**Cours LSTAT2340 – Traitement statistique de données Omiques**

**Projet 5: Modèles de classification et données métabolomiques**

**Objectif**

Dans ce projet, vous appliquez les méthodes de classification sur des données métabolomiques.

**Votre deliverable**

Vous déposez sur Moodle votre code R Markdown ainsi que le fichier html généré.

**Description des données**

* Dans ce projet, vous disposez de données d’analyse par H-NMR. Ce sont donc des spectres et chaque spectre est discrétisé en 238 descripteurs (les variables donc) et a comme axe des x les ppm.
* On a analysé par RMN des échantillons de sang de 3 patients (A, B et C). Chaque patient a donné son sang trois fois (à 3 semaines d’intervalle). Nous avons donc 9 échantillons de sang. Chacun de ces échantillons a ensuite été séparé en 3 puis chaque sous-échantillon résultant (il y en a donc 27) a été passé 3 fois dans le spectromètre. Nous disposons donc en tout de 81 spectres.
* Les spectres sont nommés comme suit : PBS2\_E1R3 est la 3 ième analyse du 1ième sous échantillon obtenu à partir du second prélèvement de sang du patient B.
* Votre «mission» est de trouver, via plusieurs modèles de classification, ce qui différentie les échantillons de sang de 2 de ces 3 patients et éventuellement (mais c’est un objectif secondaire) d’observer des éventuelles variations entre les échantillons d’un même patient dans le temps.
* Les modèles seront potentiellement très simples car, vous le verrez, les patients se différentient très bien mais il faudra trouver par quoi !

**Consignes plus précises**

* Les données sont dans le fichier csv CarreLatinMBXmanualPPMINV.csv.
* Le fichier CarreLatinMBX\_Student.Rmd vous donne déjà du code pour le lire, dessiner les spectres et effectuer une première ACP. Faites en le tourner pour déjà vous faire une petite idée !
* Votre analyse statistique est à réaliser en R avec rédaction d’un petit rapport en R markdown.
* Les étapes de votre analyse doivent comprendre les points suivants :

1. Sur base de la PCA, choisissez les 2 patients que vous avez envie de comparer. Vous travaillez donc ensuite uniquement sur les données les concernant.
2. Ajustez au minimum les modèles de classification suivants sur ces données : logistique stepwise, PLS-DA, OPLS-DA, régression logistique pénalisée.
3. Comparez des performances des 4 modèles selon une série de critères.
4. Identifiez, pour chaque méthode une liste de variables (ppm) ou de pics du spectre qui vous permettent de séparer le mieux les 2 patients et comparez de cette liste.
5. Donnez vos remarques éventuelles supplémentaires sur ce que vous voyez dans ces données.

Basez-vous sur les codes Rmd des analyses de données q\_PCR et Gasoline que nous vous avons donnés pour réaliser vos analyses (usez et abusez du copier-coller ☺). La petite difficulté pour bien présenter les données est que dans les données on ne dispose que de peu de variables et donc on peut toutes les nommer et les dessiner dans les sorties. Par contre ici vous avez des spectres et donc beaucoup de sorties (les paramètres et loadings) devront être dessinées comme des courbes. C’est ce qui était fait dans le code gasoline.

**Délais**

Votre rapport Rmarkdown est à rendre pour la semaine 9 (ou avant bien-sûr). Vous pouvez nous poser des questions durant les congés de Pâques.