'autres compétiteurs et partagés sur Kaggle afin d'essayer d'améliorer la performance de leurs modèles. Plus en détails, une première partie sera consacrée à l'exploration des données afin de se familiariser avec les données. Ensuite la partie modélisation sera discutée, en s'appuyant sur un traitement de données proposé par d'autres utilisateurs, en essayant de le modifier légèrement. Une moyenne de prédiction de différents modèles sera réalisée afin d'espérer un gain de performance. Pour finir, la soumission des résultats et les scores obtenus sur différents essais sera présentée.

II. Présentation du projet

1. Contexte

Novozymes le leader mondial du marché des solutions biologiques, trouve des enzymes dans la nature et les optimise pour une utilisation dans l'industrie. Dans l'industrie, les enzymes remplacent les produits chimiques et accélèrent les processus de production. Ils aident ces clients à faire plus avec moins, tout en économisant de l'énergie et en générant moins de déchets. Les enzymes sont largement utilisées dans les détergents à lessive et à vaisselle où elles éliminent les taches et permettent un lavage à basse température et des détergents concentrés. D'autres enzymes améliorent la qualité du pain, de la bière et du vin, ou augmentent la valeur nutritionnelle des aliments pour animaux. Les enzymes sont également utilisées dans la production de biocarburants où elles transforment l'amidon ou la cellulose de la biomasse en sucres qui peuvent être fermentés en éthanol. Ce ne sont que quelques exemples car nous vendons des enzymes à plus de 40 industries différentes. Comme les enzymes, les micro-organismes ont des propriétés naturelles qui peuvent être utilisées dans une variété de processus.

Cependant, de nombreuses enzymes ne sont que marginalement stables, ce qui limite leurs performances dans des conditions d'application difficiles. L'instabilité diminue également la quantité de protéines pouvant être produites par la cellule. Par conséquent, le développement d'approches informatiques efficaces pour prédire la stabilité des protéines présente un énorme intérêt technique et scientifique.

La prédiction informatique de la stabilité des protéines basée sur les principes de la physique a fait des progrès remarquables grâce à des méthodes avancées basées sur la physique telles que FoldX, Rosetta et autres. Récemment, de nombreuses méthodes d'apprentissage automatique ont été proposées pour prédire l'impact sur la stabilité des mutations sur les protéines en fonction du modèle de variation des séquences naturelles et de leurs structures tridimensionnelles. De plus en plus de structures protéiques sont résolues grâce à la récente percée d'AlphaFold2. Cependant, la prédiction précise de la stabilité thermique des protéines reste un grand défi.

Dans ce concours, Novozymes nous invite à développer un modèle pour prédire/classer la thermostabilité des variants enzymatiques sur la base de données expérimentales de température de fusion, obtenues du laboratoire de criblage à haut débit de Novozymes. Vous aurez accès aux données des publications scientifiques précédentes. Les données de thermo stabilité disponibles s'étendent des séquences naturelles aux séquences modifiées avec des mutations simples ou multiples sur les séquences naturelles. En cas de succès, vous aiderez à résoudre le problème fondamental de l'amélioration de la stabilité des protéines, en rendant l'approche de conception de protéines nouvelles et utiles, comme les enzymes et les thérapeutiques, plus rapidement et à moindre coût.

2. Objectif

Les enzymes sont des protéines qui agissent comme des catalyseurs dans les réactions chimiques des organismes vivants. L'objectif de ce concours est de prédire la thermostabilité de variants enzymatiques. Les données de thermostabilité (température de fusion) mesurées expérimentalement comprennent des séquences naturelles, ainsi que des séquences modifiées avec des mutations simples ou multiples sur les séquences naturelles. Comprendre et prédire avec précision la stabilité des protéines est un problème fondamental en biotechnologie. Ses applications incluent l'ingénierie enzymatique pour relever les défis mondiaux en matière de durabilité, de neutralité carbone et plus encore. Des améliorations de la stabilité des enzymes pourraient réduire les coûts et augmenter la vitesse à laquelle les scientifiques peuvent itérer sur les concepts.

