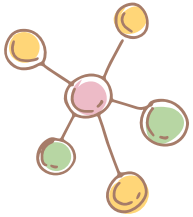


Prédiction de la stabilité des enzymes Novozymes

kaggle

LO OUSMANE, Machine Learning Engineer

Table of Contents

01 Présentation de la compétition

02 Exploration des données

03 Traitement des données

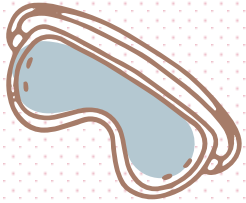
04 Modélisation

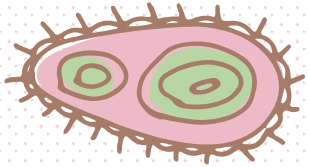
05 Soumission des résultats

06 Conclusion

01

Présentation de la compétition





Présentation de la compétition

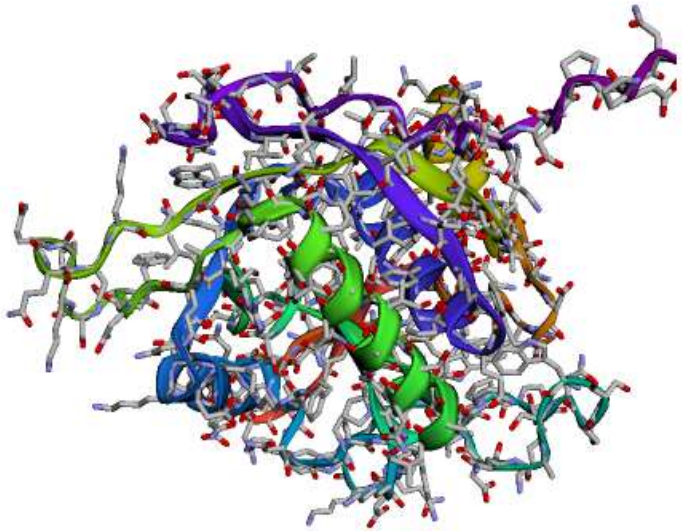
Contexte

- Novozymes leader mondial des solutions biologiques
  - Recherche des enzymes
  - Les optimize
  - Utilisation dans l'industrie

Objectif

- Prédire la thermostabilité de variants enzymatiques
  - Traitement des des données
  - Développer des modèles pour la prediction

Données



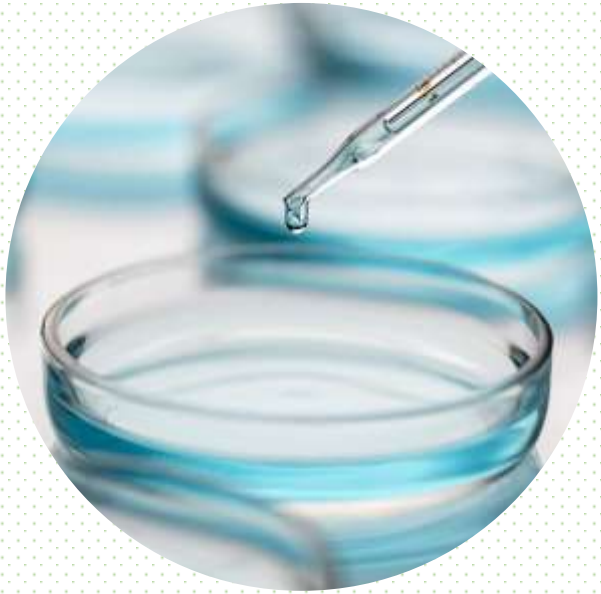
Structure tridimensionnelle de l'enzyme

On dispose des données:

