Université Sultan Moulay Slimane
Ecole National des Sciences Appliquées
Khouribga

Module : Data Mining & Texte Mining

1ère Année Master Big Data & Aide à la Décision
Semestre 2 / Année universitaire 2018/2019
Feuille de Travaux Pratiques N° 4

OBJECTIF DE L'ACTIVITE PRATIQUE :

*Dans ce TP, vous allez manipuler le logiciel **KNIME** pour effectuer la tâche de regroupement (Clustering) en utilisant les trois algorithmes classiques : **K-Means**, **AHC** (Ascendant Hierarchical Clustering) et **DBSCAN**.*

L'algorithme K-means

1. Créez un nouveau workflow.
2. Ajoutez les quatre nœuds suivants : « **Table Creator** », « **K-Means** », « **Color Manager** » et « **Scatter Plot** », puis les connecter comme indiqué dans la figure 1.

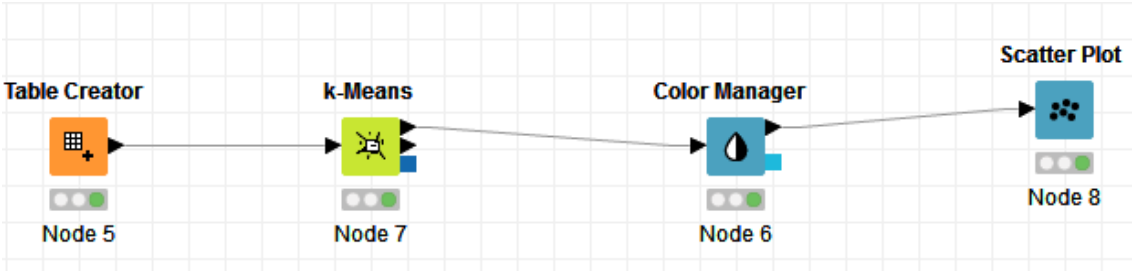
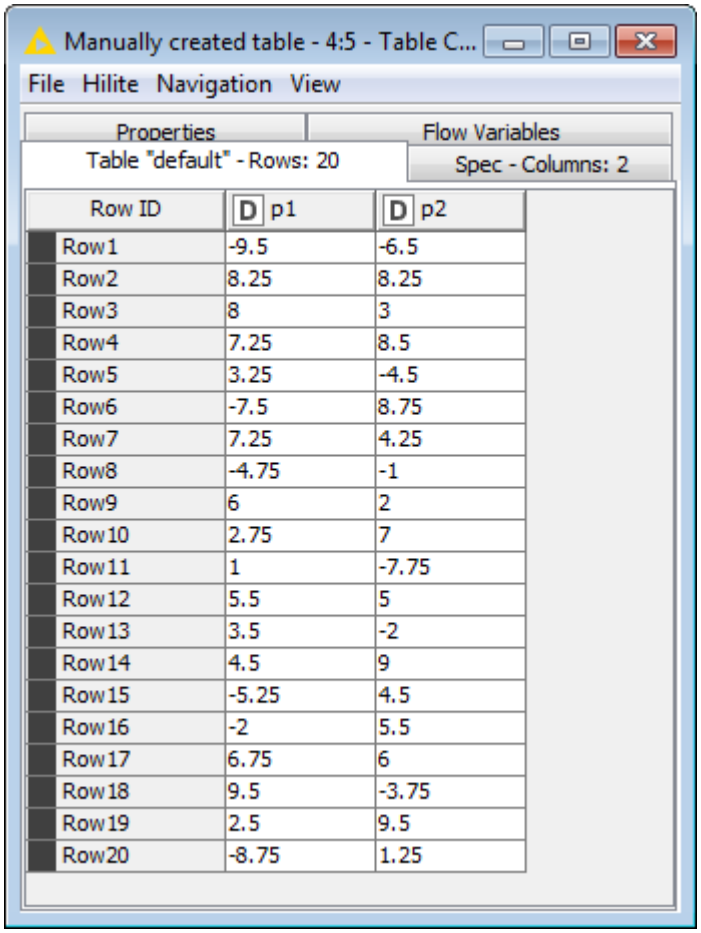


Fig. 1. Les nœuds du projet N° 1.

3. Configurez le nœud « **Table Creator** » pour saisir une matrice de données X représentant un univers Ω formé de $n = 20$ individus, où chaque individu est décrit par $p = 2$ variables. Exécutez ce nœud (le résultat doit être semblable à celui de la figure 2).
4. Configurez le nœud « **K-Means** » par le choix du nombre de clusters (paramètre $K = 2$), le nombre d'itérations de l'algorithme (le fixer à 5) et les colonnes à inclure (voir figure 3).
5. Exécutez l'ensemble des nœuds, puis visualisez le résultat à l'aide du nœud « **Scatter Plot** » (Commande « **View : Scatter Plot** »). Figure 4.