III. La base de données

1. Origine et description des données

Dans ce concours, nous étions invités à développer des modèles capables de prédire le classement de la thermostabilité des protéines (mesurée par le point de fusion, `tm`) après une mutation et une délétion d'acides aminés en un seul point.
Pour l'ensemble d'apprentissage, les données de thermostabilité des protéines (température de fusion expérimentale) comprennent des séquences naturelles, ainsi que des séquences modifiées avec des mutations simples ou multiples sur les séquences naturelles. Les données proviennent principalement de différentes sources d'études publiées telles l'atlas Meltome - stabilité du protéome thermique à travers l'arbre de la vie.

Les données ci-dessous ont été mis à notre disposition pour la compétition :

- **train.csv** - les données d'entraînement, avec les colonnes suivantes :
  - `seq_id`: identifiant unique de chaque variant protéique
  - `protein_sequence`: séquence d'acides aminés de chaque variant protéique. La stabilité (mesurée par `tm`) d'une protéine est déterminée par sa séquence protéique.
  - `pH`: l'échelle utilisée pour spécifier l'acidité d'une solution aqueuse sous laquelle la stabilité des protéines a été mesurée. La stabilité d'une même protéine peut changer à différents niveaux de pH.
  - `data_source`: source où les données ont été publiées
  - `tm`: colonne cible. Étant donné que seule la corrélation de Spearman sera utilisée pour l'évaluation, la prédiction correcte de l'ordre relatif est plus importante que les `tm` valeurs absolues. (Plus élevé `tm` signifie que la variante protéique est plus stable.)
- **train_updates_20220929.csv** - lignes corrigées dans le train, veuillez consulter ce message du forum pour plus de détails

- **test.csv** - les données de test ; votre tâche est de prédire la cible `tm` pour chacun `protein_sequence`(indiqué par un unique `seq_id`)
- **sample_submission.csv** - un exemple de fichier de soumission au format correct, avec des `seq_id` valeurs correspondant à **test.csv**
- **wildtype_structure_prediction_af2.pdb** - la structure tridimensionnelle de l'enzyme répertoriée ci-dessus, comme prédit par AlphaFold

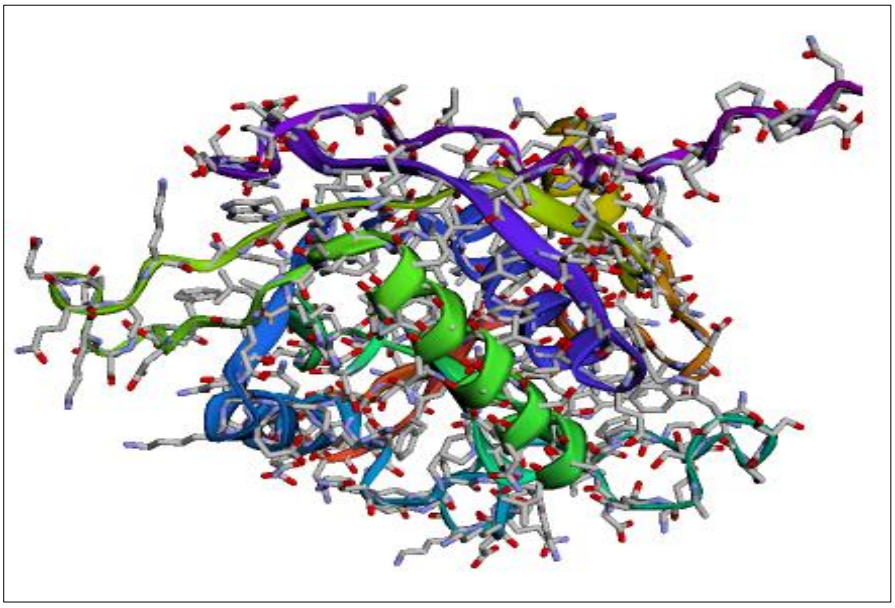

Figure 1: Structure tridimensionnelle de l'enzyme

2. Préparation des données

Pour préparer les données que dois utiliser notre modèle, nous avions effectué des extractions de features par embeding en utilisant le modèle ProBERT afin de transformer les variables catégorielles en variables continues.

