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° 2

OBJECTIF DE L'ACTIVITE PRATIQUE :

*Dans ce TP vous allez manipuler le logiciel **KNIME** pour effectuer des tâches de prétraitement de données (Valeurs manquantes, valeurs extrémales, normalisation des données).*

Création de données par le nœud « Table Creator »

Le nœud « **Table Creator** » de **KNIME** vous permet de générer des données d'entrée ad-hoc d'une manière très simple et rapide. L'avantage est que ces données seront enregistrées dans votre flux de travail courant.

1. Lancez **KNIME** (assurez-vous qu'il démarre avec un workflow vide).
2. Ajoutez un nœud « **Table Creator** » au workflow (**IO** / **Other** / **Table Creator**). Voir la figure 1.

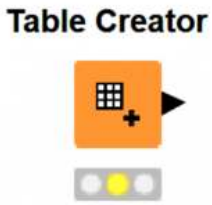

Fig. 1. Le nœud « Table Creator ».

3. Faites un double-clique sur le nœud « **Table Creator** » et sélectionnez la commande « **Configure** » à partir du menu contextuel. Dans la boîte de dialogue (figure 2) qui vient de s'ouvrir, cliquez dans la zone de la 1ère cellule du tableau (1ère ligne / 1ère colonne). Remplissez 20 valeurs numériques dans la 1ère première colonne en laissant 5 cellules vides (**valeurs manquantes**). Validez en cliquant sur le bouton « **Ok** ».
4. Exécutez le nœud, puis visualisez le résultat. Remarquez la présence de symboles d'interrogation « ? » : ils indiquent des **valeurs manquantes** (figure 3).