Row ID	D p1	D p2
Row1	-9.5	-6.5
Row2	8.25	8.25
Row3	8	3
Row4	7.25	8.5
Row5	3.25	-4.5
Row6	-7.5	8.75
Row7	7.25	4.25
Row8	-4.75	-1
Row9	6	2
Row10	2.75	7
Row11	1	-7.75
Row12	5.5	5
Row13	3.5	-2
Row14	4.5	9
Row15	-5.25	4.5
Row16	-2	5.5
Row17	6.75	6
Row18	9.5	-3.75
Row19	2.5	9.5
Row20	-8.75	1.25

Fig. 2. La matrice de données *X* à saisir.

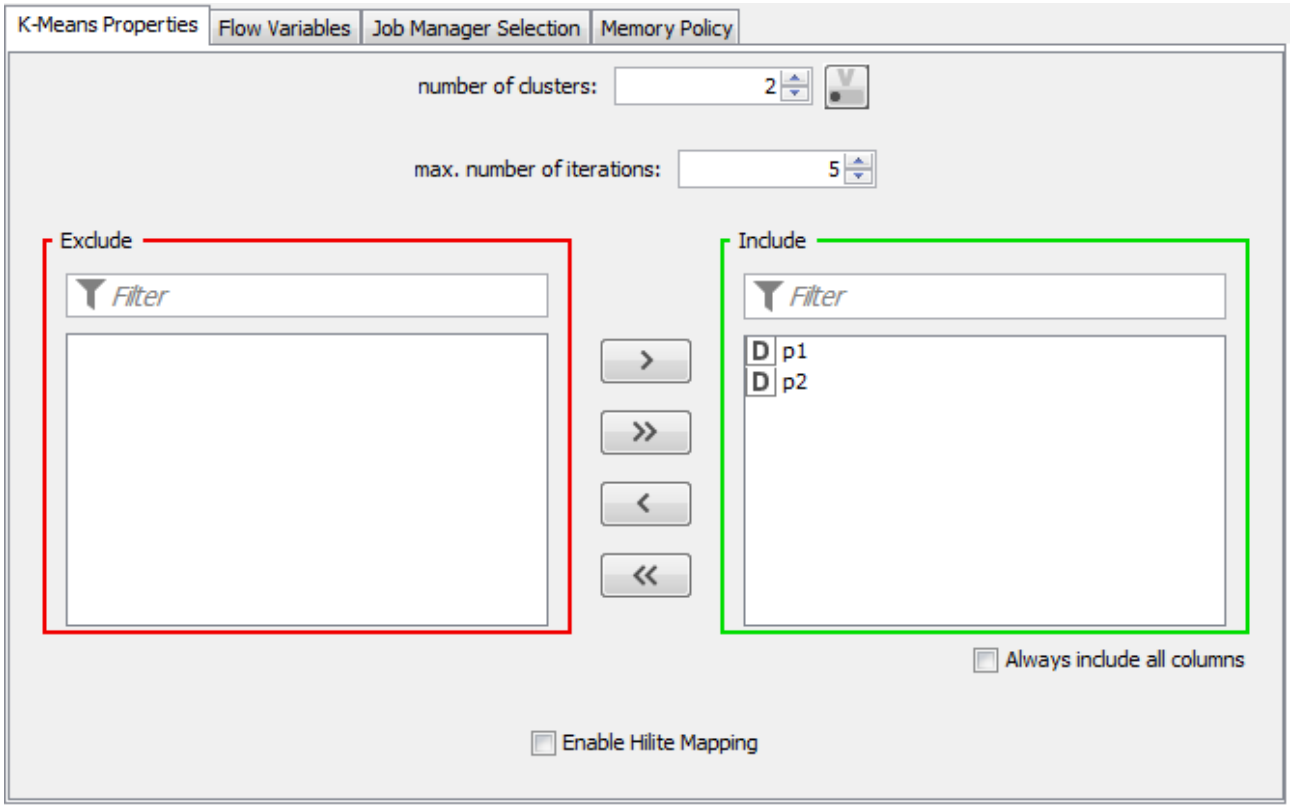


Fig. 3. La boîte de dialogue du nœud « K-Means ».