ProtBERT est un modèle pré-entraîné sur des séquences protéiques utilisant un objectif de modélisation de langage masqué. Il est basé sur le modèle BERT, qui est pré-entraîné sur un large corpus de séquences protéiques de manière auto-supervisée. Cela signifie qu'il a été préformé uniquement sur les séquences de protéines brutes, sans qu'aucun humain ne les étiquette avec un processus automatique pour générer des entrées et des étiquettes à partir de ces séquences de protéines.

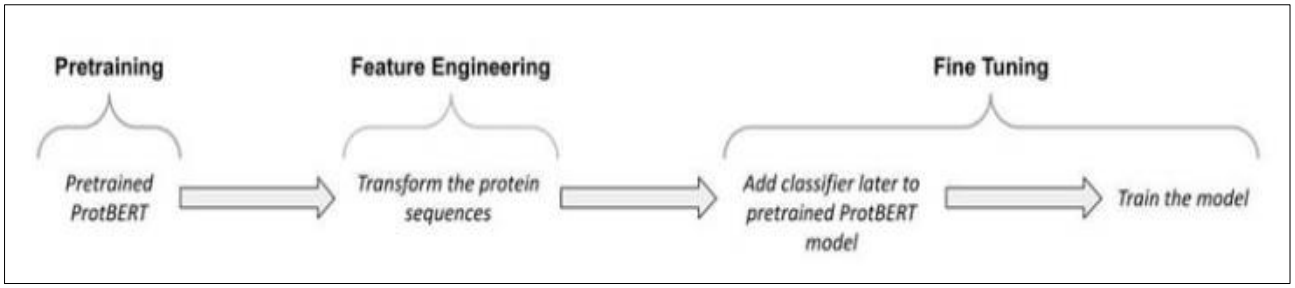

Figure 2: Vue d'ensemble de la solution

On a essayé de faire de feature engineering pour déterminer la longueur de séquence de protéine mais également déterminer une séquence unique d'acide aminé c'est à dire des variables allant de la lettre A jusqu'à Y. Nous avons combiné par la suite les données obtenues avec extraction de caractéristiques et celles de feature engineering.

IV. Exploration des données

L'analyse du pH montre une neutralité importante des solutions pour la plupart des données car la distribution du pH est plus importante au pH égale à 7(voir Figure 3). Quant à la distribution des tm on constate qu'elle ne suit pas une loi normale et qu'elle semblerait peut-être à une loi log normale (Figure 4). On constate aussi une distribution relation relativement faible quand la température devient de plus en plus importante.