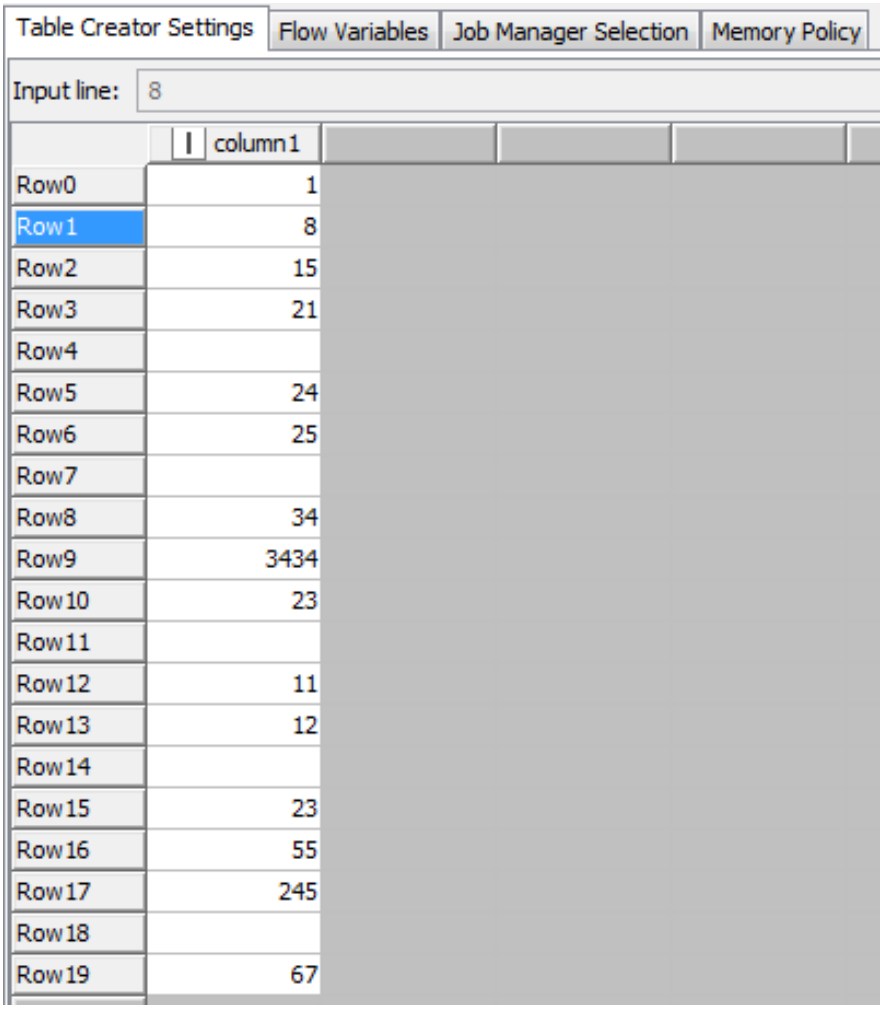

Fig. 2. Configuration du nœud « Table Creator ».

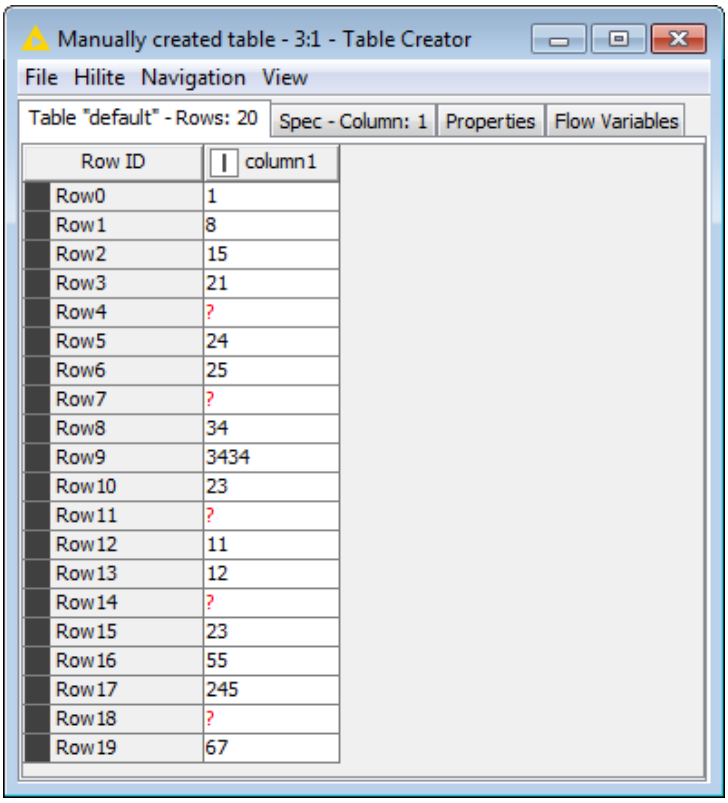

Fig. 3. Table obtenue après exécution du nœud « Table Creator ».

Manipulation des données manquantes

5. Ajoutez un nœud « **Missing Value** » au workflow (**Manipulation** / **Column** / **Transform** / **Missing Value**). Voir la figure 4.

Fig. 4. Le nœud « Missing Value ».

6. Reliez le nœud « **Missing Value** » avec la sortie du nœud « **Table Creator** », puis procédez à sa configuration. Choisissez une méthode pour traiter les valeurs manquantes en sélectionnant une entrée dans la liste de choix (figure 5). Par exemple, « **Do nothing** » pour ignorer les valeurs manquantes, « **Fix value** » pour saisir manuellement une valeur, « **Most Frequent Value** » pour affecter la valeur la plus fréquente dans la colonne, etc. Exécutez le nœud et voyez le résultat.