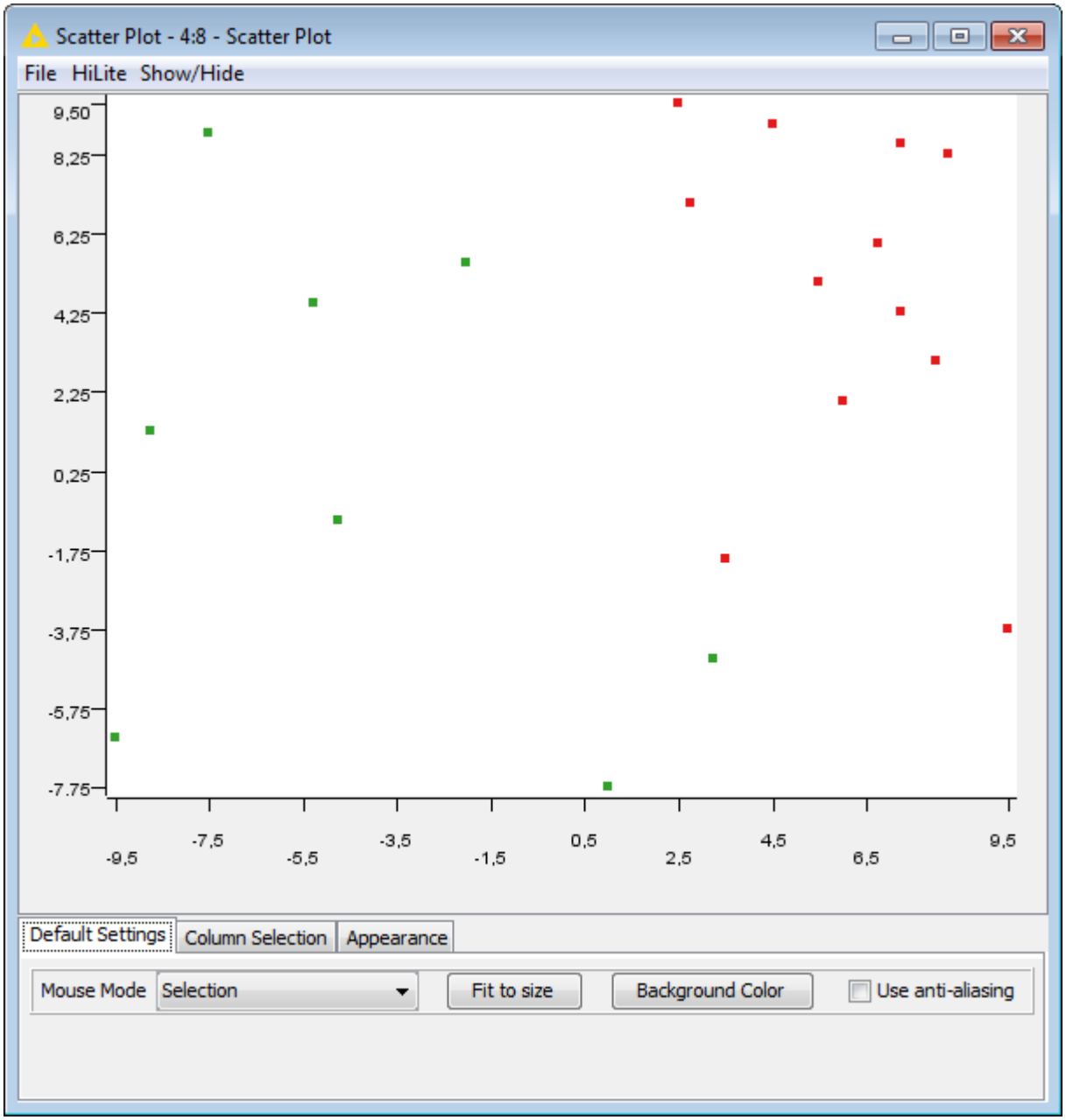


Fig. 4. Résultat de l'application de l'algorithme K-Means.

L'algorithme AHC (*Ascendant Hierarchical Clustering*)

1. Créez un nouveau workflow.
2. Ajoutez les deux nœuds suivants : « **Table Creator** » et « **Hierarchical Clustering** », puis les connecter comme indiqué dans la figure 5.

Fig. 5. Les nœuds du projet N° 2.

6. Configurez le nœud « **Table Creator** » pour saisir une matrice de données X représentant l'univers $\Omega = \{1, 2, 3, 4\}$ de 4 individus, où chaque individu est décrit par $p = 2$ variables. Exécutez ce nœud (le résultat doit être semblable à celui de la figure 6).

