Projet n°4

Le contexte :
VSS : **volume apparent de distribution**, est un paramètre caractérisant la distribution d'une molécule dans le corps humain. On mesure la concentration sanguine de la substance active en question. Le volume de distribution est le volume dans lequel serait dissous la quantité administrée de substance active (c'est-à-dire contenue dans le médicament) pour donner la concentration obtenue dans le compartiment sanguin. C'est un volume fictif.

La question :
Le but de ce projet est de mettre en place une méthode de prédiction du **log(VSS) d**'une petite molécule à partir de descripteurs calculés sur cette dernière.

Les données :

Pour cela on dispose d'un jeu de données de 633 petites molécules décrites par 278 descripteurs calculés par 3 logiciels:
– 98 descripteurs calculés par le logiciel MOE (fichier MOE_ligMod.csv)
– 45 descripteurs calculés par le logiciel Material Studio (fichier MaterialStudio_ligMod.csv)
– 135 descripteurs calculés par le logiciel Volsurf (fichier volsurf_ligMod.csv)
Ces descripteurs peuvent-être classés en 3 classes :
– descripteurs 1D : accessible *via* la formule brute de la formule (ex: $C_2H_5NO_2$). Ils reflètent des propriétés très générales de la molécule tels que le poids moléculaire, le nombre d'atome …
– descripteurs 2D: calculés à partir de a structure 2D de la formule (cf. Figure 1). Ils donnent des informations sur les propriétés physico-chimiques des petites molécules (ex: nombre de donneurs de liaisons H, nombre d'anion …), caractérisent la taille, la ramification de la molécule …
– descripteurs 3D: calculés à partir de la structure 3D de la molécule (cf. Figure 2). Ces descripteurs renseignent sur la géométrie de la molécule (agencement des atomes (moment principal d'inertie, surface accéssible au solvant...). Ils sont basés sur les propriétés physico-chimiques de la molécule, sur son champs d'interaction moléculaire…
Les valeurs de la Vss pour les 633 molécules se trouvent dans le fichier Vss_ligMod.csv.

Guide d'analyse :
- penser à éliminer les variables inutiles (non informatives ou redondantes),
- justifier soigneusement les méthodes choisies,
- dans le cas où il y a des paramètres à choisir, prendre soin de les déterminer par validation croisée,
- les résultats terminaux seront donnés en utilisant la validation croisée sur le jeu d'apprentissage et le jeu de validation,
- conclure sur les avantages et inconvénients des méthodes utilisées.