INF4117 : TP examen
Durée : 2 heures

Soit les données stockées dans le fichier « Wholesale customers data.csv ». Les attributs sont décrits de la manière suivante :

1. FRESH: annual spending (m.u.) on fresh products (Continuous);
2. MILK: annual spending (m.u.) on milk products (Continuous);
3. GROCERY: annual spending (m.u.)on grocery products (Continuous);
4. FROZEN: annual spending (m.u.)on frozen products (Continuous) DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
5. DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);
6. CHANNEL: customersâ€™ Channel - Horeca (Hotel/Restaurant/CafÃ©) or Retail channel (Nominal)
7. REGION: customersâ€™ Region â€“ Lisnon, Oporto or Other (Nominal)

TAF

1. Les attributs « CHANNEL » et « REGION » même étant codés comme des entiers sont catégoriels. Utiliser le one-hot encoding pour binariser ces attributs
2. Appliquer les algorithmes kmeans avec k= 2 et DBSCAN sur ces données, et visualiser les cluster dans un graphique
3. Calculer le **silhouette score**, le **davies-bouldin score**. Et comparer les deux résultats.
4. Supposer que la colonne « CHANNEL » représente la classe de chaque exemple, appliquer une mesure supervisée pour évaluer les résultats du clustering.