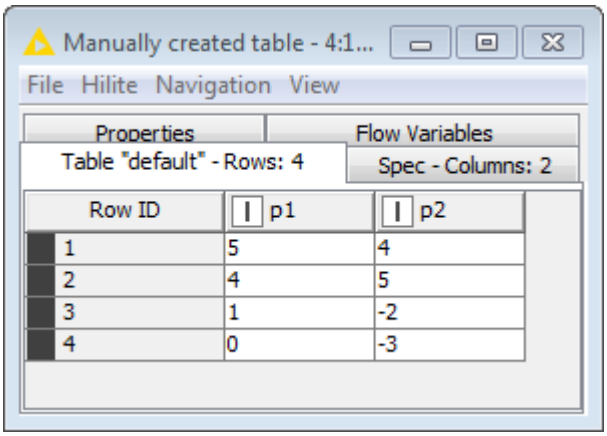


Fig. 6. La matrice de données *X* à saisir.

7. Configurez le nœud « **Hierarchical Clustering** » en effectuant notamment les choix sur le nombre de cluster de sortie, la fonction de distance et le type de liaison (quelle méthode utiliser pour mesurer la distance entre les clusters). Voir figure 7 comme exemple.

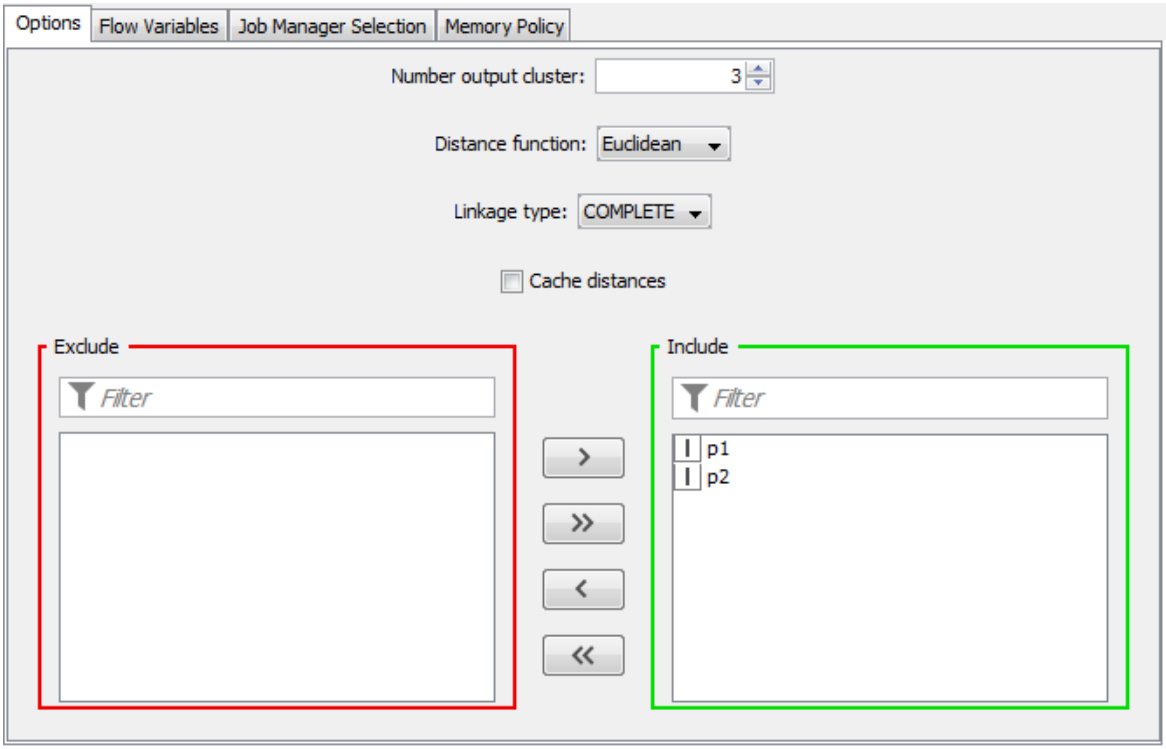


Fig. 7. La boîte de dialogue du nœud Hierarchical Clustering.

8. Exécutez l'ensemble des nœuds, puis visualisez le résultat à l'aide du nœud « **Hierarchical Clustering** » (Commande « **View : Dandrogram/Distance View** »). Voir figure 8.

