TD découverte du logiciel RATTLE-GUI

Aide à l'exécution de Rattle :
Ouvrir Rstudio.
Ecrire les deux lignes de commandes suivantes pour lancer Rattle :
library(rattle)
rattle()

Si vous voulez utiliser Rattle depuis votre ordinateur personnel, télécharger les packages « rattle » et « stringr ».

Exercice 1 : méthode prédictive

Description du jeu de données : *échantillon de 100 individus prélevés d'une étude du CHU d'Angers décrits par leur âge, leur poids, leur taille, le nombre de verre d'alcool consommé par jour, le fait qu'ils soient fumeur ou non-fumeur et le fait qu'ils ronflent ou non (1= ronfle ; 0= ne ronfle pas).*

1. Importer le jeu de données « dataRonfle ».
2. Découverte des données à l'aide de statistiques descriptives :
   → Faire un diagramme en barre sur la variable ronfle (nécessite le recodage de « ronfle » en type « catégorique »).
   → Tester la corrélation entre l'âge et l'alcool.
   → Trouver combien de femmes ronflent dans le fichier.
3. Séparer le jeu de données en 2 avec une partie apprentissage et une partie test (70% ; 30%).
4. Trouver combien de personnes sont contenues dans l'échantillon d'apprentissage.
5. Créer un arbre à partir de l'échantillon d'apprentissage en prenant la variable « ronfle » comme cible. Sortir le graphique correspondant.
   Quelle variable influe le plus le modèle ?
6. En sachant que la modalité positive est ronfle=1, calculer la sensibilité, la précision et taux d'erreur.
7. Quelle serait la prédiction pour l'individu P0108 ?
8. Évaluation du modèle : Faire la courbe ROC afin d'évaluer si le modèle est bon.
9. Réaliser une régression logistique afin de comparer les deux méthodes. Laquelle des méthodes prédictives est la meilleure d'après la courbe ROC ?
   Observez-vous des résultats similaires pour les deux méthodes ?
10. Exporter le code R qu'a généré Rattle en arrière-plan.

Exercice 2 : Apprentissage non-supervisé

Le jeu de données « Mammifères » présente les caractéristiques de 55 mammifères. Les variables disponibles sont le poids (en kg), le poids du cerveau (en g), le temps de sommeil (en heures), l'espérance de vie (en années), la durée de gestation (en jours), l'indice de prédation (de 1-moins susceptible d'être la proie à 5-plus susceptible d'être la proie), l'indice d'exposition pendant le sommeil (de 1-l'animal dort dans un endroit protégé à 5-l'animal est exposé pendant son sommeil) et l'indice de danger global (de 1-peu en danger à cause des autres animaux à 5-très en danger à cause des autres animaux).

1. Importer les données. Indiquer à Rattle que la variable « Espèce » est l'identifiant et que la variable « Danger global » est une variable d'entrée (et non une variable cible). Faire attention à décocher la case « Partition » (automatiquement cochée) avant d'exécuter les modifications.
2. Afficher les données.
3. Donner le poids moyen des mammifères du jeu de données et réaliser un boxplot de la durée de gestation.
4. Réaliser la matrice des corrélations.
5. Normaliser les données afin de travailler avec des variables de même unité.
6. Afficher le graphique d'évolution de l'inertie intra-classe en fonction du nombre de classes.
7. A l'aide de ce dernier, déterminer le nombre de classes et réaliser l'analyse des k-means correspondantes. Représenter graphiquement les résultats.
8. Caractériser les clusters.
9. Pour affiner l'analyse, construire une CAH et afficher le dendrogramme.
10. Exporter les données sous un fichier .csv et les importer sous Excel (à partir de l'onglet Evaluer).
11. Croiser les résultats issus des k-means et de la CAH afin de comparer la cohérence des groupes obtenus.