Université de Lille **Méthodes d'Apprentissage**
M2 FCA - Data Sciences obouaziz@univ-lille.fr.fr

TP. 6
Clustering : Kmeans et CAH

Exercice
On considère à nouveau le jeu de données `load_digits` disponible dans `sklearn.datasets`.

1. Importer le jeu de données. Visualiser rapidement les variables disponibles. Isoler une matrice contenant les différentes images (stockées sous forme de vecteurs de pixels) et la tracer pour quelques individus (voir `plt.matshow`).
2. **Kmeans** (a) Importer le fonction `KMeans` et l'appliquer aux données. Faites varier le nombre de classes et l'initialisation et regarder comment cela influe sur l'inertie intra-classe. Commenter.
   (b) Faire tourner l'algorithme des Kmeans sur 10 classes et comparer les résultats obtenus avec les vrais labels. Commenter. *On pourra regarder la fonction `crosstab`.*
3. **CAH** (a) Importer et appliquer aux données la fonction `AgglomerativeClustering`.
   (b) Tracer le dendrogramme associé (regarder l'aide `https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html`).
   (c) Combien de classes retenir ?
   (d) Dans le cas d'une CAH à 10 classes comparer 1) les résultats obtenus avec les vrais labels et 2) les résultats obtenus avec ceux de la méthode Kmeans.
4. **Méthode mixte** Appliquer la méthode mixte aux données. Combien de classes retenir ? Conclure.

Exercice 2 : Classification des villes européennes [R]

On cherche à regrouper les villes européennes en fonction du climat. Pour cela, on a relevé les températures mensuelles moyennes, la température annuelle moyenne et l'amplitude de variation des températures, ainsi que la latitude et la longitude de 35 villes.

1. Importer la base de données `temperatures.csv`.
2. Proposer une classification de la base de données, à l'aide de l'algorithme des K-means et de l'algorithme de CAH, et sans tenir compte de la variable `Region`. Combien de classes retient-on ? Comparer les classes obtenues avec les deux méthodes.
3. Décrire les classes obtenues, et proposer une interprétation des résultats.