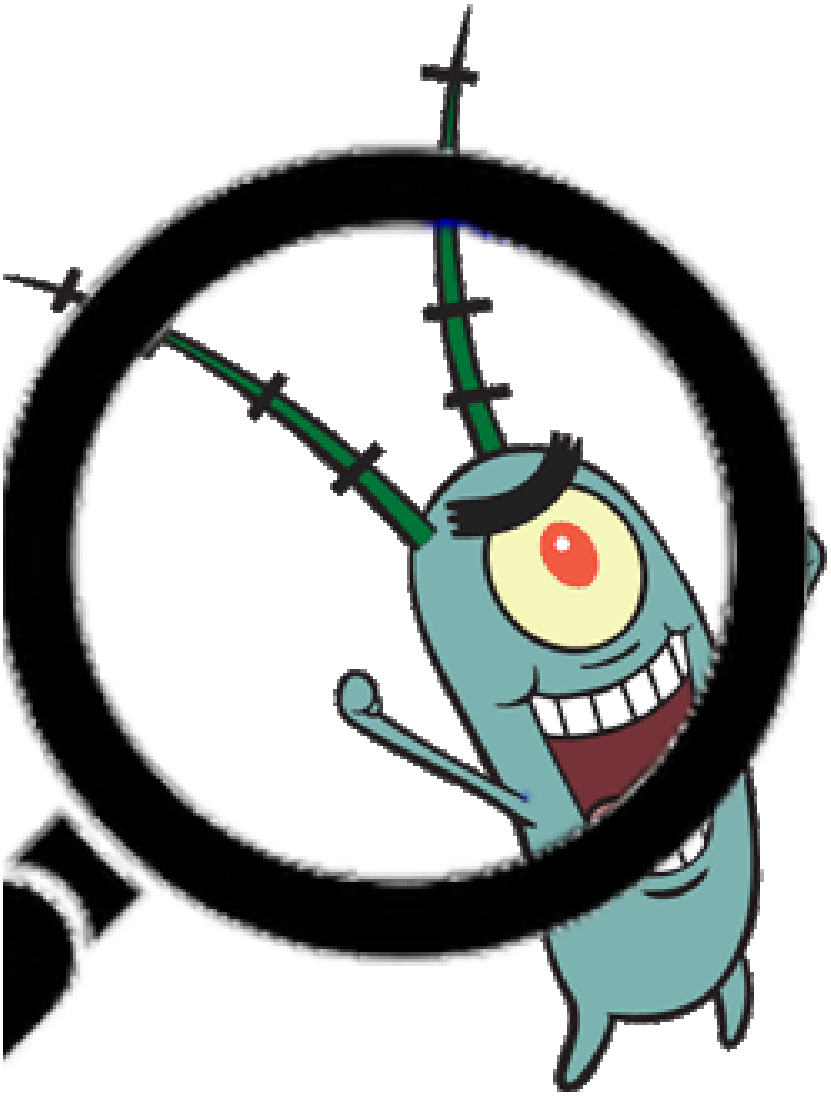
GAIASAVERS

Rapport Groupe GREENFORCE

Membres : *Léo RESSAYRE, Riyad MEDJADI, Céline YAN, Circé CARLETTI*

URL du challenge : https://codalab.lri.fr/competitions/623

Repo GitHub du projet : https://github.com/ShionMirai/GREEN



CONTEXTE ET DESCRIPTION DU PROBLÈME

Sur Terre, plus précisément sous l'eau, il existe une très grand diversité des espèces. Il faut les protéger et c'est pour cela que nous étudions les planctons et nous déterminons la variété des espèces dans chaque endroit. En effet, les planctons sont un très bon indicateur de la biodiversité.

Notre problème est un problème de classification multiclasse. Nos données sont des photographies de plancton.

Les deux parties que nous avons effectuées en scindant le groupe de 4 en 2 sont le preprocessing qui prépare les données à leur utilisation par le classifieur, et le travail sur la sélection du modèle.

PREPROCESSING : Céline YAN & Riyad MEDJADI

Pour la partie preprocessing du projet, qui a pour but d'optimiser l'efficacité du classifieur en modifiant les données, nous avons commencé par réfléchir à un moyen de détecter les données aberrantes ("outliers"), qui ne sont pas pertinentes. Pour cela, nous avons choisi d'utiliser l'algorithme "Isolation Forest" pour prédire ces exceptions. Il a fallu réduire le nombre de dimensions des données X_train jusqu'à deux dimensions uniquement, à l'aide de l'algorithme Principal Component Analysis (PCA). Pour l'affichage des exceptions, nous avons commencé par choisir d'utiliser mathplot, mais nous avons finalement préféré l'utilisation de scatterplot.

En faisant ensuite varier le nombre de dimensions, nous avons pu conclure que le meilleur score possible (qui était d'environ 0.6108) était atteint lorsque la dimension était égale à 50. L'objectif ici était de prédire Y_hat_test.

Enfin, la dernière partie du travail consistait à enlever les fonctionnalités superflux dans les données, en faisant une sélection des fonctionnalités, puis de mettre en valeur les fonctionnalités les plus importantes.

MODÈLE : Circé CARLETTI & Léo RESSAYRE

Afin de mener à bien notre partie du projet, nous avons décidé d'effectuer notre travail en deux parties. Nous avons d'abord cherché à faire fonctionner quelques modèles, afin de nous approprier le code donné, mais aussi de voir le type de résultats que nous

pouvions obtenir avec différents types de modèles. Ensuite, nous avons créé une boucle permettant de faire fonctionner de nombreux modèles les uns après les autres, dans le but de trouver celui qui donnerait le meilleur résultat.

Ainsi, nous avons commencé par tester différents modèles (Nearest Neighbors Classification, One-by-one, Classification Decision Tree, AdaBoost), et repérer quels « Types » de modèles seraient les plus prometteurs : les modèles de type ensemble.

Par la suite, nous avons défini plusieurs modèles et les avons tous fait fonctionner les uns après les autres, puis avons affiché les résultats de chacun, ce qui nous a permis d'en extraire le modèle le plus efficace. C'est finalement en modifiant le paramètre de ce modèle que nous avons pu obtenir nos résultats finaux, et décider du modèle ainsi que des paramètres à utiliser.

RÉSULTATS PRÉLIMINAIRES

Le score obtenu au final (entre 0.6 et 0.7) est assez acceptable en conclusion des deux parties effectuées, même s’il peut encore être amélioré davantage.