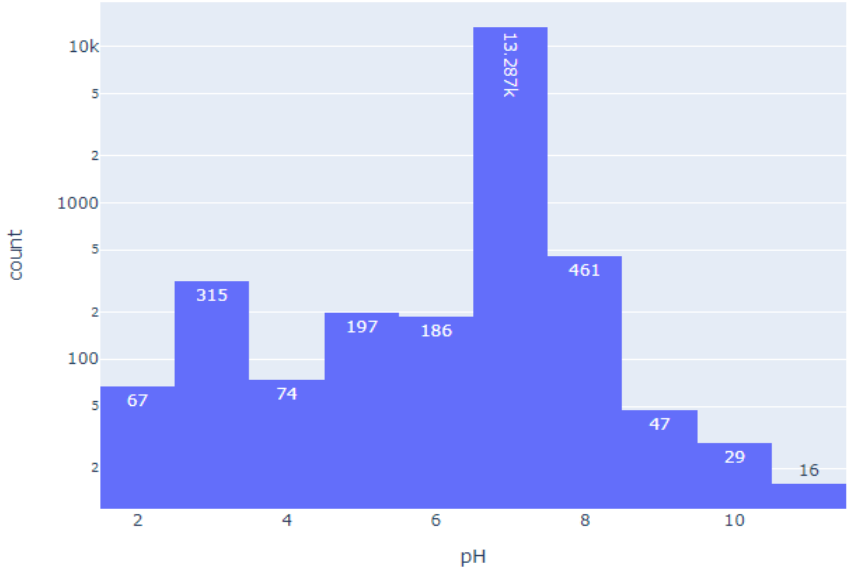
- Les données d'entraînement avec 5 variables:
  - seq_id, protein_sequence, pH, data_source
  - tm: colonne cible
- Les données d'entraînement mis à jour
- Les données de test
- Les données de soumission
- Les données pour la structure tridimensionnelle de l'enzyme