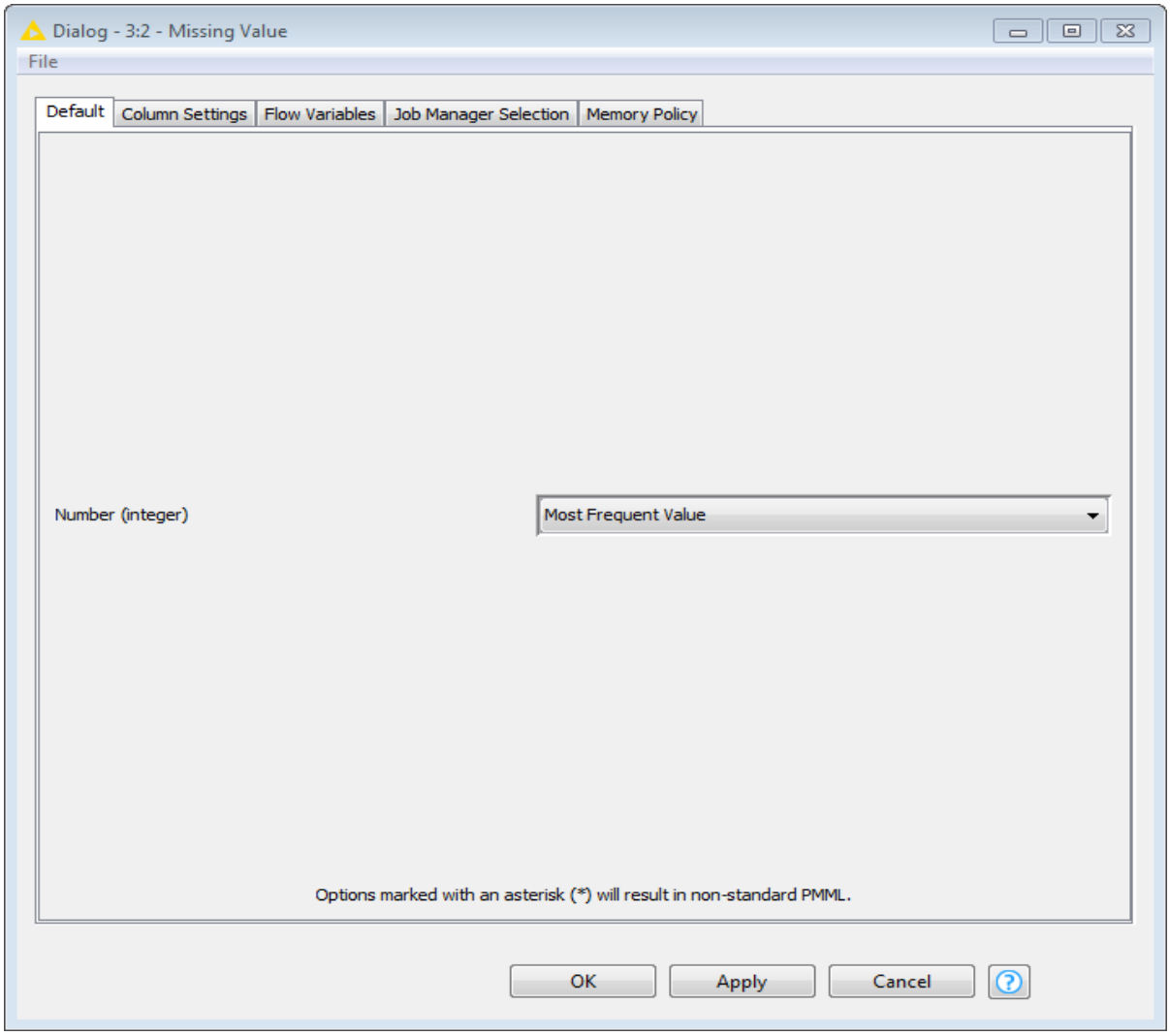

Fig. 5. Options pour le traitement des valeurs manquantes.

Manipulation des données extrémales

7. Ajoutez un nœud « **Numeric Outliers** » au workflow (**Analytics** / **Statistics** / **Numeric Outliers**). Voir figure 6.

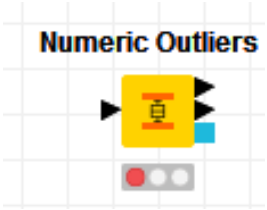

Fig. 6. Le nœud « Numeric Outliers ».