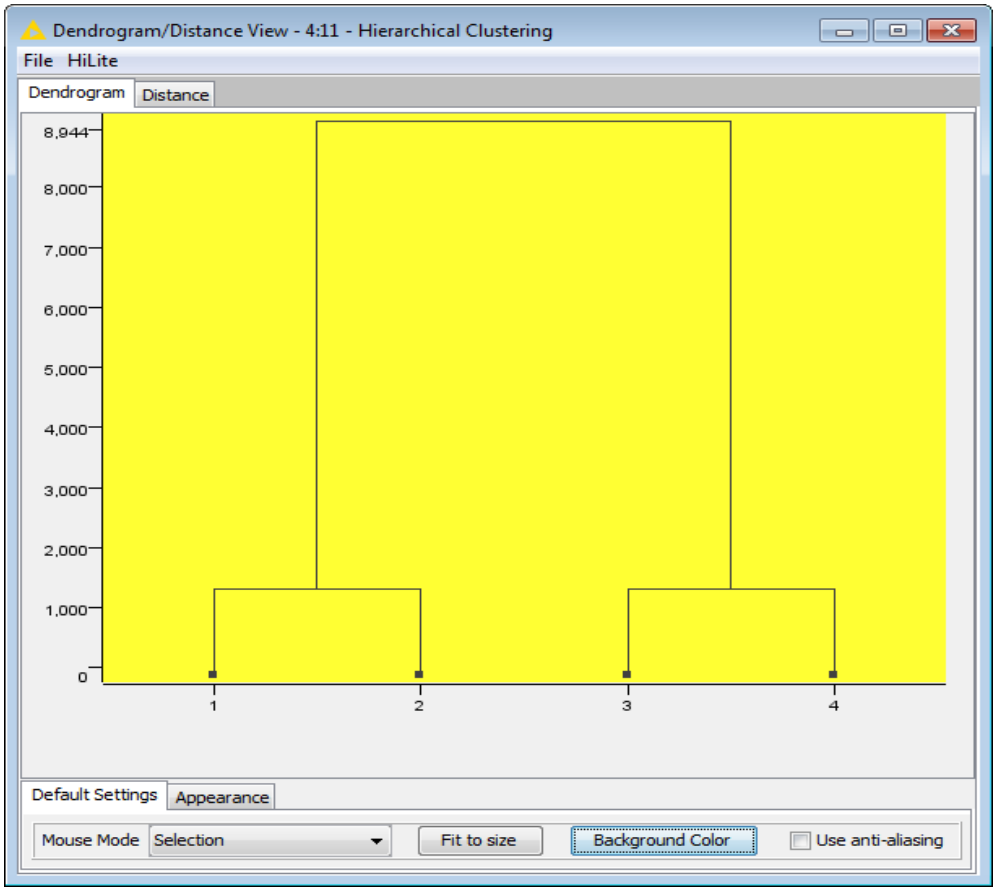


Fig. 8. Résultat de l'application de l'algorithme AHC.

L'algorithme DBSCAN

1. Créez un nouveau workflow.
2. Ajoutez les cinq nœuds suivants : « **Table Creator** », « **Numeric Distances** », « **DBSCAN** », « **Color Manager** » et « **Scatter Plot** », puis les connecter comme indiqué dans la figure 9.

