La tâche qui nous incombe ce premier semestre relève du Machine Learning et de la classification multi-classes, deux domaines documentés par le corps enseignant Cours de Mme Demangeot mais également dans des contextes plus généraux http://cazencott.info/dotclear/public/lectures/IntroML_Azencott.pdf et https://fr.wikipedia.org/wiki/Apprentissage_automatique.

Notre premier modèle non naïf est le KNN. D'abord uniquement sur les coordonnées géographiques des espèces puis en prenant en compte l'ensemble des données abiotiques. Afin d'obtenir un modèle convenable le cours de Madame Demangeot cité précédemment nous a été utile. Nous avons aussi consulté les vidéos suivantes :
https://www.youtube.com/watch?v=LOj9yGrgU4Y&ab_channel=AIforyou-MorganGautherot

https://www.youtube.com/watch?v=P6kSc3qVph0&ab_channel=MachineLearnia

Notre second modèle non naïf est l'algorithme de forêts aléatoires, qui va construire plusieurs arbres de classification à partir des données bioclimatiques, abiotiques ou des deux afin de décider si une espèce est présente ou non selon les paramètres de la zone géographique où elle se trouve. Avant d'appréhender ces notions en cours nous avons pu consulter des vidéos (https://www.youtube.com/watch?v=v6VJ2RO66Ag&t=306s&ab_channel=NormalizedNerd) pour mieux comprendre le fonctionnement d'une random forest puis nous avons essayé de l'appliquer à nos données sur Rstudio avec la library "randomForest" en suivant ce lien https://thinkr.fr/premiers-pas-en-machine-learning-avec-r-volume-4-random-forest/ mais également sur avec pyhton en utilisant "scikit learn" (https://www.codementor.io/@agarrahul01/multiclass-classification-using-random-forest-on-scikit-learn-library-hkk4lwawu)

Mots clés utiles et présents dans les liens :

- Apprentissage automatique
- Arbres de régression
- Classification supervisée / non-supervisée
- Distribution d'espèces
- Kaggle
- KNN
- Machine Learning
- Multi-classes
- RandomForest
- Scikit-learn
- SVM

Il s'agit là d'un défi Kaggle terminé, c'est-à-dire avec de nombreuses soumissions qui reprend une problématique partagée par de nombreux chercheurs en biologie, en statistiques et en Machine/Deep Learning.
Plusieurs équipes ont participé au challenge kaggle en 2023, nous donnant des pistes de réflexions ainsi que des exemples de modèles avec le score obtenu. Sont notamment répertoriés les meilleurs scores des KNN et random forest que nous avons aussi implémentés. Nous pouvons alors aujourd'hui comparés nos résultats à ceux existants.
Random Forest environnemental : nous -> 0. 244 kaggle-> 0.188
KNN environnemental : nous -> 0. 223 kaggle-> 0.166
Nous essayons de construire un modèle de distribution d'espèces végétales le plus performant possible à partir des données mises à notre disposition mais il existe d'autres notions plus larges

comme la capacité de déplacement d'une espèce ou d'autres types de données permettant de construire ce genre de modèle de façon plus globale https://damariszurell.github.io/SDM-Intro/