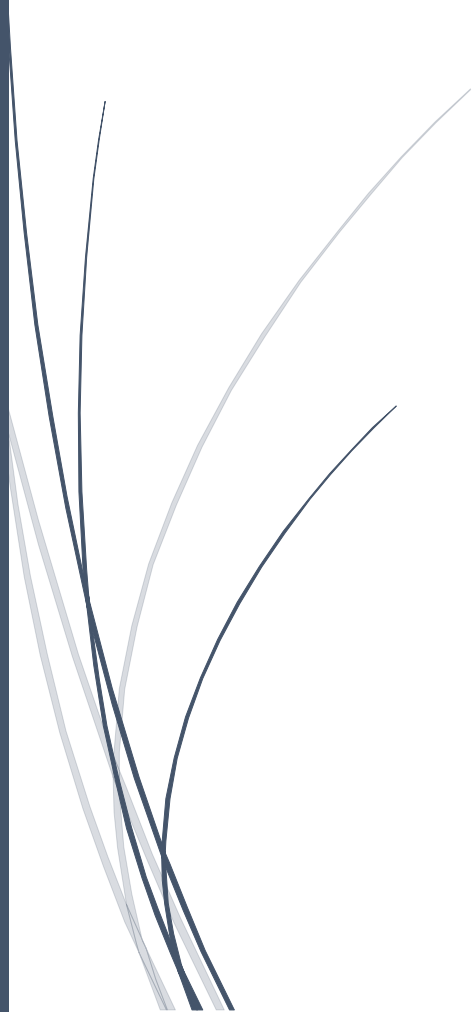
2018/2019

ETUDE DE CAS 1

Classification de phytoplanctons

FOUTSE YUEHGOH
CHIBANE LYDIA



Introduction :

Après recherche bibliographique sur le sujet des phytoplanctons, on a trouvé que : « chaque classe de phytoplancton est efficace à des longueurs d'onde différentes »,

Ainsi, en se basant sur cette information, on a proposé de travailler sur la problématique suivante :

Classification par sélection de variables pour confirmer ou infirmer cette hypothèse.

1. Importance des méthodes de sélection de variables :

On pourrait penser que plus on augmente le nombre de variables décrivant chaque observation de l'échantillon, plus on dispose d'informations concernant ces observations et plus on en facilite et on en améliore la classification. Cependant, la qualité de la classification ne dépend pas du nombre d'informations à disposition mais de la pertinence de ces informations.

Parmi les variables à disposition, il s'avère souvent que seules certaines d'entre elles contiennent la structure d'intérêt des observations. En d'autres termes, ces variables pertinentes suffisent à distinguer les différences de caractéristiques entre les observations et à les regrouper en classes.

2. Enumération des différentes méthodes de sélection de variables :

Deux types de procédure de sélection de variables existent :

- Les méthodes "filter": où la sélection se fait en amont du processus de classification.
- Les méthodes "wrapper": la sélection de variables est insérée au sein du processus de classification. Ces méthodes ont l'avantage de ne pas dissocier le problème de sélection de variables et classification pour mieux appréhender et interpréter les variables et leurs rôles.

On peut en citer certaines de ces méthodes pour la sélection de variables et qu'on utilisera pour notre étude :

a. Missing value ratio :

Ayant une base de données, en préalable de toute étude, on doit faire une analyse exploratoire. Si lors de cette étape, on trouve des données manquantes, on a recours à plusieurs solutions :

- Faire l'étude en gardant l'échantillon tel qu'il est avec les données manquantes.
- Eliminer la variable qui a beaucoup de valeurs manquantes.
- Mais si la variable avec beaucoup de valeurs manquantes est cruciale à l'étude, on supprime les individus ayant des données manquantes.
- Faire l'imputation qui est de remplacer ces données manquantes (nan) par des moyennes par exemple ou tout autre indicateur.

On pourra fixer un seuil précis sur le pourcentage de données manquantes pour une variable, si une des variables dépasse ce seuil en nombre de données manquantes, on pourra l'éliminer sous hypothèse qu'elle ne rapporte pas autant d'informations que les autres.

b. Low variance filter :

Si une variable a une variance proche du zéro donc elle n'améliore pas le modèle donc on pourra l'éliminer.

c. High correlation filter :

S'il y'a une forte corrélation entre deux variables, ceci signifie qu'elles ont des tendances similaires et peuvent porter la même information.

d. Random forest :

Un des algorithmes les plus utilisés pour la sélection de variables (feauture selection).

