4.1 La base de données `college` contient les variables suivantes :

- `prive` : variable binaire indiquant si l'institution d'enseignement est privée (1) ou pas (0)
- `napplications` : nombre d'applications pour les études
- `nadmission` : nombre de personnes admises parmi `napplications`
- `ninscrits` : nombre d'étudiant·e·s inscrit·e·s à l'université
- `m10p` : pourcentage des admis provenant d'une école secondaire classé parmi les 10% meilleures
- `m25p` : pourcentage des admis provenant d'une école secondaire classé parmi les 25% meilleures
- `tempsplein1c` : nombre d'étudiant·e·s à temps plein au premier cycle
- `tempspart1c` : nombre d'étudiant·e·s à temps partiel au premier cycle
- `fraiscolexternes` : frais de scolarité pour étudiant·e·s externes (hors état)
- `fraisres` : frais pour les résidences et l'hébergement
- `fraislivres` : frais moyens pour les ouvrages obligatoires
- `fraisperso` : frais pour des dépenses personnelles
- `pourcentdoctorat` : taux de membres du corps enseignant détenteurs d'un Ph.D.
- `pourcentterminal` : taux de membres du corps enseignant détenteurs d'un diplôme terminal
- `ratioetudprof` : rapport du nombre d'étudiant·e·s versus professeur·re·s
- `pourcentdonationdiplome` : pourcentage des diplômé·e·s qui font des dons à l'institution
- `depenseparetud` : dépenses liées à l'enseignement, par étudiant·e
- `tauxdiplom` : taux de diplomation
- `nom` : nom de l'établissement

Le but de l'exercice est de bâtir un modèle prédictif pour le nombre annuel de demandes d'admission.

(a) Faites une analyse exploratoire des variables explicatives :

- quelles variables devraient êtres exclues de la modélisation? Justifiez votre réponse
- comparez la variable réponse avec les autres variables : y a-t-il des transformations qui améliorerait l'ensemble de variables candidates : interactions, création de variables dychotomiques, transformations (racines carrée, transformation logarithmique, etc.)? Vérifiez s'il y a des variables catégorielles encodées comme des variables numériques.

(b) Scindez la base de données en échantillon avec données d'entraînement (environ 2/3 des données) et échantillon de validation; utilisez le germe aléatoire `60602`. Sélectionnez un modèle à l'aide d'une des méthodes couvertes, mais en basant votre choix sur l'erreur moyenne quadratique évaluée sur l'échantillon de validation.

(c) Répétez la sélection, cette fois en prenant comme critère pour l'erreur moyenne quadratique évaluée par validation croisée (aléatoire) à cinq plis.

(d) Créez un tableau avec le nombre de coefficients de votre modèle final et un estimé de l'erreur moyenne quadratique obtenu par validation externe ou croisée. Commentez sur le meilleur modèle parmi les combinaisons.

Compilé le 18/10/2021 à 11:22