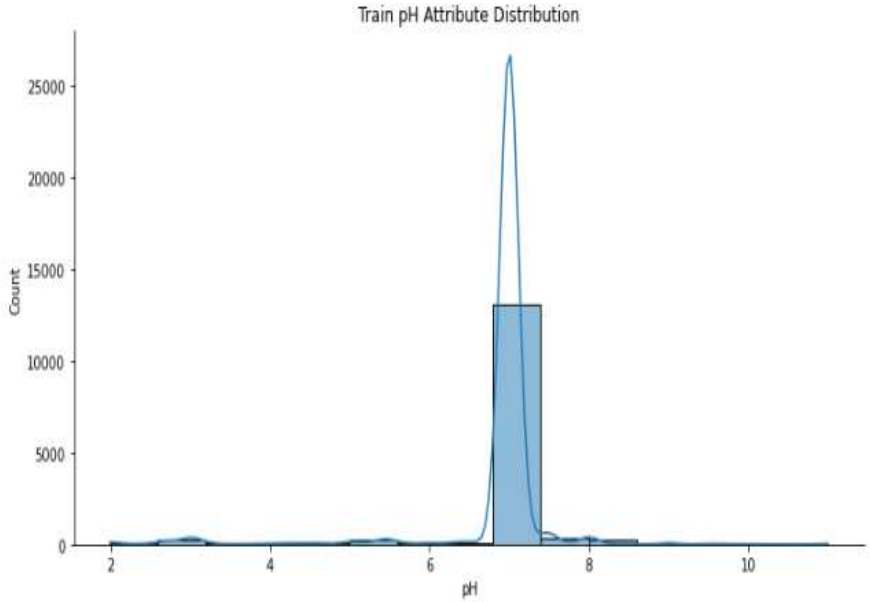
02

Exploration des données



Exploration

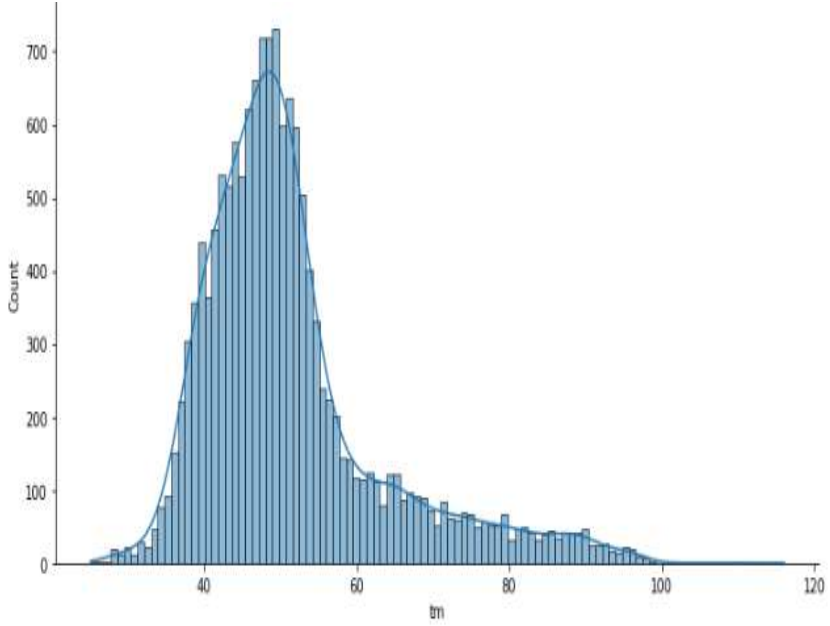




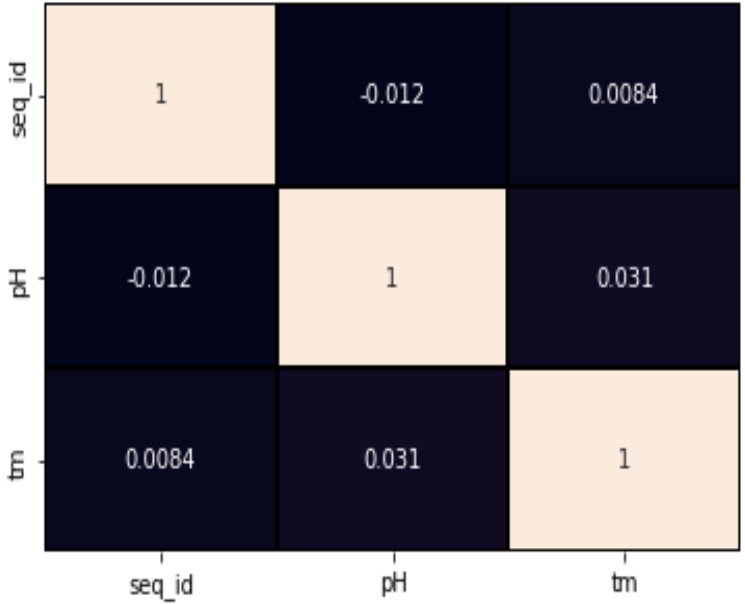
Distribution du pH

Exploration

Distribution du tm

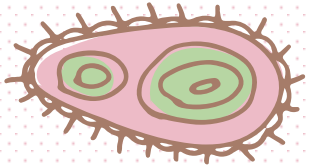


Corrélation des variables



03

Traitement des données



Feature extraction

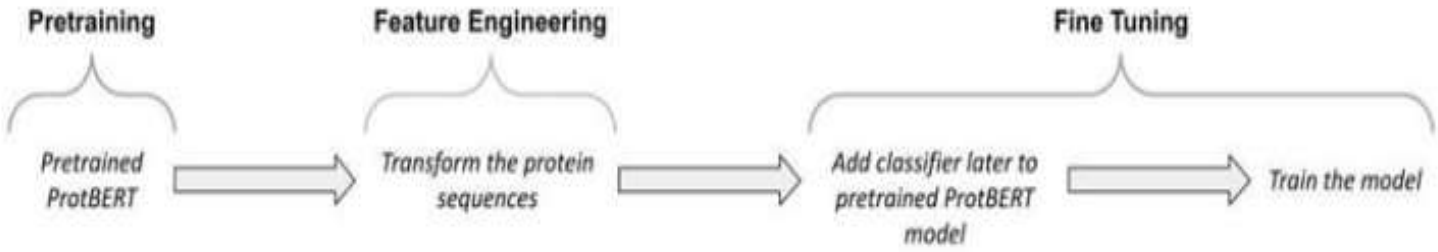
- Utilisons des techniques de PNL pour la classification des séquences de protéines.
- Interpréter les séquences protéiques comme des phrases et leurs constituants, les acides aminés, comme des mots
- Extraction avec Probert

