**Codon usage**

A00188646 – Alex Savoie

Projet de session pour le cours INFO4301 Apprentissage machine – Automne 2021

Lien vers la présentation : …………………………

**Table des matières**

[**1.** **Introduction** 3](#_Toc86584679)

[**2.** **Base de données** 3](#_Toc86584680)

[**3.** **Algorithmes** 3](#_Toc86584681)

[**3.1** **Algorithme X** 3](#_Toc86584682)

[**3.2** **Algorithme Y** 3](#_Toc86584683)

[**3.3** **Algorithme ….** 3](#_Toc86584684)

[**4.** **Optimisation des hyperparamètres** 3](#_Toc86584685)

[**4.1** **Algorithme X** 3](#_Toc86584686)

[**4.2** **Algorithme Y** 3](#_Toc86584687)

[**4.3** **Algorithme ….** 3](#_Toc86584688)

[**5.** **Résultats** 4](#_Toc86584689)

[**6.** **Conclusion** 4](#_Toc86584690)

[**7.** **Références** 4](#_Toc86584691)

# **Introduction**

Mise en contexte et induction du sujet/objectif spécifique de ce projet

# **Base de données**

Notre base de données contient un total de 13028 données et de 69 colonnes. Nos 13028 données correspondent à 13028 différentes espèces vivantes uniques. Notre première colonne, ‘Kingdom’, correspond au règne associé à l’espèce. C’est cette colonne qui sera notre étiquette pour notre classification. Il existe 11 règnes différents possible : 'arc'(archaea), 'bct'(bactérie), 'phg'(bactériophage), 'plm' (plasmide), 'pln' (plante), 'inv' (invertébré), 'vrt' (vertébré), 'mam' (mammifère), 'rod' (rongeurs), 'pri' (primates), et 'vrl'(virus). La deuxième colonnes, nommé DNAType, est la composition génomique. En raison des objectifs de cette expérience, cette colonne sera tout simplement ignoré pour le projet. La colonne SpeciesID est le ID de l’espèce en question. Puisque toutes les données dans SpeciesID sont uniques pour chacune des espèces, cette colonne n’apporte aucun intérêt pour la création de notre algorithme. La prochaine colonne, Ncodons, corresponds aux nombres totaux de codons utilisé qui sert à calculer la fréquence de chaque codon dans une espèce spécifique. Encore une fois, cette colonne ne présente aucune utilité pour notre projet. La colonne SpeciesName donnent tout simplement le nom de chaque espèce. Pour la même raison que la colonne SpeciesID, elle est ignorée pour ce projet. Les 64 prochaines colonnes sont la fréquence des 64 codons en décimal avec 5 chiffres après la virgule. Les fréquences varient de 0 à 1. Donc, en théorie, la somme de la fréquence des 64 codons devrait être égale à 1. Ce sont ces 64 colonnes qui seront utilisé comme caractéristique pour nous aider à prédire le règne d’une espèce.

Dans la base de données, il avait deux données corrompues qui causait problème. Puisque notre base de données contient déjà un grand nombre de données, ces deux données ont tout simplement été laissé tomber. Un autre problème que présente notre base de données est l’inégalité de la distribution de nos classes. Certaines classes, comme celles des bactéries et des virus, contiennent des milliers de données alors que des classes, comme celle des plasmides contient seulement une dizaine de données. Cela pourrait causer certains problèmes lors de l’analyse de nos résultats plus tard. Le graphe ci-dessous démontre bien cette inégalité.

GRAPHE

Avant d’implémenter les différents algorithmes, il important de mentionner que les données ont été divisé de façon que 80% de celles-ci seront utilisé pour le training et l’autre 20% pour le testing. Les données ont par la suite été standardiser en utilisant le StandardScaler fournit par ScikitLearn , (z = (x-u) / s).

# **Algorithmes**

## **Régression logistique**

Pour débuter, la régression logistique était notre premier algorithme à essayer. Pour sa première exécution, j’ai conservé les paramètres par défaut de ScikitLearn. La régression logistique nous donnait une exactitude d’environ 86%. En observant le graphe ci-dessous, on aperçoit que l’exactitude converge après une dizaine d’itération.

GRAPHE

Essayons d’autre algorithme en espérant obtenir des meilleurs résultats.

## **Random Forest**

Le deuxième algorithme implémenté est le classificateur Random Forest. Celui-ci nous amène à un meilleur résultat que la régression logistique. Notre exactitude est augmentée à 89,2% et notre F1-score est passé de 86 à 89%. À noter que, encore une fois, les paramètres par défaut ont été choisi.

## **Support Vector Machine (SVC)**

Encore une fois, notre score à continuer de s’améliorer. Avec le SVC, notre accuracy a augmentée à 93,3% et notre F1-score à 93%. Il est aussi important à noter, que jusqu’à maintenant, aucun de nos algorithmes a réussi à bien classifier un seul plasmide. Rappelons-nous que les plasmides était l’une des classes les moins fréquentes dans notre base de données. Donc, ce n’est pas vraiment surprenant.

## **Neural Network**

Le réseau neurone démontrait des résultats assez similaires au SVC. Celui-ci avait une exactitude de 92,8% et un F1-score de 93%. Cependant, le réseau neurone à mieux réussi à classifer les classes ayant beaucoup de données comme les bactéries et les vertébrés. Le réseau neurone était le premier algorithme à réussir à classifier un plasmide.

## **KNN (k-nearest neighbor)**

Finalement, le dernier algorithme testé fut le KNN. Ce dernier présentait des résultats très similaire au réseau neurone avec une exactitude de 92,8% et un F1-score de 93%.

# **Optimisation des hyperparamètres**

## **Algorithme X**

Décrire les optimisations des hyperparamètres effectuées pour cet algo et les résultats préliminaires associés. Donner suffisamment de détails pour démontrer le travail effectué.

## **Algorithme Y**

Décrire les optimisations des hyperparamètres effectuées pour cet algo et les résultats préliminaires associés. Donner suffisamment de détails pour démontrer le travail effectué.

## **Algorithme ….**

Décrire les optimisations des hyperparamètres effectuées pour cet algo et les résultats préliminaires associés. Donner suffisamment de détails pour démontrer le travail effectué.

# **Résultats**

Afficher; tableau comparatif des meilleurs résultats par algo, courbes

Important : discuter, analyser et comparer vos meilleurs résultats

# **Conclusion**

Conclusion décrivant les résultats et démontrant votre compréhension du travail effectué.

# **Références**

1. Référence 1
2. Référence 2
3. ……..