8. Reliez ce nœud à la sortie du nœud « **Missing Value** », comme dans la figure 7.

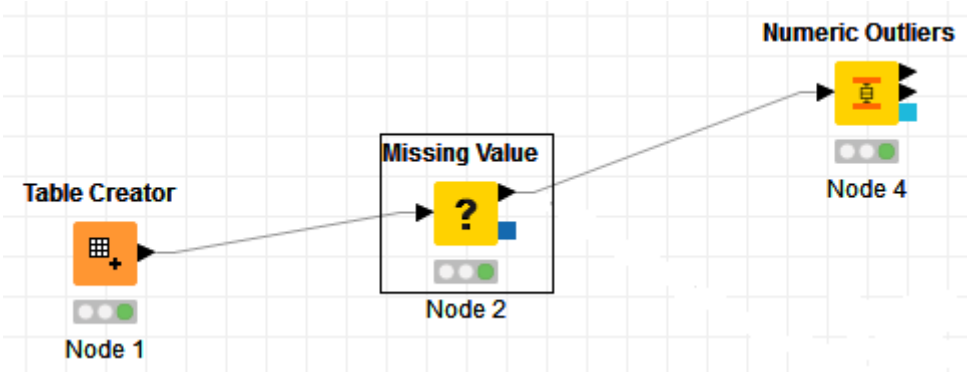

Fig. 7. Les trios nœuds reliés.

Le nœud « **Numeric Outliers** » détecte et traite les **valeurs aberrantes** pour chacune des colonnes sélectionnées individuellement au moyen d'une plage interquartile (IQR). Pour détecter les valeurs aberrantes pour une colonne donnée, les premier et troisième quartiles (Q_1, Q_3) sont premièrement calculés. Une observation est signalée comme aberrante, si elle se situe en dehors de la plage

$$R = [Q_1 - k \times (IQR), Q_3 + k \times (IQR)]$$

avec $IQR = Q_3 - Q_1$ et $k \geq 0$. La valeur $k = 1,5$ est la plus petite valeur de R. Si une observation est marquée comme une valeur aberrante, vous pouvez la remplacer par une autre valeur ou supprimer / conserver la ligne correspondante. Les valeurs manquantes contenues dans les données seront ignorées, c'est-à-dire qu'elles ne seront ni utilisées pour le calcul des valeurs aberrantes ni signalées comme des valeurs aberrantes.

9. Configurez le nœud « **Numeric Outliers** » en : (i) précisant les colonnes à traiter, (ii) indiquant les réglages généraux, (iii) précisant les traitements à effectuer aux valeurs extrémales. Voir figure 8.