Il est livré avec une fonctionnalité intégrée, et donc pas besoin de programmer ceci séparément.

Cet algorithme nous aide à sélectionner un plus petit sous-ensemble de variables.

e. Backward feature elimination :

Le principe de cette méthode est le suivant :

- Faire la classification en utilisant toute les variables et voir les performances du modèle.
- Calcul des performances du modèle précédent mais dont on élimine une des variables, on fait ceci pour chacune des variables (une à la fois) donc ayant p-variables, on fera p-apprentissage.
- Choisir la variable pour laquelle les performances sont restées les mêmes que le modèle où on a utilisé toutes les variables. On l'élimine.
- On répète ce processus jusqu'à ce qu'aucune variable ne peut être éliminée.

f. Forward feature elimination :

Ce processus est l'opposé du précédent :

- On fait l'apprentissage en utilisant une variable à la fois, donc p-apprentissages.
- On choisit la variable qui donne de meilleures performances.
- A chaque fois, on rajoute une variable à la fois et en sélectionner le couple de variables qui donne les meilleures performances.
- On répète ce processus jusqu'à ce qu'on obtienne une évolution dans la performance pas très importante.

3. Application des méthodes de filtrage pour la sélection de variables sur la base de données relative aux phytoplanctons :

Notre objectif d'étude c'est d'aboutir à une classification cohérente avec ce qu'on a déjà dans notre base de données en n'utilisant que les longueurs d'ondes. Pour cela, on applique les différentes méthodes précédentes pour la sélection de variables pour en déduire si vraiment les longueurs d'ondes seules sont pertinentes pour la classification.

a. Missing value ratio :

Dans ce cas de figure, on ne sait pas si les variables ayant un grand pourcentage de données manquantes sont cruciales ou pas, donc cette méthode est à écarter de l'étude.

b. Low variance filter :

On travaille sur les données imputées, on calcule les variances des différentes variables et celles ayant une variance proche du zéro peut être éliminée. On trouve dans ce cas que la SST est la variable qui a la plus petite variance par rapport aux autres.

c. High correlation filter :

Les plus fortes corrélations qu'on trouve (étude bidimensionnelle faite dans le jalon précédent) sont celles des longueurs d'ondes entre elles. Ceci dit elles peuvent porter des informations de tendances similaires.

Notre travail consiste à chercher si on peut réduire les différentes variables mises à disposition en n'utilisant que les longueurs d'ondes pour la classification. On remarque que les méthodes élaborées jusqu'à présent nous informent que la variable SST peut être éliminée en utilisant « variance low filter » et aussi en utilisant le « high correlation filter », on peut en éliminer une des longueurs d'ondes. On va utiliser d'autres méthodes qui nous montrent l'importance de chaque variable pour la classification des phytoplanctons.

Dans la suite, nous allons faire la classification avec toutes les variable avec différentes méthodes et évaluer les performances, puis faire la sélection de variables et évaluer encore afin d'observer l'évolution.

4. Classification des phytoplanctons et sélection de variables

Nous allons utiliser les données de la première semaine comme notre train data set, et celle de la deuxième semaine pour tester le modèle ceci avec différentes méthodes :

- Données de Test et apprentissage :

On va travailler sur les données par semaine pour manque de ressources de calcul. Donc la première semaine sera notre base pour l'apprentissage et une autre semaine comme base de test (ce choix est aléatoire).

- La classification :

✓ Classification par arbres de décision :

L'arbre de décision n'est autre qu'un ensemble de règles pour la classification où la décision de la classe à laquelle appartient l'individu est définie par une suite de tests associés aux attributs, les tests étant organisés de manière arborescente.