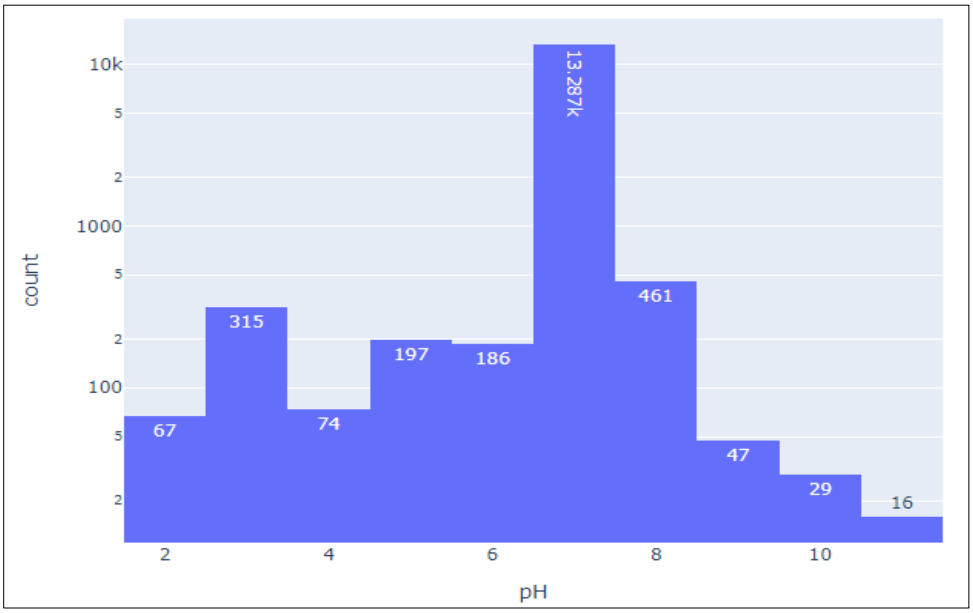

Figure 3: Histogramme des pH

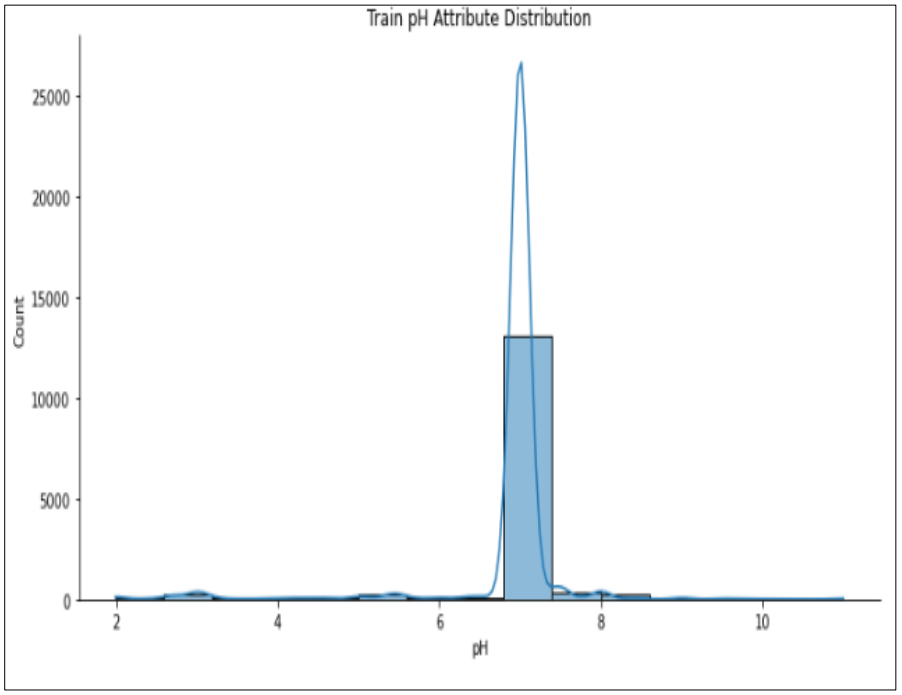

Figure 4: Distribution du pH

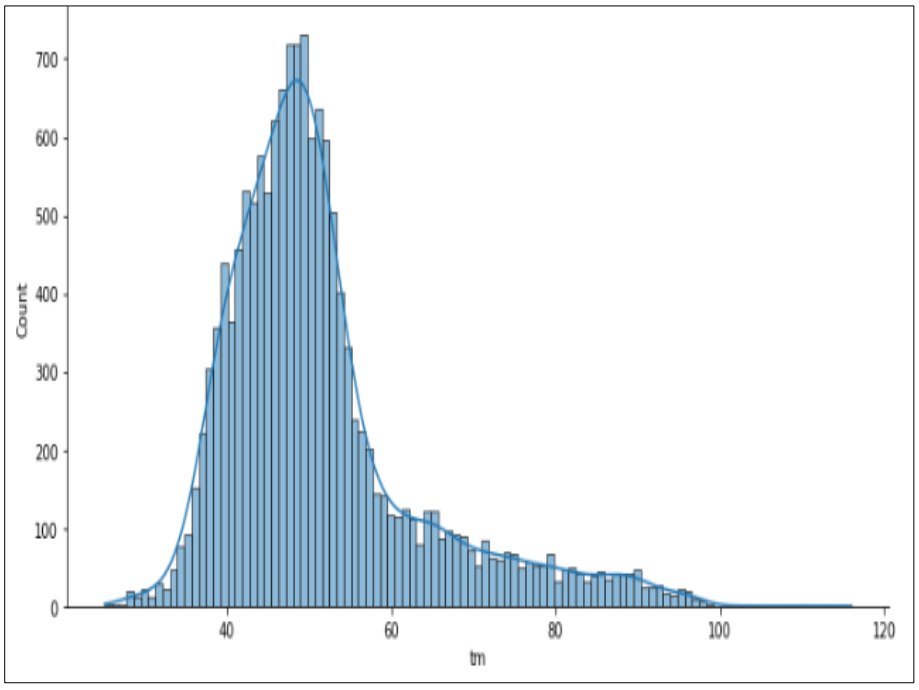

Figure 5: Distribution des tm

V. Modélisation

1. Modèle et performance

Après le traitement des données c'est à de l'extraction des données et features engineering et de la combination de ces dernières, on a choisi d'utiliser le modèle XGBOOST pour la prédiction des thermostabilités des enzymes. Les techniques basées sur le cross validation et de l'optimisation des hyperparamètres n'ont pas été explorées sur ce projet vu la taille importante des données qui demande trop de calcul et le temps très peu octroyé pour réaliser ce projet.

L'évaluation se base sur le coefficient de sperman également appelé rhô ou r(s) qui évalue l'association entre deux variables mesurées dans une échelle ordinale. Ce test statistique est réalisé à partir des rangs contrairement au coefficient de corrélation de Pearson qui se fait sur les valeurs. Il s'interprète de la même manière qu'un coefficient de corrélation de Pearson : une valeur positive (maximum = +1) indique une variation simultanée dans le même sens, une valeur négative (minimum = -1) une variation simultanée en sens inverse. Une valeur nulle indique que les deux classements n'ont rien à voir l'un par rapport à l'autre.

Nous avons obtenu de très bonnes performances avec le modèle car notre erreur de prédiction avec la métrique MAE est relativement grande (5.58) et que le coefficient Sperman est à 57.9%.

L'analyse des résidus montre que les résidus suivent une loi normale et qu'on a un écart de 22% entre le R^2 obtenu sur les données d'entrainement et celles de test. Ce qui révèle un potentiel phénomène de surapprentissage (Figure 6).