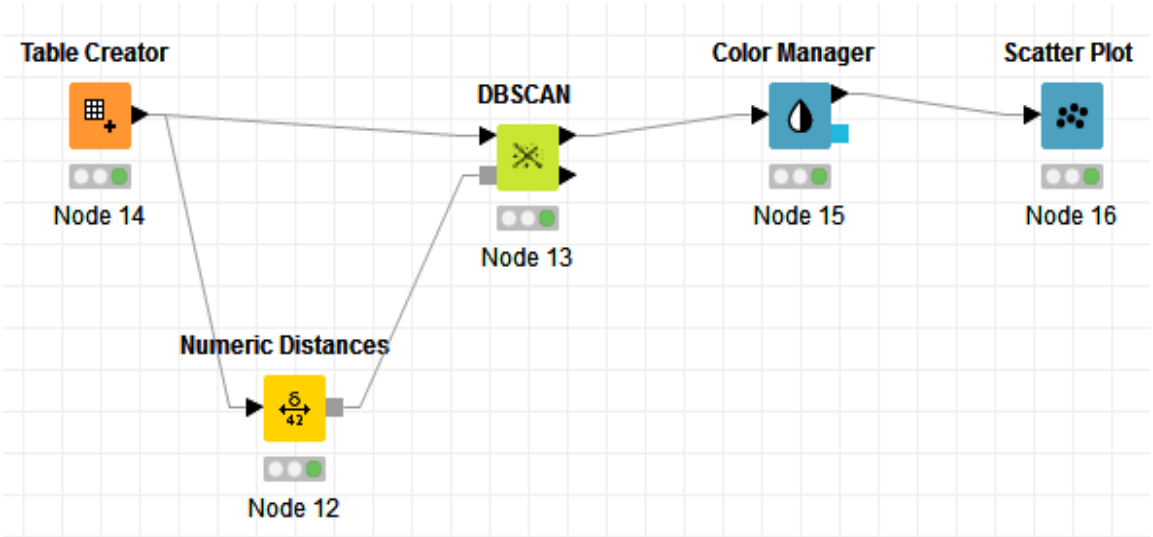
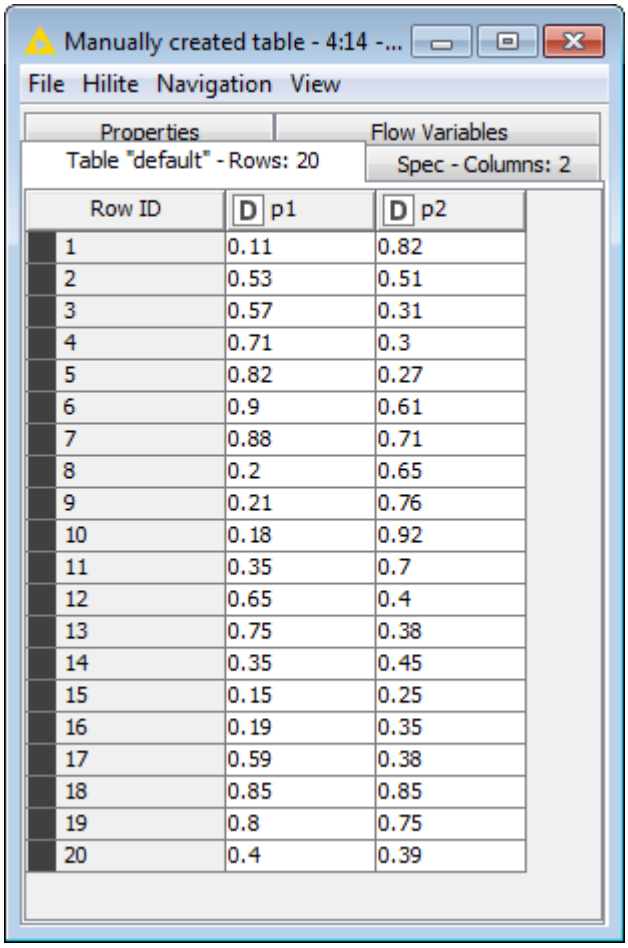


Fig. 9. Les nœuds du projet N° 3.

3. Configurez le nœud « **Table Creator** » pour saisir une matrice de données X représentant un univers Ω formé de 20 individus, où chaque individu est décrit par $p = 2$ variables. Exécutez ce nœud (le résultat doit être semblable à celui de la figure 10).



Row ID	D p1	D p2
1	0.11	0.82
2	0.53	0.51
3	0.57	0.31
4	0.71	0.3
5	0.82	0.27
6	0.9	0.61
7	0.88	0.71
8	0.2	0.65
9	0.21	0.76
10	0.18	0.92
11	0.35	0.7
12	0.65	0.4
13	0.75	0.38
14	0.35	0.45
15	0.15	0.25
16	0.19	0.35
17	0.59	0.38
18	0.85	0.85
19	0.8	0.75
20	0.4	0.39

Fig. 10. La matrice de données X à saisir.

9. Configurez le nœud « **Numeric Distances** » en effectuant notamment les choix sur la fonction de distance et les colonnes à inclure (voir figure 11).
10. Configurez le nœud « **DBSCAN** » en fixant les valeurs des deux paramètres ε et ***MinPts*** (voir figure 12).
11. Exécutez l'ensemble des nœuds, puis visualisez le résultat à l'aide du nœud « **Scatter Plot** » (Commande « **View : Scatter Plot** »). Voir figure 13.