Après apprentissage, on visualise les performances du modèle obtenu, sur les deux ensembles de test et apprentissage, comme suit : (dans ce cas la base test 1 c'est la 2éme semaine de l'étude)

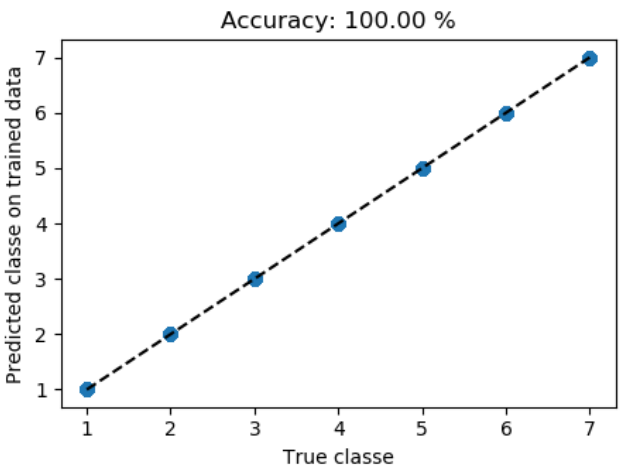


Figure 1. Performances sur la base d'apprentissage

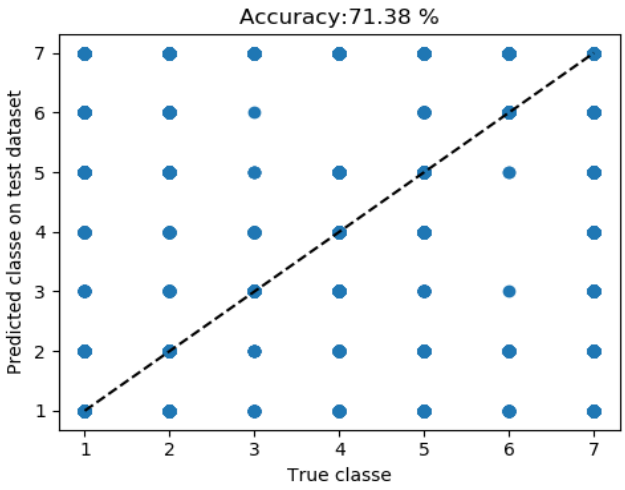


Figure 2. Performances sur la base de Test 1

✓ Classification en utilisant le Random Forest :

Dans cette partie, on construit un modèle random forest pour classifier les données selon 7 types de phytoplanctons, pour la base d'apprentissage, on a les performances suivantes :

