Enseignants	TP5	Classe
COURS : A. NAJJAR-– M. FARHAT I. BEN OTHMEN **TP : F. JENHANI- S. BIROUZA I.HAMROUNI**	**Machine Learning**	**3ème GLSI**

Partie 1 (K-moyennes)

Reprenez la base "**pima-indians-diabetes.data.csv**" utilisée lors du dernier TP. On souhaite effectuer la catégorisation des individus par l'algorithme k-moyenne.

1- **Préparation des données** : afin de pouvoir afficher les individus dans un plan (2D), ainsi que leurs classes d'appartenance on fera les restrictions suivantes :
   - Pour la description des individus, on se restreindra à deux attributs.
   - Pour le nombre des individus, on se restreindra à 100 individus.

   a- Construire le nouvel ensemble de données (100 individus et 2 attributs).
   b- Effectuer la normalisation des donnés afin que valeurs des deux attributs soient dans un intervalle [0-1].
   c- Afficher dans un graphe les individus tel que le premier attribut représente les abscisses et le deuxième attribut représente les ordonnés.

2- **Catégorisation** : Effectuer la catégorisation par la méthode des k-moyennes puis afficher le résultat de la catégorisation pour des valeurs de k =2, 3 et 4. Conclure.

On souhaite effectuer la catégorisation automatique d'un ensemble de fromages. L'objectif est d'identifier des groupes de fromages homogènes, partageant des caractéristiques similaires. Pour ce faire, la classification ascendante hiérarchique (CAH) sera utilisée.

Partie 2 (Classification Hiérarchique ascendante)

Les données sont disponibles dans un fichier "**fromage.txt**". On possède 29 observations décrites par des variables quantitatives.

1- Importer les données à partir du fichier "**fromage.txt**".
2- Afficher la dimension de l'ensemble de données et en déduire le nombre des variables utilisées.
3- Afficher les noms de ces variables.
4- Déterminer l'intervalle de variation de chacune des variables.
5- Changer l'échelle des données pour les ramener dans intervalle [0-1]. Pourquoi cette opération est-elle nécessaire ?

6- Afficher le **dendrogramme** qui correspond à une classification hiérarchique ascendante. Utiliser la distance euclidienne et tester plusieurs critères d'agrégation : "indice de Ward", "le lien minimum", "le lien maximum" et "le lien moyen". Analyser les dendrogrammes obtenus en identifiant, pour chacun des critères, le meilleur choix du nombre de classes k.

7- Effectuer la **catégorisation** en utilisant la classification hiérarchique ascendante et **visualiser** les résultats obtenus et ce en utilisant :

- Variables : seulement les 1er et 2ème attributs
- Distance : euclidienne
- Critères d'agrégation : "l'indice de Ward", "le lien maximum" et "le lien moyen".
- Nombre de classes k : entre 2 et 5.