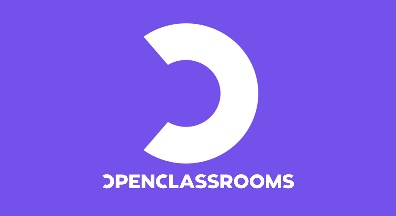
**** 

|  |
| --- |
|  |

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](#_Toc121996208)

[II. Présentation du projet 2](#_Toc121996209)

[1. Contexte 2](#_Toc121996210)

[2. Objectif 2](#_Toc121996211)

[III. La base de données 3](#_Toc121996212)

[1. Origine des données 3](#_Toc121996213)

[2. Préparation des données 3](#_Toc121996214)

[IV. Exploration des données 4](#_Toc121996215)

[V. Modélisation 5](#_Toc121996216)

[VI. Conclusion 6](#_Toc121996217)

[VII. Références bibliographiques 7](#_Toc121996218)

# 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 tabulées, la compétition « American Express - Default Prédictions » (lien ici) a été choisie.

# Présentation du projet

## 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.

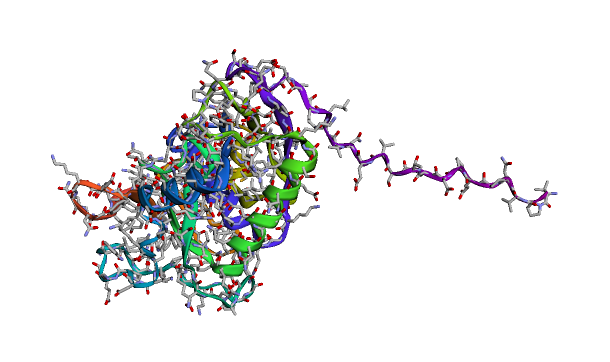
Dans ce concours, Novozymes nous invite à développer un modèle pour prédire/classer la thermostabilité des variants enzymatiques sur la base de données expérimentales de température de fusion, obtenues du laboratoire de criblage à haut débit de Novozymes. Vous aurez accès aux données des publications 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.

## 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.

# La base de données

## Origine des données



## Préparation des données

# Exploration des données

# 

# Modélisation

# Conclusion

# Références bibliographiques

[1] X. Huang, A. Khetan, M. Cvitkovic, et Z. Karnin, « Tabtransformer: Tabular data modeling using contextual embeddings », ArXiv Prepr. ArXiv201206678, 2020.

[2] A. Vaswani et al., « Attention is all you need », Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[3] G. A. Aguayo et al., « Comparison of Deep Neural Networks and Regularised Cox Regression Models in the Prediction of Neurodegenerative Diseases in the General Older Population », Available SSRN 4026085.

[4] H. Aboutalebi, M. Pavlova, M. J. Shafiee, A. Florea, A. Hryniowski, et A. Wong, « COVID-Net Biochem: An Explainability-driven Framework to Building Machine Learning Models for Predicting Survival and Kidney Injury of COVID-19 Patients from Clinical and Biochemistry Data », ArXiv Prepr. ArXiv220411210, 2022