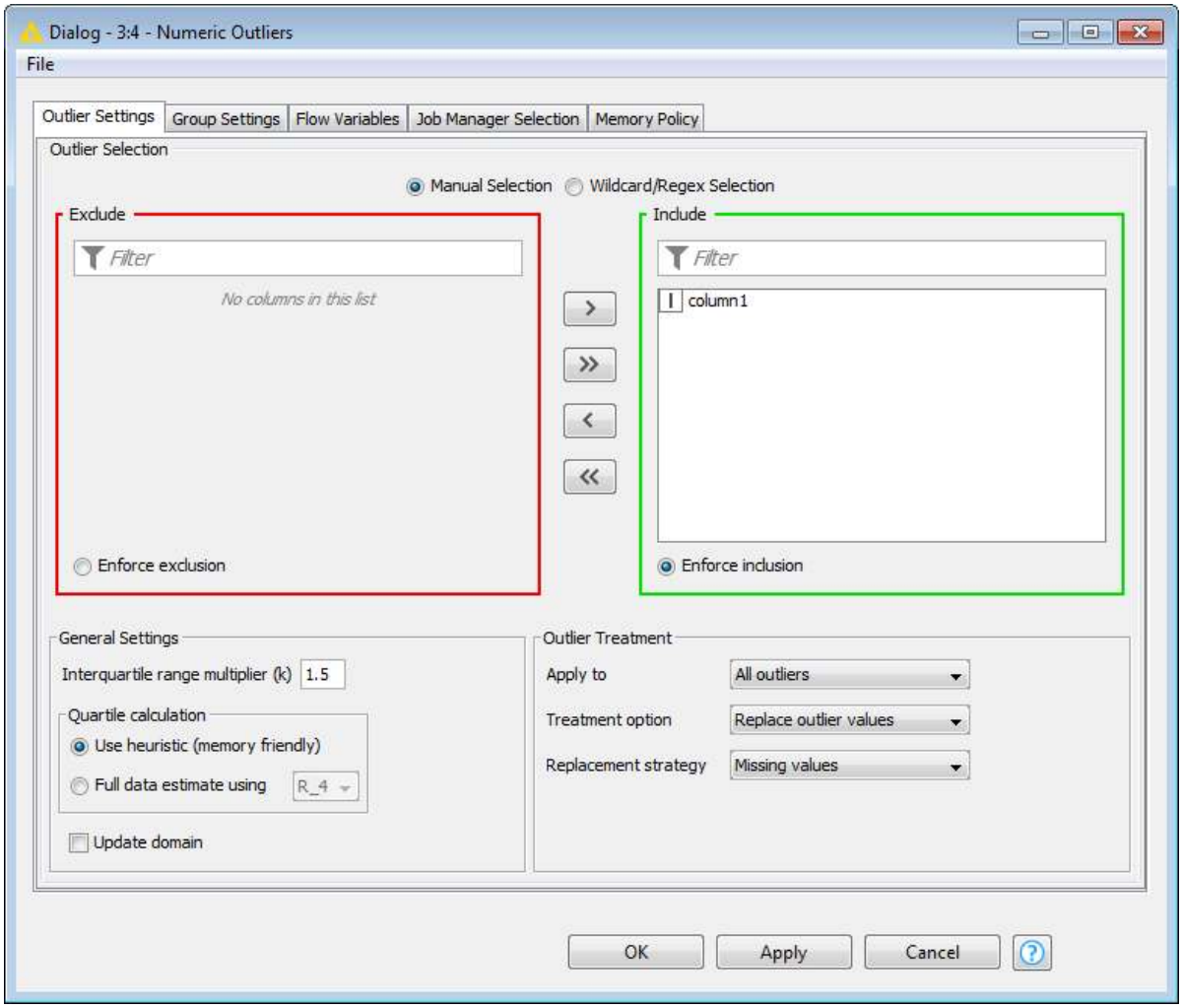

Fig. 8. La boîte de dialogue du nœud « Numeric Outliers ».

Le **calcul du quartile** permet de spécifier comment les quartiles sont calculés. Il y a deux méthodes :

(a) **Utiliser une heuristique** : cette option garantit que les quartiles sont calculés en utilisant une approche heuristique. Ce choix est recommandé pour les grands ensembles de données en raison de sa mémoire insuffisante. Cependant, pour les petits ensembles de données, les résultats de cette approche peuvent être assez éloignés des résultats exacts.

(b) **Estimation complète des données** : cette option permet en outre de spécifier le mode de calcul de la valeur réelle, qui est codée par les différents types d'estimation (LEGACY, R_1, ..., R_9).

Mettre à jour du domaine : si coché, le domaine des colonnes aberrantes sélectionnées est mis à jour.

Le **traitement des valeurs extrémales** :

(a) Le **cible du traitement** : (1) toutes les valeurs aberrantes, (2) les valeurs éloignées en dessous de la limite inférieure, (3) les valeurs éloignées au-dessus de la limite supérieure.

(b) L'**option de traitement** : trois stratégies différentes pour traiter les valeurs extrémales :

(1) **remplacer les valeurs aberrantes** : permet de remplacer les valeurs aberrantes en fonction de la "stratégie de remplacement" sélectionnée.