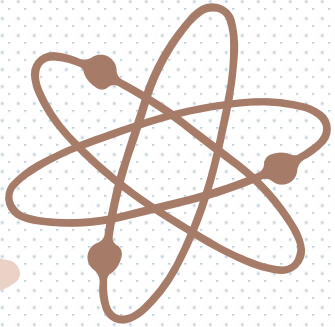
Probert Model

- Modèle pré-entraîné sur des séquences protéiques
- Basé sur le modèle BERT
- Uniquement sur les séquences de protéines brutes
- Aucun humain ne peut pas les étiquette

Architecture ProBERT

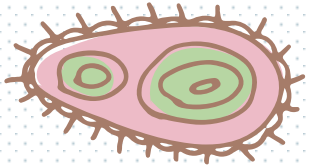
- Le message se concentre sur le réglage fin du modèle PyTorch ProtBERT (schéma suivant).
- Nous étendons d'abord le modèle ProtBERT pré-entraîné pour classer les séquences protéiques.





04

Modélisation

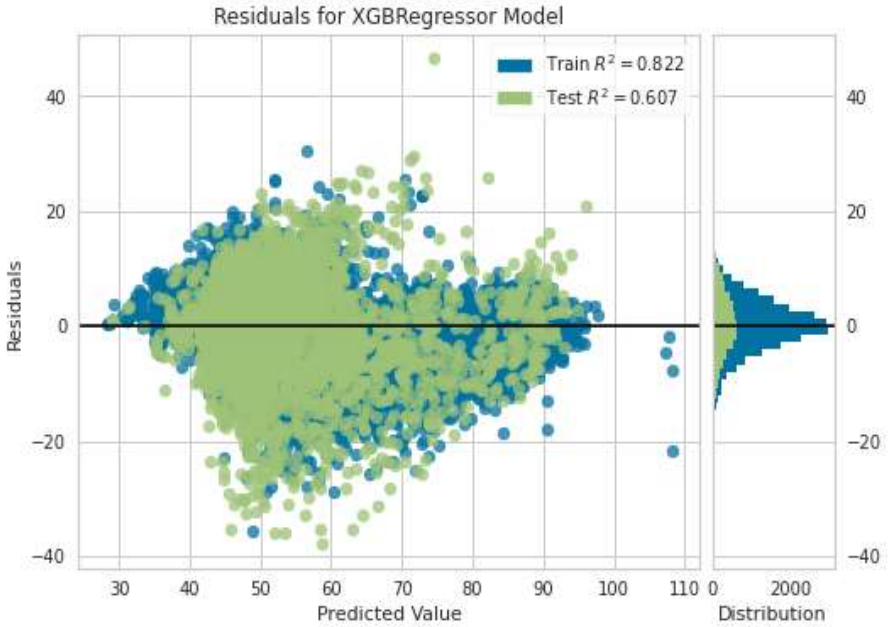


Création de Modèle

- Extraction de feature
- Feature Engineering
- Combiner les données obtenues avec extraction et celles du feature engineering
- Preparation des données d'entrainement, validation et celles de test.
- Choix de modèles pour la prédiction des tm
  - Xgboost modèle choisi
  - Validation croisée non envisager vu la taille des données
  - Pareil pour l'optimisation des hyperparametres