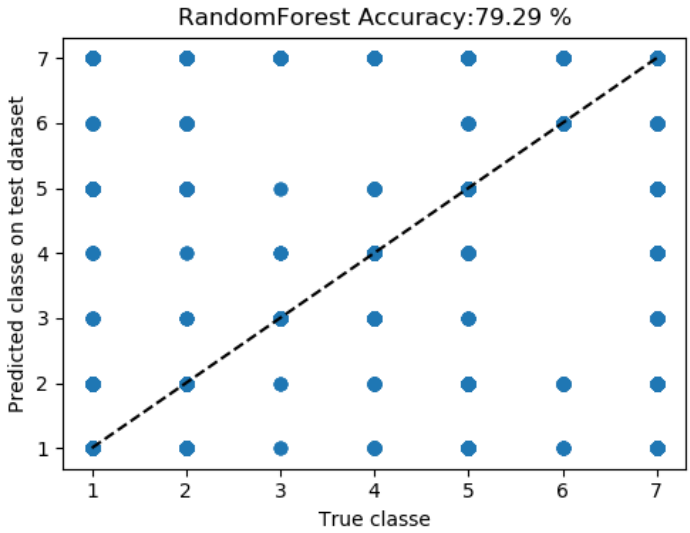


Figure3. Performance en base Test 1 en utilisant le Random Forest

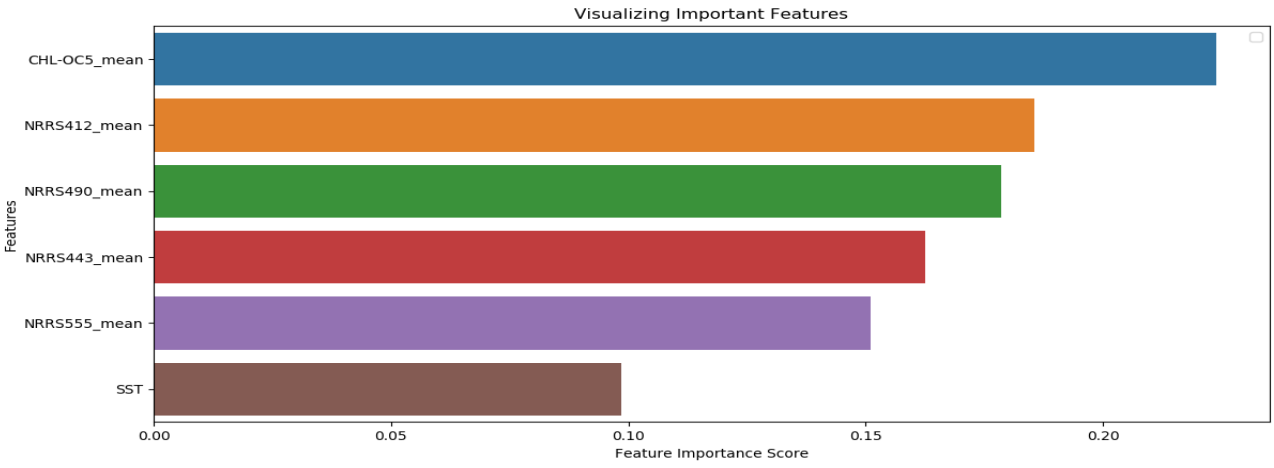
Dans ce cas de figure et pour cette base de test, on a eu des performances de 79.29 %. On continu avec cette méthode pour faire la sélection de variables et voir si on aboutit à une réponse à notre problématique.

✓ Sélection de variables :

Pour faire la sélection, nous avons utilisé :

> ExtraTreesClassifier : Cette classe implémente un méta-estimateur qui adapte des arbres randomisés à certains sous échantillon de la base de données à étudier. Et utilise le calcul de moyenne pour optimiser la précision de prédiction et contrôler le sur apprentissage.

Avec cette méthode, on obtient le graphique suivant où l'on remarque que la variable SST est la moins importante :



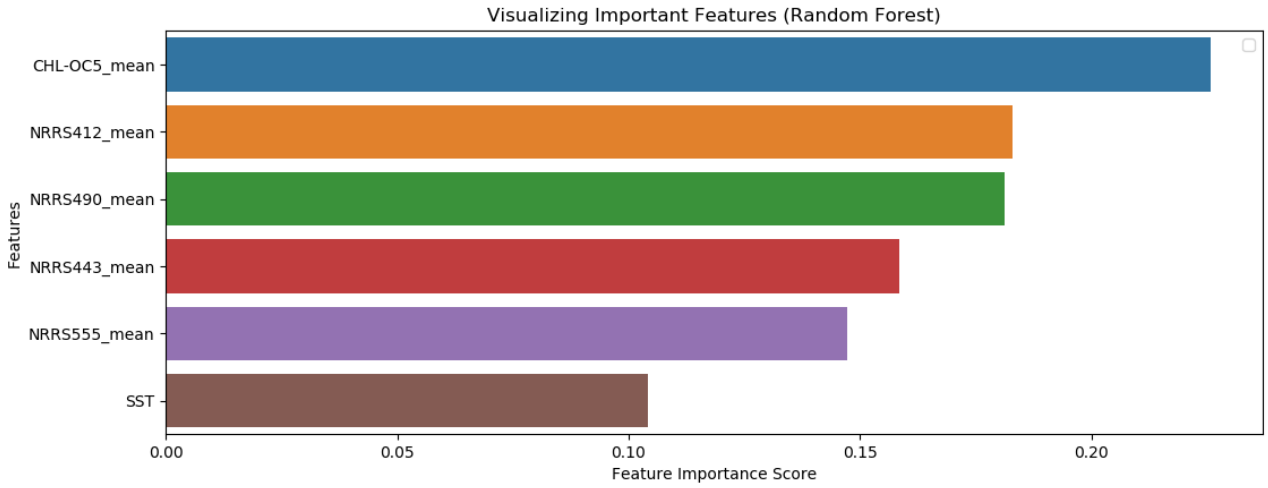


Figure4. Importance des features

En partant de ces résultats, on va exclure la variable SST et faire un autre apprentissage (principe de la méthode backward)

