TD 6 Regroupement (suite) + Autres opérations de Mise en forme de données sous RapidMiner

1. On va utiliser l'algorithme DBScan pour effectuer du regroupement
 - a) Récupérer dans RapidMiner le fichier DeuxCarres.csv à partir du site du cours
 - b) Visualiser ce jeu de données en utilisant RapidMiner
 - c) Appliquer DBScan à ce jeu de données avec les paramètres Epsilon=1.1, MinPoints=2, données numériques et distance de Manhattan
 - d) Visualiser le résultat sous forme graphique
 - e) Ré-exécuter DBScan avec Epsilon=1 et MinPoints=2
 - f) Appliquer K-means à votre jeu de données avec K=2. Comparer le résultat obtenu ici avec le résultat obtenu à la question c)

2. Reprendre le fichier DeuxCarrés vu au TD précédent. On a vu qu'avec Eps=1.1 et MinPts=2, l'algorithme DBSCAN permet de retrouver les deux carrés en les considérant comme deux groupes distincts. On a vu que K-means était incapable d'identifier ces deux carrés. Est-il possible de retrouver ces deux mêmes carrés en utilisant un algorithme hiérarchique agglomératif ? Expliquer

3. On considère le fichier « Houston 24 ... » qui se trouve dans Training Ressources/Data/Hourly Weather. Il s'agit de relevés météo relatifs à la ville de Houston. Les informations ont été enregistrées avec une granularité d'une journée. On veut utiliser ces données pour apprendre un modèle de prédiction météo (trop prétentieux comme objectif) spécifique à cette ville. Le modèle est censé se baser sur les données météo d'une période donnée pour prédire la météo du jour succédant à la période. Par exemple, si on fixe la taille de la période à 3 jours, alors pour prédire la météo de demain, nous avons besoin des données météo d'aujourd'hui, d'hier et d'avant-hier. Si la taille de la période est fixée à 1, alors nous n'aurons besoin que des données d'aujourd'hui.

- a) Transformation des températures de Fahrenheit à Celsius : Pour cela, utiliser l'opérateur `Generate attribute` en fixant le nom du nouvel attribut et la formule définissant ses valeurs. $Celsius = (Fahrenheit - 32) / 1.8$.
- b) Supprimer la colonne de la température en Fahrenheit avec l'opérateur `Select Attributes`.
- c) Agrégation de périodes : nous devons construire un jeu d'apprentissage qui
 - i agrège les données sur une période de longueur P (en prenant la moyenne par exemple) et
 - ii qui affecte la condition météo (attribut `weather description`) du jour P+1

Pour ce faire, utiliser l'opérateur `Process Windows` qui est capable de construire des enregistrements à partir d'une série en les agrégeant selon une taille de fenêtre glissante et en créant une valeur « Horizon » qui pour nous est la condition météo du jour P+1. Les enregistrements obtenus doivent faire partie d'un même jeu de données. Utiliser l'opérateur `Append` pour les concaténer.

- d) Tester les performances de différents modèles en faisant varier la valeur de P. On réalise que les prédictions sont en général insatisfaisantes. On change alors d'objectif, au lieu de chercher à prédire les conditions météo, on va essayer de prédire la température. On se focalisera sur une prédiction qualitative : Froid (< 0), Frais ($[0,10[$), Doux ($[10,25[$), chaud ($[25,35[$) et canicule si c'est plus. Utiliser l'opérateur `Discretize by user specification` qui permet de transformer des valeurs numériques en valeurs nominales. Vérifier la qualité prédictive des modèles ainsi obtenus.