Sujet de Travail Dirigé
Master 2 SISE

---

**RapidMiner**

---

rapidminer

NGO Ronald
MALOD Jeremy
HAMITOU Seifidine
YONGWE Jean-Luc

Professeur : Mr. Ricco RAKOTOMALALA

Année universitaire : 2016/2017

25 octobre 2016

Les jeux de donées de l'exercice 1 et 3 vous seront transmis par mail.

***Création d'une variable***

Importer dans votre "Repository" le jeu de données *"IMC.xlx"* , poser l'opérateur des dites données dans votre "Process" en cliquant dessus avec votre souris et le déplacant dans le "Process"

**Exercice 1**

**Question 1 :**
Retrouver l'opérateur *Generate Attribute* dans le menu déroulant des Opérators.

**Question 2 :**

Paramétrer l'opérateur précèdent dans *"Edit List(0)"* dans l'onglet *Parameters"*

**Question 3 :**
Paramétrer votre variable *IMC* en utilisant la fonction du calcul de l'Indice de Masse Corporel :

$$f_{IMC}(poids, taille) = \frac{poids}{taille^2}$$

***Classification***

**Exercice 2**

Le fichier utilisé pour cet exercice est le fichier *"iris"* présent sous RapidMiner donnant des informations sur la taille des fleurs de trois familles différentes.
Il est composé de 6 variables :

- Id (identifiant unique de l'observation)
- Famille (type de fleurs : Setosa , Virginica , Versicolor )
- a1 : longeur des sépales
- a2 : largeur des sépales
- a3 : longeur des pétales
- a4 : largeur des pétales

**Question 1 :**
Retrouverle fichier dans les repository de RapidMiner

**Question 2 :**
Observer les différentes statistiques descriptives que vous pouvez en tirer (graphiques et indicateurs) Essayer de décrire ce que vous retrouvez, notamment avec les nuages de points

**Question 3 :**
Renommer les variables a1, a2, a3, a4 dans des noms explicits

**Question 4 :**
Scinder le fichier en deux échantillons, un échantillon apprentissage et un échantillon test (l'opérateur *Validation* va vous aider)

**Question 5 :**
Créer un modèle à l'aide de la méthode des k-plus proches voisins sur l'échantillon d'apprentissage

**Question 6 :**
Enregistrer le modèle et l'appliquer sur l'échantillon test

**Question 7 :**
A l'aide l'outil performance, visualiser la matrice de confusion

**Question 8 :**
Changer les paramètres de l'opérateur et observer les résultats de la matrice de confusion

***Régression Linéaire***

**Exercice 3**

Importer les données intitulées *"BDD Voitures.xlx"* . Vous etes libre d'utiliser l'opérateur *"Read CSV"* ou aller dans l'onglet *"Add Data"* et les importer depuis votre arborescence où vous les avez téléchargés depuis votre boite mail.

***Partie 1 : Régression linéaire RapidMiner***

**Question 1 :**
Scinder l'échantillon en deux partitions. L'échantillon d'apprentissage qui sert à la détermination du modèle devra comprendre $\frac{2}{3}$ des observations tandis que l'échantillon test, le tiers restant. ( Blending -> Examples ->sampling . . . )

**Question 2 :**
Utiliser l'opérateur de régression linéaire sur l'échantillon d'apprentissage. Implémenter le modèle (voir du coté de l'opérateur *Scoring*) et l'appliquer sur l'échantillon test.

**Question 3 :**
Que pouvez-vous dire de la performance de notre modèle (voir du côté de l'opérateur *Performance*). Que pouvez-vous dire de nos variables explicatives, de la qualité du modèle ?

***Partie 2 : Régression linéaire script R***

**Question 4 :**
Télécharger le package R. Voir le lien *" Get More Operators "*. Vous le trouverez plus facilement dans l'onglet " Top Downloads ".

**Question 5 :**
Ouvrez un nouveau process et faites appel à votre jeu de données. Scinder votre échantillon comme vue précédemment.

**Question 6 :**
Faites appel au script R afin d'insérer votre code. Celui-ci comprendra une fonction dont l'execution devra lancer une regression linéaire du prix moyen sur toutes les autres variables (explicatives). (Vue en cours de Programmation R).

**Question 7 :**
Ajouter une nouvel opérateur *Execute R* implémentant cette fois-ci un code qui devra ajouter une colonne à notre tableau de données contenant nos prédictions. Ainsi trouvé. Exécuter votre programme.

**Question 8 :**
Comparer les performances de ce modèle avec celles de la partie 1. Qu'en concluez-vous ?