(2) **supprimer les lignes aberrantes** : supprime toutes les lignes des données d'entrée contenant au moins une valeur aberrante dans l'une des colonnes sélectionnées.

(3) **supprimer les lignes non aberrantes** : conserve uniquement les lignes des données d'entrée contenant au moins une valeur aberrante dans l'une des colonnes sélectionnées.

(c) La **stratégie de remplacement** : définit deux stratégies différentes pour remplacer les valeurs éloignées :

(1) **Valeurs manquantes** : remplace chaque valeur aberrante par une valeur manquante.

(2) **Valeur permise la plus proche** : remplace la valeur de chaque valeur aberrante par la valeur la plus proche comprise dans l'intervalle autorisé R. Si le type de colonne est un entier, la valeur de remplacement est l'entier le plus proche compris dans l'intervalle autorisé.

10. Exécutez le nœud « **Numeric outliers** » et observez les résultats après chaque configuration.

Normalisation des données

11. Ajoutez un nœud « **Normalizer** » au workflow (**Manipulation** / **Column** / **Transform** / **Normalizer**). Figure 9.

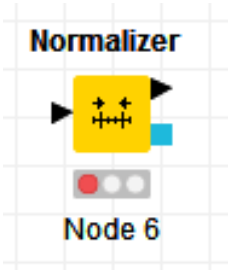

Fig. 9. Le nœud « Normalizer ».

12. Reliez ce nœud à la sortie du nœud « **Missing Value** », comme dans la figure 10.

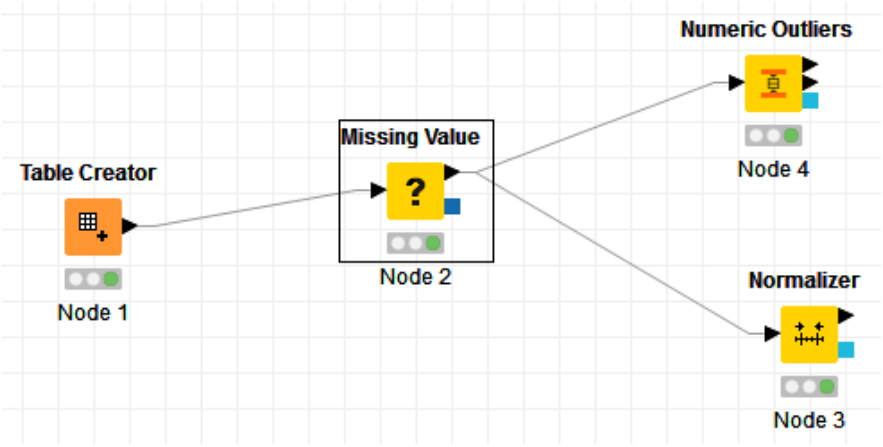

Fig. 10. Les 4 nœuds du projet.