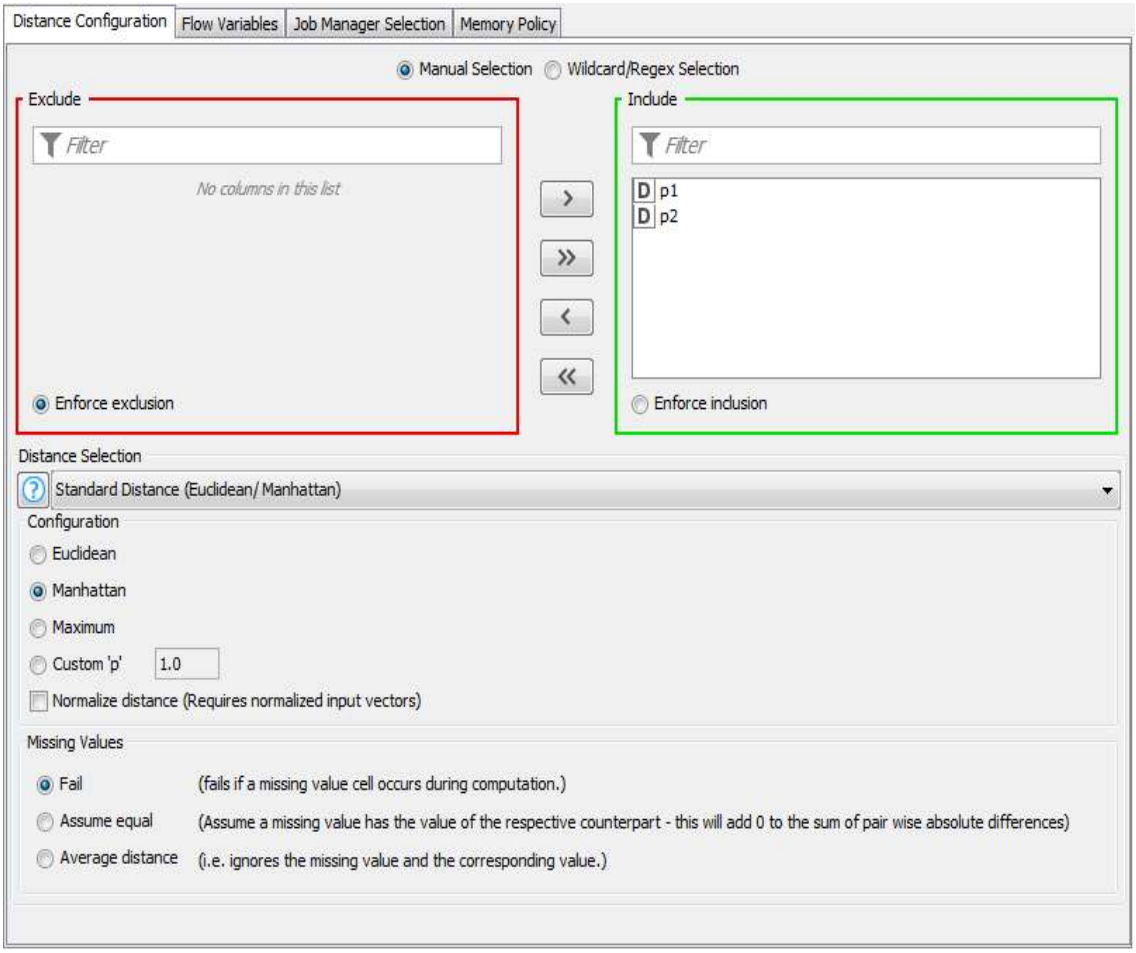


Fig. 11. La boîte de dialogue du nœud « Numeric Distance ».

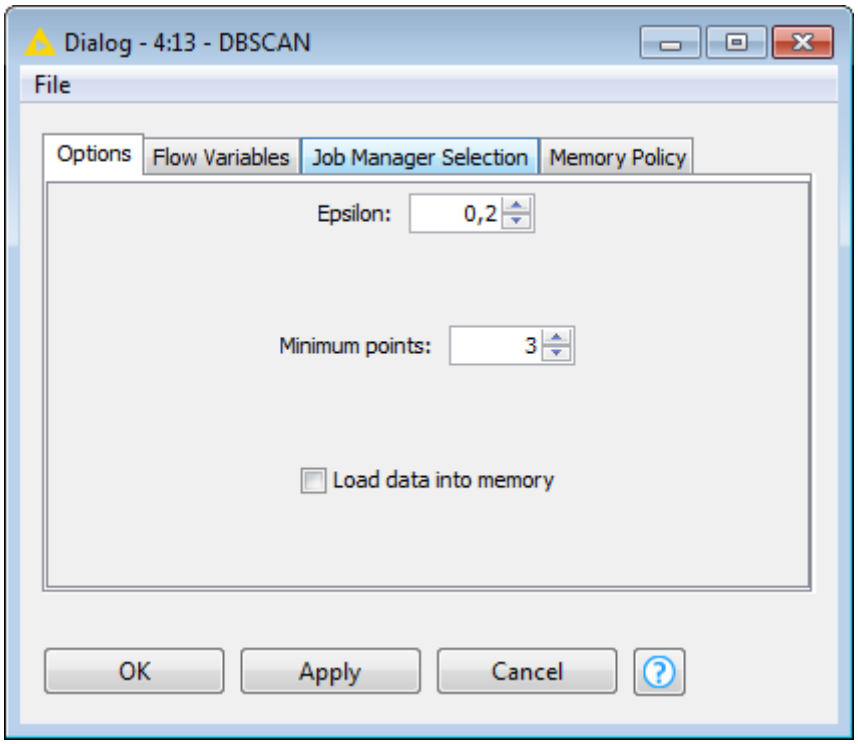


Fig. 12. La boîte de dialogue du nœud « DBSCAN ».

