JULIEN JACQUES UNIVERSITÉ DE LYON

Fouille de données

Licence Professionnelle CESTAT - Examen du 11 avril 2019

Durée : 1h30 - Tous documents autorisés

Document à rendre En fin d'examen, vous enverrez par email à `julien.jacques@univ-lyon2.fr` un unique document, intitulé `votrenom.pdf`, contenant le compte-rendu de votre travail, les graphiques, les résultats, vos interprétations et conclusions ... ainsi que les codes R utilisés pour répondre à cet examen.
Pensez à conserver également une copie de sauvegarde de votre examen.

Données Nous allons travailler sur une base de données décrivant 6224 individus américains décrits par 15 variables :

`http://eric.univ-lyon2.fr/~jjacques/Download/DataSet/adult_sample.data`

Les 15 variables sont les suivantes :

- Age : âge
- CSP : catégorie socio-professionnelle
- ScoreDemo : un score démographique
- Diplome : le type de diplôme
- ScoreDiplome : un score construit en fonction du type de diplôme
- StatutMarital : le statut marital
- Profession : la profession
- SituationFamiliale : la situation familiale
- Ethnie : l'origine ethnique
- Genre : le genre
- Economies : le montant des économies
- Dettes : le montant des dettes
- HeureSemaine : le nombre d'heures travaillées par semaine
- PaysOrigine : le pays d'origine
- Revenus : montant des revenus (supérieur ou inférieur à 50k$)

Travail à effectuer

1. Importer les données. Identifier les variables quantitatives et les variables catégorielles. Imputer les données manquantes s'il y en a.
2. Réaliser une ACP sur les variables quantitatives, et essayer d'interpréter les deux premiers axes.
3. Réaliser un clustering en deux groupes en n'utilisant que les variables quantitatives. Essayer d'interpréter vos clusters.
4. Nous allons maintenant chercher à prédire la variable revenus, qui est catégorielle :
   (a) Tirer aléatoirement 1000 données qui nous servirons de base de test, en initialisant la graine aléatoire de votre ordinateur à l'aide de la commande `set.seed(1)`. Les données restantes serviront d'apprentissage.
   (b) Réaliser une classification avec un arbre de décision, en n'utilisant que les variables quantitatives. Evaluer la qualité de la prédiction sur les données tests.
   (c) Faites de même en utilisant cette fois les variables quantitatives et catégorielles. Comparer la qualité de la prédiction avec la question précédente.
   (d) Question bonus : essayer d'améliorer votre prédiction à l'aide d'une forêt aléatoire.