Problématique : le but est de créer 2 modèles portants sur les niches potentielles et réalisées => niche potentielle = conditions favorables au développement d'une espèce / niche réalisée = niche ou l'espèce a été observée.

Le modèle constant proposé par la fac a obtenu un score de 0,16 -> jouer sur nombre d'espèce pour améliorer le nôtre.

Les données abiotiques bio1, bio2, … correspondent à des variables standardisées dont la signification est retrouvable dans le sujet.

K-voisins : ne pas retourner toutes les espèces mais celles présentes au moins 40% des fois (50% en théorie mais validateur Kaggle plus indulgent sur le surplus d'espèce que l'absence).

- Utilisation de la distance haversienne pour prendre en compte la courbure de la terre -> K voisins plus logique sur la longitude : variations de températures très déterminantes sur la répartition
- K-voisins abiotiques (pas géographiques) =niches potentielles

Les données présences only sont biaisés car favorables à des observations rares -> ne doivent pas diminuer les probabilités d'autres espèces en considérant comme absentes. Les précédents modèles ne peuvent pas être réellement être améliorés avec la présence only, ces données sont plutôt intégrées dans des modèles probabilistes de ML -> scikit learn

- ⇨ Construction d'une matrice des co-occurrence : si espèce présentes combien de fois autres espèces sont observées. On suppose répartition des espèces continues donc chaque vecteur correspond à une fonction de densité ? Score K-voisins X ligne matrice = proportion d'espèce attendue
- ⇨ Scikitlearn – Random forrest : tri automatique du modèle sur les axes d'informations et les privilégie sur prédiction

Avis sur interface Web : sélection d'une zone avec maintien souris puis prédiction dans toute la zone.