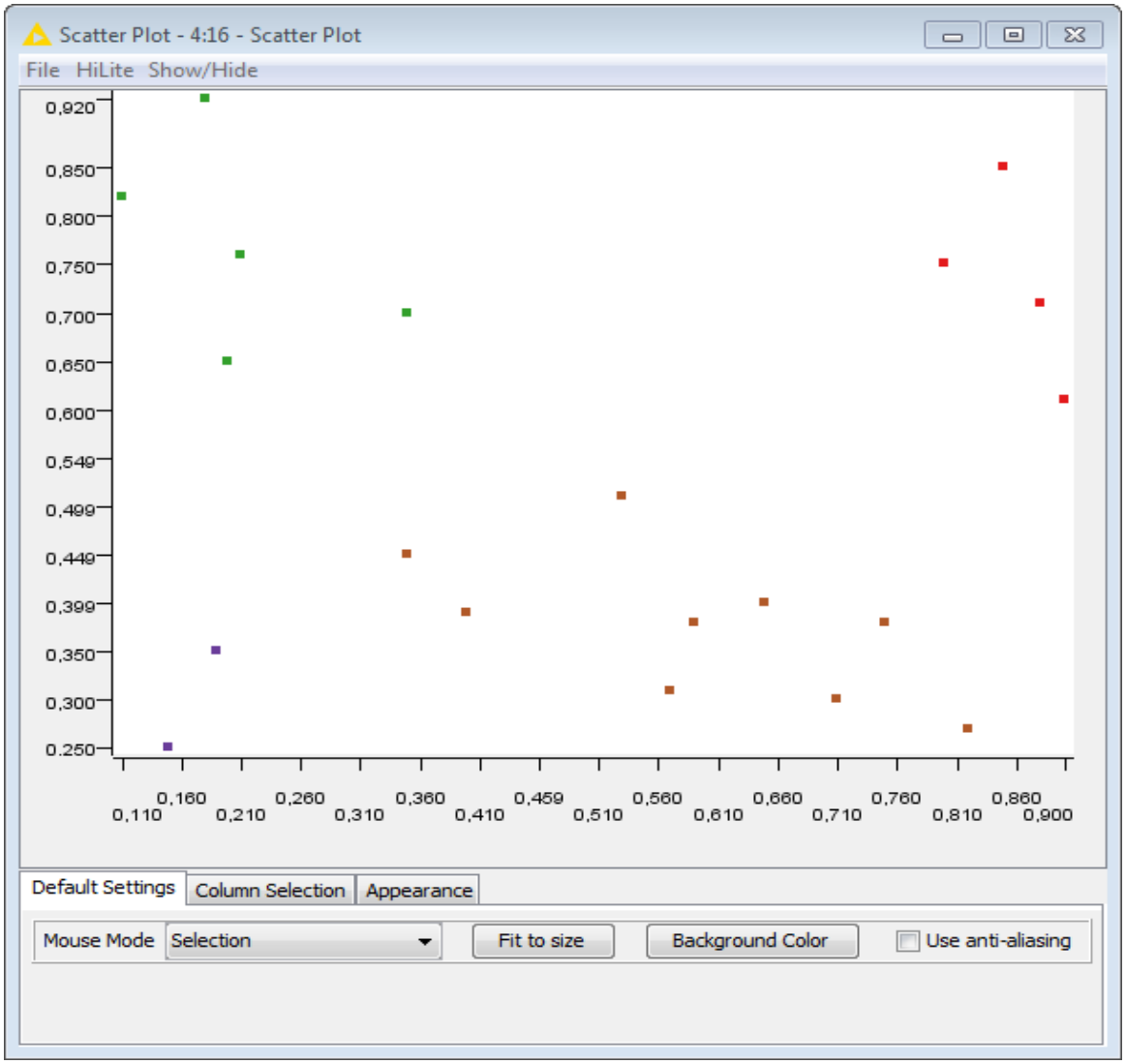


Fig. 8. Résultat de l'application de l'algorithme DBSCAN.