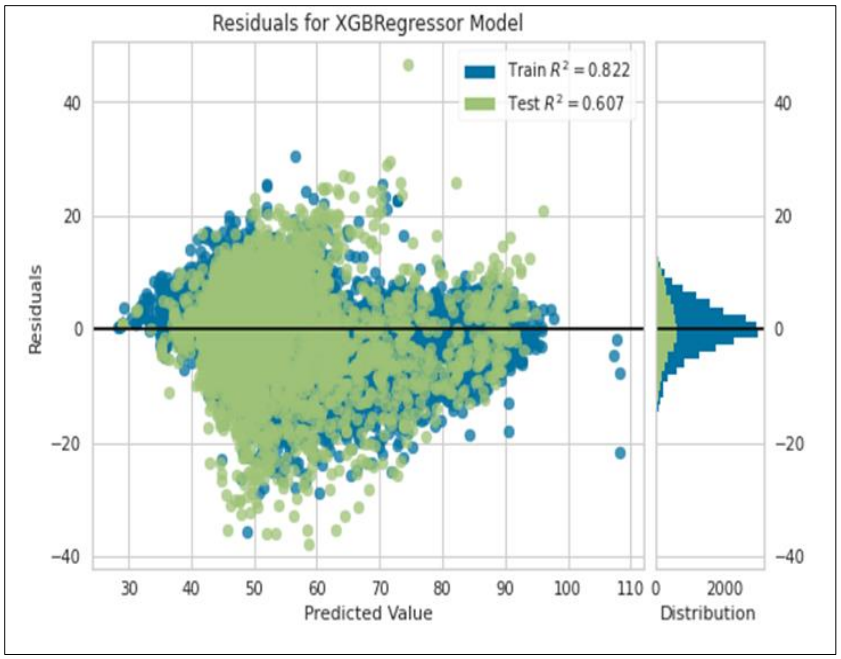

Figure 6: Analyse des résidus

2. Importance des features

L'analyse de l'importance des features montre que le pH, la séquence de type M et la longueur des protéines vont être un des points importants de notre modèle. La présence d'une séquence d'acide aminé de type A, K, Y, I, T, Q, H, C et L va avoir un rôle à jouer même si cela sera moins important (Figure 7). Le pH impacte positivement sur notre modèle (Figure 8).

