Parcours Informatique Médicale
DFGSM3
Semestre 1 - 2018-2019

Le sujet comporte 4 volets. Chaque binôme doit réaliser une des 4 parties.

Le projet devra être **rendu** au plus tard le **4 février 2019 à 08h**.
La **soutenance** aura lieu le **6 février** pendant les heures de cours.

Les résultats doivent être présentés sous la forme d'une application R Shiny dont la démonstration sera faite lors de la soutenance orale.
Les codes sources de l'application, et un rapport de 5 pages (hors page de titre) par binômes seront envoyés aux enseignants selon des modalités définies ultérieurement.

La soutenance durera 15 minutes au total par binôme, dont au plus 5 minutes de démonstration.

Les enseignants sont disponibles pour vous apporter des éclaircissements, des explications ou vous guider dans votre travail. Cela d'autant plus que vous les solliciterez AVANT la date limite.

Sujets

1 a) Nettoyage d'un jeu de données

Les données proviendront d'un jeu de données au format CSV (comma separated values)

- Une première étape consiste à nettoyer ces données (cf TP)
    - identification des types de données (numériques, catégorielles binaires, catégorielles n-aires...)
    - identification de données aberrantes et nettoyage automatisé
    - identification de données hors bornes
    - L'étape de nettoyage et de préparation des données devra faire l'objet de la production d'un rapport de *log* automatisée.

1 b) Préparation d'un jeu de données

- Une deuxième étape est la création de nouvelles variables adaptées à l'analyse prévue. Parmi les méthodes courantes, on retiendra :
    - Calcul de délais à partir de deux variables temporelles (au format ISO e.g. 2017-01-16)

- Transformation de variables booléennes à partir de variables catégorielles (une variable à *n* (>2) classes est transformée en *n* variables à 2 classes contenant la même information)
- Création d'une variable catégorielle à partir d'une variable numérique :
    - selon un (des) seuil(s) choisi(s) par l'utilisateur
    - selon des seuils déterminés statistiquement (quartiles par exemple)

2) Analyse exploratoire

Analyse descriptive et visualisation des

- Données numériques
- Données catégorielles

La distribution des variables individuelles sera observable. Vous permettrez par la visualisation le croisement de deux variables.

3) Analyses statistiques univariées

Vous permettrez à l'utilisateur d'analyser les relations entre deux variables. Votre application devra choisir les tests statistiques adaptés en fonction des types de variables et des conditions d'applications de ces tests.
Vous guiderez l'utilisateur dans l'interprétation des résultats.

Ressources disponibles

Vous pourrez vous aider des documents utilisés pendant les cours et de ressources en lignes comme par exemple :

- Choix de tests statistiques : https://marne.u707.jussieu.fr/biostatgv/
- Tutoriel Shiny: http://shiny.rstudio.com/tutorial/
- ggplot2: http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html, http://www.cookbook-r.com/Graphs/
- R: http://www.cookbook-r.com

Coordonnées des enseignants: (merci de ne pas téléphoner à 23h le dimanche)

Antoine Neuraz: antoine.neuraz@aphp.fr, tel: 01.71.39.65.85, mobile: 06.24.62.23.55
Bastien Rance: bastien.rance@aphp.fr, tel: 01.56.09.59.85, mobile: 06.14.89.16.35
Maxime Wack: maxime.wack@aphp.fr, tel: 01.56.09.23.63, mobile: 06.98.80.60.36