Résultats du modèle

Analyse des résidus

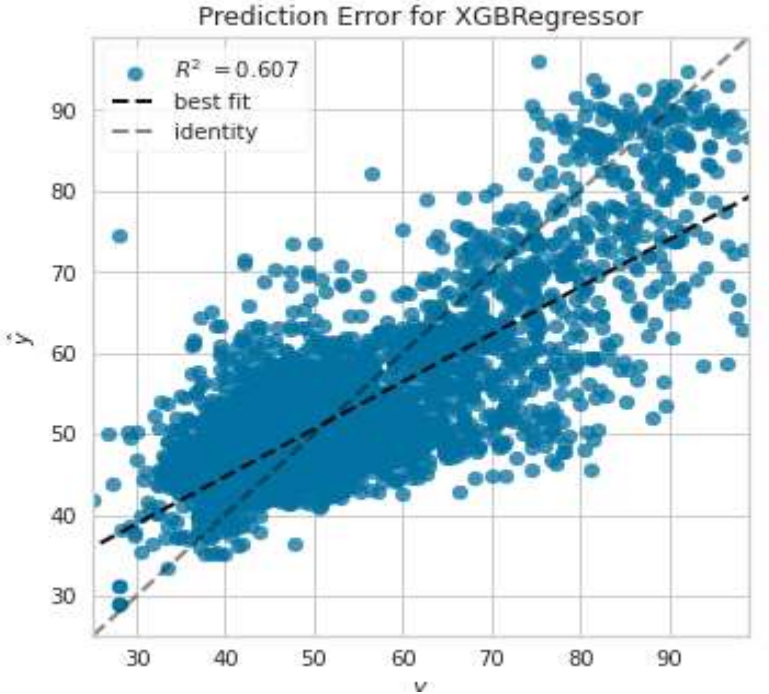


Evaluation des performances

- MAE train: 3.78
- MAE test: 5.58
- R^2 train: 82.2%
- R^2 test : 60.7%

Résultats du modèle

Prédiction correcte

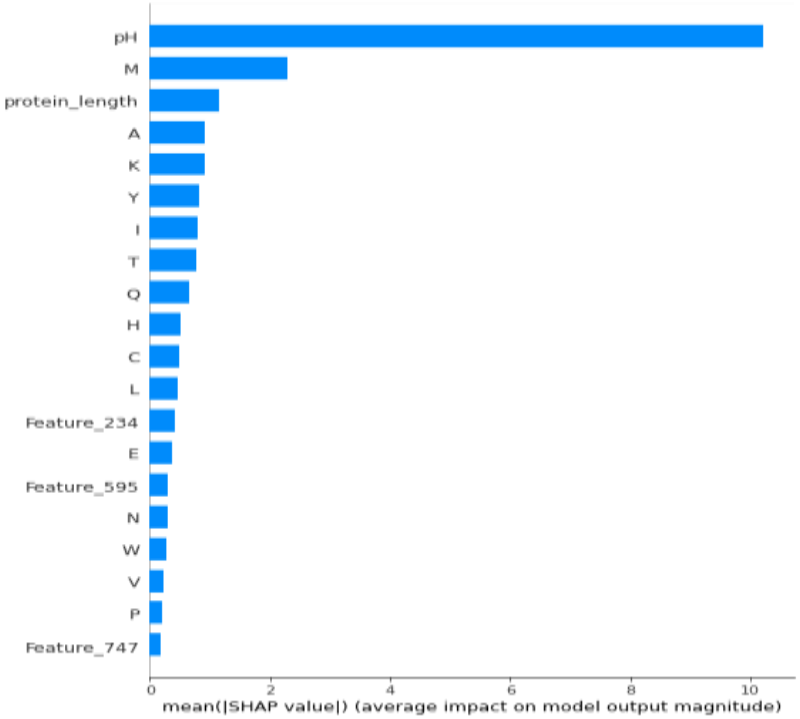


Corrélation

- Spearman train: 82.1%
- p-value : 0.0

- Spearman validation : 57.9%
- p-value : 0.0

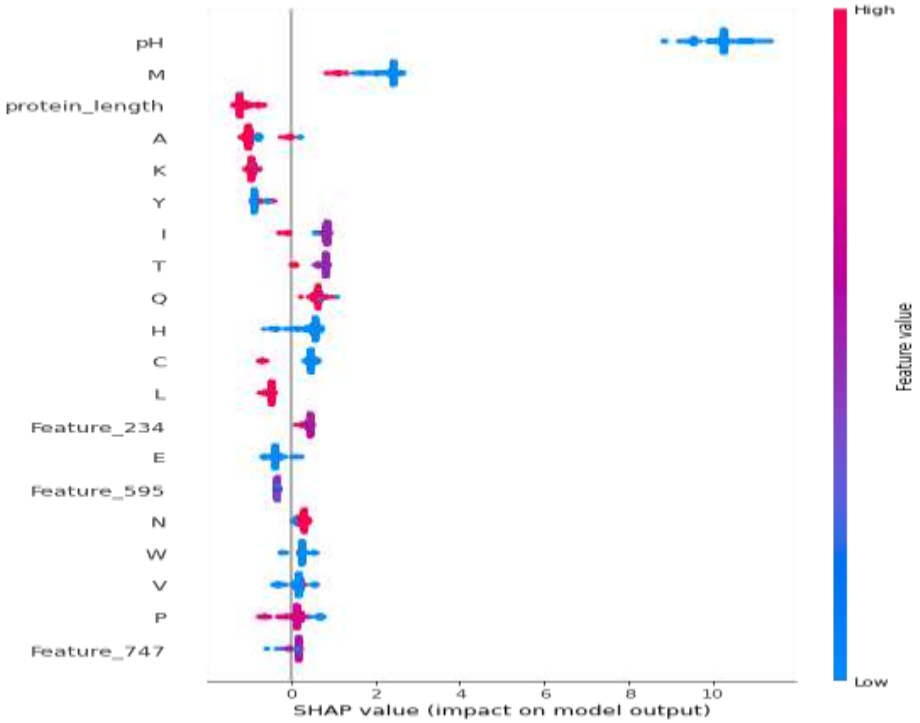
Features importance



- Le pH joue considerablement sur la prediction des tm
- La longueur des proteins participle faiblement sur la prédiction
- Les sequences A,K,Y ont un effet sur la prédiction

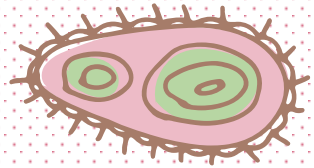
Features importance

- Le pH influe positivement sur la prediction des tm
- La longueur des proteins participle faiblement sur la prédiction



05

Soumission des résultats



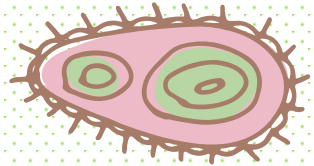
Données soumission

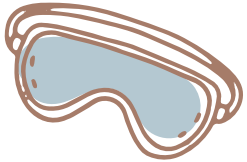
Seq id	tm pred
31390	63.644640
31391	63.644640
31392	63.194424
31393	60.611930

- Prediction des tm pas top
- Des erreurs de predictions trop grands
- Grande difference entre les tm preded et les valeurs de tm initiaux

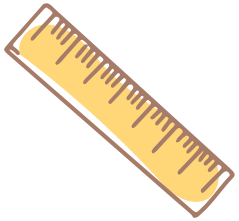
04

Conclusion





- Participation à ma 1ère compétition Kaggle
- Expérience avec les très grosses bases de données
- Amélioration du score d'un autre participant
- Améliorations à envisager
  - Intégrer la compétition assez tôt pour avoir le temps de développer plus en détails les modèles
  - Réaliser les calculs sur une machine avec plus de mémoire



Thanks!

Do you have any questions?
ousmanelo78@gmail.com
06 44 06 89 45

Please keep this slide for attribution

CentraleSupélec