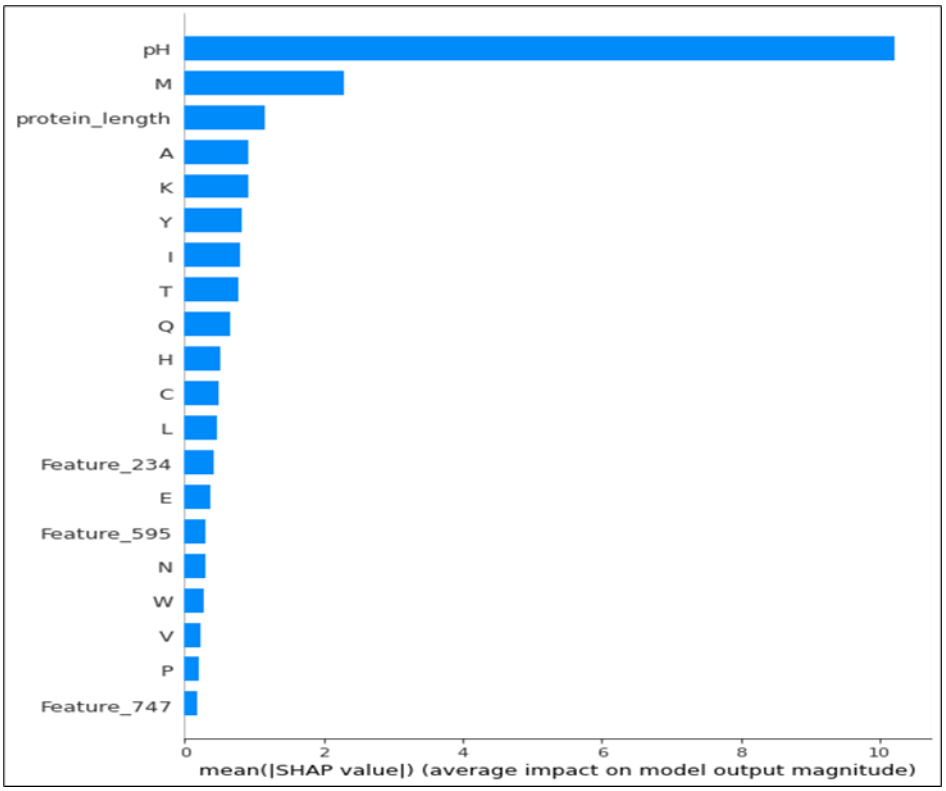

Figure 7: Importance des features via shap

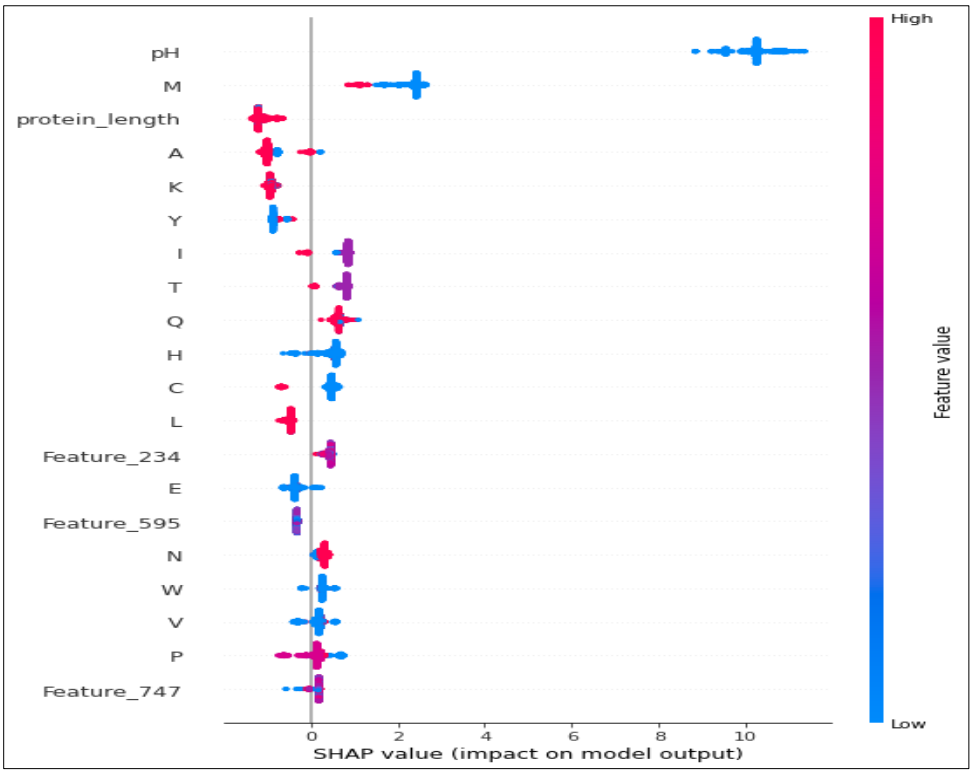

Figure 8: Impact des features

Tableau 1: Données de soumission

Seq id	tm preded
31390	63.644640
31391	63.644640
31392	63.194424
31393	60.611930
31394	63.672363

VI. Conclusion

Lors de ce projet j'ai participé à ma première compétition Kaggle, ce qui m'a permis de découvrir ce monde de la compétition. Le choix a été fait de baser mon travail sur ce qui a été proposé par d'autres compétiteurs afin d'améliorer les performances. Pour cela j'ai décidé d'une méthode méthodes de traitements de données et un type de modèle afin de faire de l'ensemble des prédictions. Cette stratégie a permis d'améliorer légèrement les performances, en passant d'un score de 0.613 à 0.62235. De mon point de vue, compte tenu de la taille de la base de données, je pense que pour obtenir des performances supérieures il est nécessaire de travailler sur des notebooks avec d'avantage de RAM. Lors de ma prochaine compétition, je pense qu'il serait important de rejoindre la compétition plus tôt afin d'avoir plus de temps pour développer des modèles. Également, il est important de choisir des compétitions avec des bases de données de tailles plus raisonnable pour éviter cette problématique de mémoire.

Références

[1] Novozymes, 2022 : Prédiction de la stabilité des enzymes, compétition

[2] Chris DEOTTE, Kaggle Grandmaster: XGBoost - 5000 Mutations 200 PDB Files lien kaggle

[3] Rishabh R Rahatgaonkar, Kaggle Contributor : Novoenzyme ProtBert + XGBoost lien kaggle

[4] AWS, Amazon : Fine-tune and deploy the ProtBERT model for protein classification using Amazon SageMaker ici

[5] Eodalisa, logiciel et étude : Coefficient de correlation sur les rangs (Rho de Spearman) ici

[6] Chimie générale : Thermostabilité, Polymères thermostables ici