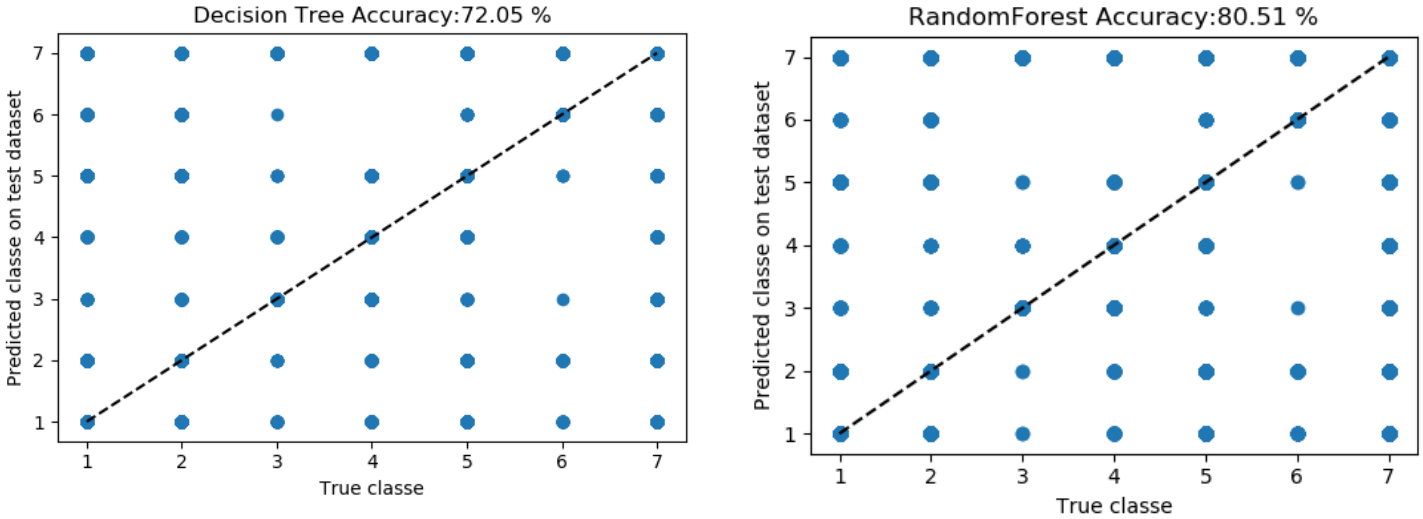
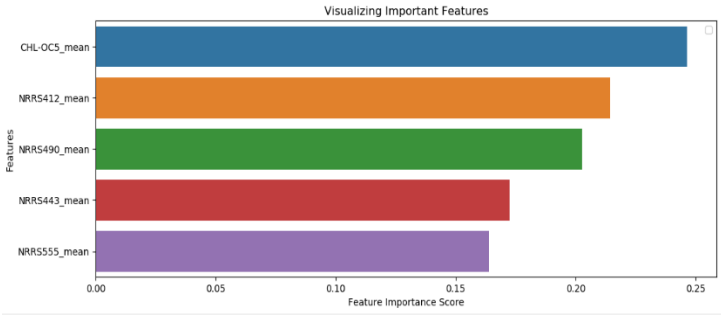
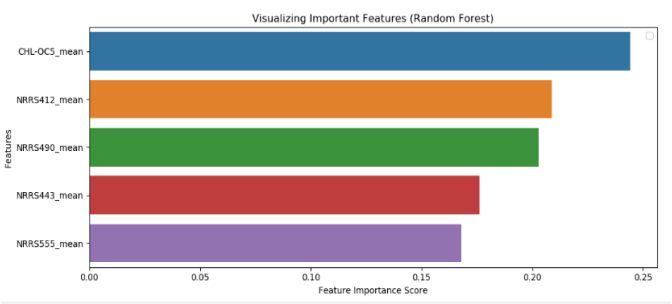


Figure5. Les performances sur les bases de test

On remarque une amélioration de la classification 2% pour Random Forest et 1% pour Arbre de décision

