# Notes de lecture pour travail en cours sur modèle de ML développés pour la reconnaissance des catégories de plancton

1. Analyse et représentation.

Fichiers : “CC\_analysis.jpnb” puis “trophie\_filtered.jpnb”

Fichier de première approche et analyse des données présentées, notamment des CC.

> les lignes

Visualisation des données. Pose des différentes problématiques.

Identification des marqueurs CC qui ne sont présents que dans un seul mode trophique (phototrophe, mixotrophe (CM) et zootrophe).

1. Préprocessing + ML log et lasso

Fichier : Full\_data Add\_label (changer le nom…)

Dans ce fichier, on retrouve l’étape complète de pré – processing (dont la mise en place des étiquettes) + modèle ML régression logistique avec régularisation lasso L1

Pre-processing

2.1 labélisations des matrices.

Librairie utilisée : pandas et numpy.

Prendre l’ensemble d’une matrice et concaténer les labels y correspondants, ces étiquettes se trouvant dans un autre fichier, vérifier la correspondance exacte de ces labels.

En effet, suivant les matrices, leurs tailles sont différentes et nous pourrons trouver des éléments non correspondants entre les deux matrices (variables X et labels y).

Main file :

Full data\_add\_label

Input :

La matrice que nous voulons traiter, par exemple « abundance\_matrix\_80\_100.RED.csv »

Étiquettes : « target\_name.csv » (il s’agit juste de la colonne extraite du fichier des méta-données)

output :

Étiquettes triées : « target\_filtered.csv »

Fichier de données avec labellisation triée : « data\_filtered »

* 1. En option : labellisation par regroupement de labels, notamment des mixotrophes, ainsi que tri

Et tri sur sur la représentation des CC par échantillons.

Par exemple : création des matrices 3 CC et 4 CC qui ne va garder que les CC qui sont présents 3 X et plus, et 4 X et plus.

(permet d’affiner l’approche, un peu plus de pertinence et de réduire la taille des matrices).

Trophie filtered. jpnb

somme\_trophie\_fil\_transposed.jpnb

matice\_3C\_treatment.jpnb

pour les créations des matrices 3cc et 4cc. Optionnelles.

2.3 Application du modèle de ML régression logistique avec régularisation en Lasso.

fichier : Full data\_add\_label

Soit on utilisera les fichiers qui viennent d’être créés, soit importation possible de fichiers précédemment créés.

1. Modèle Decision Tree classifier et Random Forest avec PCA.

Fichier: forest\_M3\_decision\_tree. jpnb

Import des fichiers : matrice\_3C\_treated.csv

Fichier pré-processé avec filtrage des CC à 0 représentants et réduction en gardant les CC représentés += 3 X

Le premier modèle qui est un Decision Tree Classifier simple sert de modèle de base et de comparaison pour les autres modèles qui seront traités à la suite. C’est-à-dire que leur hyperparamètre doit permettre un meilleur score.

Deux modèles suivent de Random Forest suite à une réduction en PCA.

Avec deux méthodes de recherches en hyperparamétrages. (GridSearch CV et RandomizedSearchCV)

1. A rendre : un document word exposition et développent de la problématique.

+ conclusion / piste d’amélioration.