Série 4 des Travaux pratiques en Fouille de données : Clustering

Travail préparatoire

1. Créer un document Excel avec deux feuilles ; renommer les feuilles exo1, exo2. Sauvegarder le document sous le nom *TP4_nom1_nom2.xls* où *nom1* et *nom2* désignent vos noms respectifs.

2. Les réponses aux questions doivent être rédigées dans le document Excel en dessous des données, regroupées par exercice et numérotées conformément à la série.

Exercice 1 : Importer le fichier texte ***diabetes.txt*** disponible sur la plateforme de e-learning depuis le campus virtuel vers le fichier Excel, feuille exo1 et supprimer toutes les colonnes à l'exception des trois colonnes *Plas*, *Pres* et *Skin*.

1. Répartir les individus en quatre groupes aléatoirement en essayant d'avoir des groupes de cardinalités proches.

Individu	**Groupe**
Valeur 1	G1
Valeur 2	G2
…	

2. Calculer l'écart-type à l'intérieur de chaque groupe pour chaque variable. Structurer la réponse pour chaque groupe comme suit (pensez à utiliser les filtres)

Groupe G1

Individu	**Plas**	**Pres**	**Skin**
1			
2			
…			
Ecart-type			

…..

Exercice 2 : Lancer le logiciel Tanagra

- Créer un nouveau diagramme (file/new). Sauvegarder le diagramme sous le nom *TP3_nom1_nom2.tdm*.
- Dans le choix du dataset, sélectionnez le fichier texte ***Diabetes.txt***. L'élément Dataset s'affiche dans Tanagra.
- Cliquer sur l'icône de définition de statuts (4 flèches colorées) pour l'afficher sous *Dataset* et double-cliquer sur *Define status 1*. Choisissez les trois variables **plas** et **pres** et **skin**.
- Dans la partie en bas de Tanagra, sélectionnez l'onglet *Clustering*. Repérer *K-means* et glisser son icône sur l'icône *Define Status 1*. Une nouvelle icône *K-means 1* s'affiche en dessous de *Define status 1*.
- En utilisant le bouton droit sur l'icône *K-means 1*, paramétrez K-means en gardant les valeurs

par défaut, sauf pour la normalisation des distances (choisir *none*). Aussi, dans l'onglet *Results*, cocher l'option Table Anova. Faire le travail pour K=4.

- Exécuter l'algorithme en double-cliquant sur l'icône K-Means1.
- Dans l'onglet *Data visualization*, choisir *View dataset* et le glisser sur l'icône *K-means 1*. L'icône *View dataset 1* apparaît. Double-cliquer dessus pour afficher le tableau du résultat.

1. Quel est le nombre d'objets par cluster. Expliquer la différence avec la répartition de l'exercice 1 ?

- Refaire le même travail de clustering K-means sur l'icône « Define_status 1 » pour 3 clusters.

2. Remplir le tableau suivant (manuellement) à partir des résultats du clustering et des résultats de l'exercice 1.

Moyenne des écart-type	Sans K-means (4 clusters)	K-means 4 clusters	K-means 3 clusters
Plas	Moy(ET1+ET2+ET3+ET4)=…	Moy(ET1+ET2+ET3+E4)=…	Moy(ET1+ET2+ET3)=…
Pres	Idem	Idem	Idem
Skin	Idem	Idem	Idem

- selon votre analyse du tableau, répondre aux questions suivantes

3. Quelles remarques peut-on faire sur les différences entre les écarts-types. Justifier les réponses.

    a. selon l'utilisation ou pas de K-means
    b. selon le nombre de clusters

4. Quel est selon vous le meilleur clustering (sans clustering ou avec clustering pour quel nombre K ?)

Modalité d'envoi des solutions : Le fichier Excel et le fichier *.tdm* doivent être envoyés à l'adresse email suivante : doulkifli.boukraa@univ-jijel.dz. La date limite d'envoi sera communiquée ultérieurement.