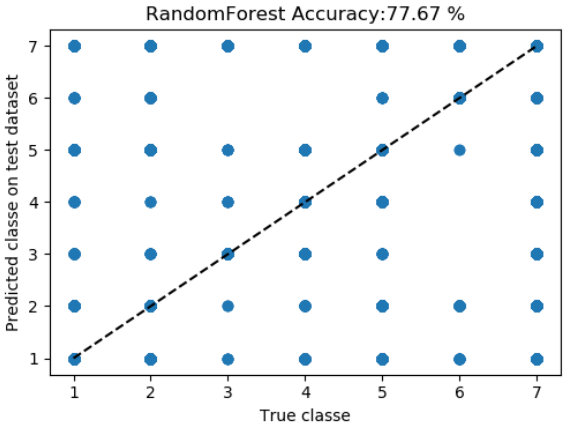
Même procédure que tout à l'heure, on retrouve ces résultats :



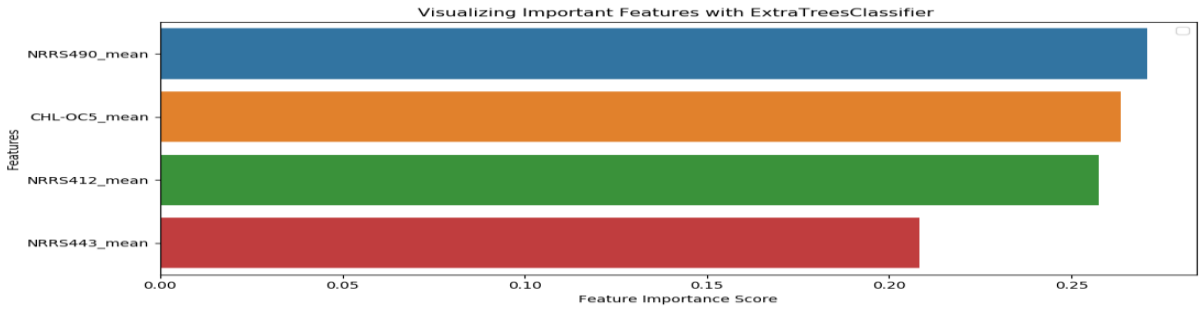


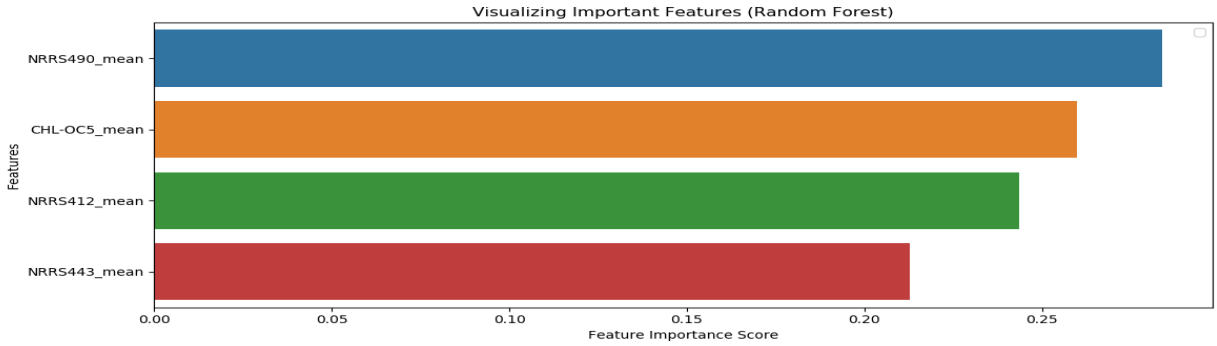
A partir de ces graphes, on en déduit que la longueur d'onde NRRS555 est la moins importante.

Nous avons donc retiré une longueur d'onde comme suggérer par la sélection de variable, on obtient les performances suivantes avec nos différentes méthodes :



On remarque une perte de performance.

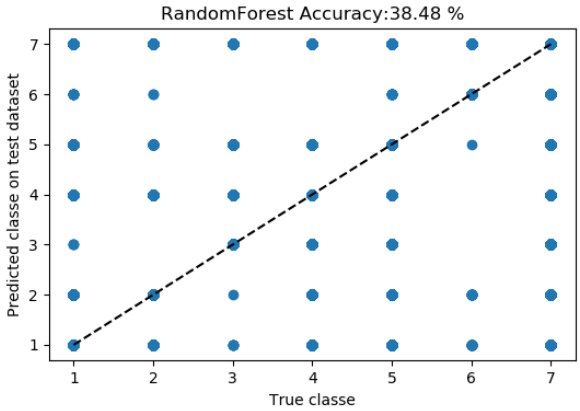




Avec ce qu'on observe sur les graphiques après élimination d'une première longueur d'onde, on retrouve que la chlorophylle est une variable très importante pour la classification et qu'on ne peut s'en passer pour cette étude. Et donc comme réponse à notre problématique, les longueurs d'ondes seules ne suffisent pas pour identifier différents types de phytoplanctons.