13. Configurez le nœud « **Normalizer** », comme dans la figure 11.

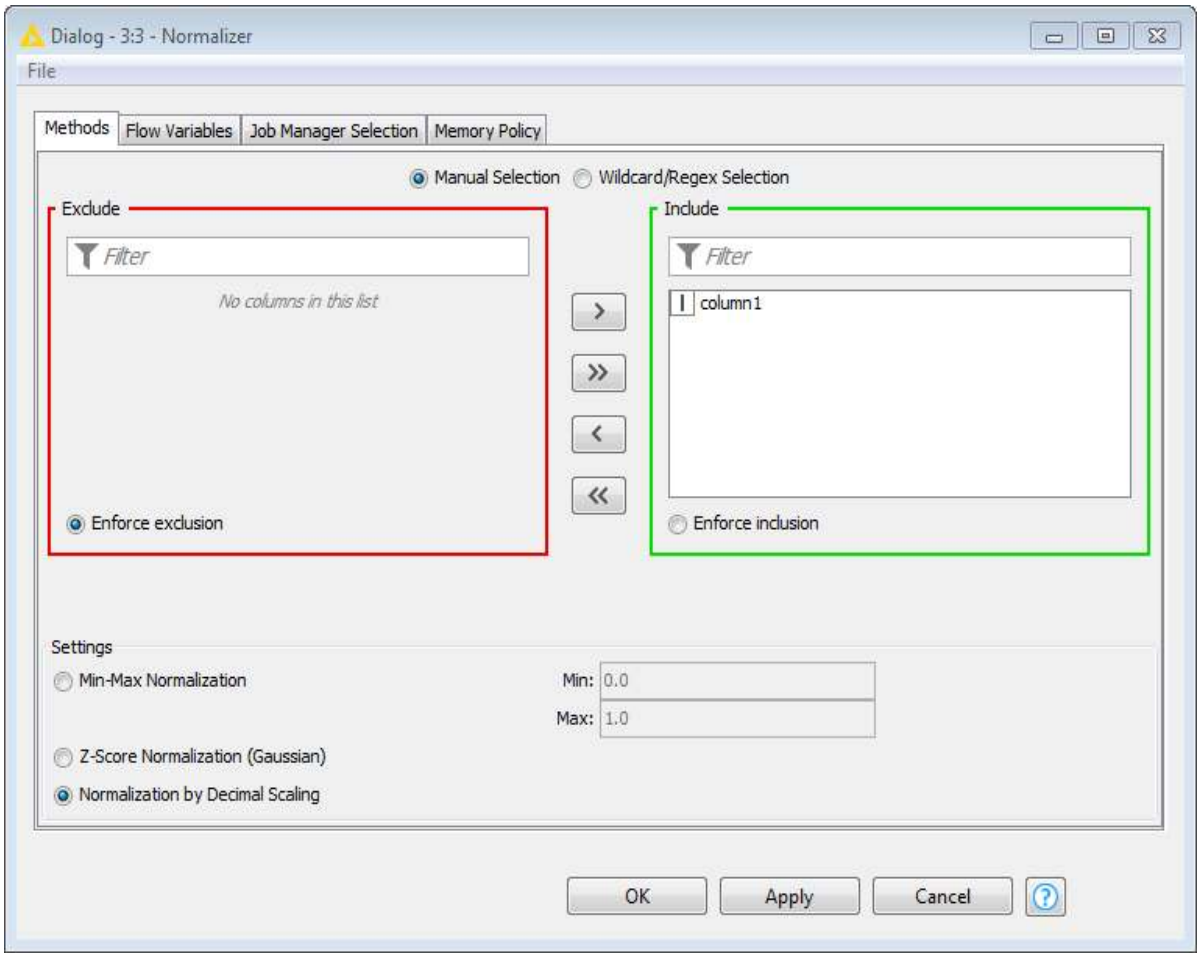

Fig. 8. La boîte de configuration du nœud « Normalizer ».

Ce nœud **normalise** les valeurs de toutes les colonnes de type numérique. Dans la boîte de dialogue, vous pouvez choisir les colonnes sur lesquelles vous souhaitez travailler. Les **méthodes de normalisation** suivantes sont disponibles dans la boîte de dialogue :

(a) **Normalisation Min-Max** : applique une transformation linéaire de toutes les valeurs telles que le minimum et le maximum de chaque colonne soient tels qu'ils sont donnés.

(b) **Normalisation du Z-score** (gaussien) : applique une transformation linéaire telle que les valeurs de chaque colonne correspondent à une distribution gaussienne (0,1), c'est-à-dire que la moyenne est 0,0 et que l'écart-type est 1,0.

(c) **Normalisation par mise à l'échelle décimale** : la valeur maximale dans une colonne (à la fois positive et négative) est divisée par 10 « j » fois jusqu'à ce que sa valeur absolue soit inférieure ou égale à 1. Toutes les valeurs de la colonne sont ensuite divisées par 10^j.

14. Exécutez le nœud « **Normalizer** » en essayant plusieurs méthodes, et observez les résultats après chaque configuration.