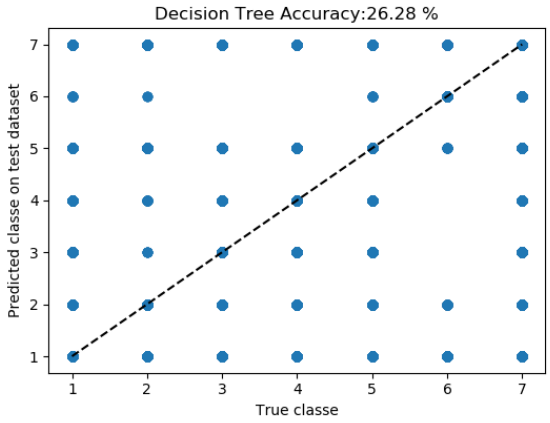
Dans cette étape, nous allons prendre une autre base de données de test et on prend la 20ieme semaine.

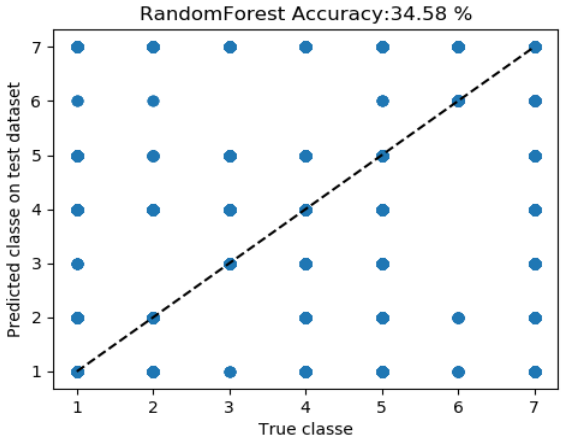
Cette première figure représente les performances sur cet ensemble test après un apprentissage fait en utilisant toutes les variables



On remarque bien une chute énorme en performance, voir on ne peut pas généraliser le modèle.

On va retirer SST et refaire notre classification et on obtient :





La suppression de SST semble ne pas être une bonne idée ici.

Cette méthode de classification n'est pas bonne et on pourrait penser que ces mauvaise performances sont due à :

- Données d'entrainement sur une courte durée, semaine 1 et test sur semaine 20. Dans ses deux période, il y'a surement eu une évolution énorme du phytoplancton, ainsi en une semaine on ne peut pas avoir toute les données pour entrainer notre modèle à connaitre tos les patterns.
- La suppression de données manquante pourrais résulter a diffèrent endroit du aux fait que les nan ont aussi été présent a plusieurs endroits differents.

Conclusion :

La méthode du random forest sur des données de courtes durées donne de meilleures performances que l'arbre de décision. Elle nous offre un classifieur avec de bonnes performances, et nous fait également une bonne sélection de variables, pour trouver réponse à notre question « est-ce-que les 4 longueurs d'ondes de notre base de données suffisent pour classifier les phytoplanctons ? ». On a trouvé que la variable chlorophyle-CH5 est très importante pour cette étude voire c'est la variable la plus informative sur les classes. Mais, on pourra s'en passer de la variation de température pour cette classification car elle ne rapporte pas assez d'information pour notre modèle.

Ces conclusions ont été faites en travaillant sur les données dont on a supprimé complétement les Nan.

Serait-ce la méthode de classification qui n'est pas adéquate ou bien la manière dont on a traité les valeurs manquantes ?

Bibliographie :

Méthodes de filtrage de variables :

https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/?fbclid=IwAR3PCvUnVxyV_bbcPsphWQM2xjyDsG-6sseO9DAXWtX_QLTslcq3Arlj5ew

Arbres de décision :

https://towardsdatascience.com/decision-trees-and-random-forests-for-classification-and-regression-pt-1-dbb65a458df